Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Cognitive challenges in human-AI collaboration: Investigating the path towards productive delegation

Andreas Fügener, Jörn Grahl, Wolfgang Ketter
University of Cologne, andreas.fuegener@uni-koeln.de, grahl@wiso.uni-koeln.de, ketter@wiso.uni-koeln.de
http://www.uni-koeln.de

Alok Gupta
University of Minnesota, gupta037@umn.edu, https://carlsonschool.umn.edu/faculty/alok-gupta

We study how humans make decisions when they collaborate with an artificial intelligence (AI): each instance of a classification task could be classified by themselves or by the AI. Experimental results suggest that humans and AI who work together can outperform the superior AI when it works alone. However, this only occurred when the AI delegated work to humans, not when humans delegated work to the AI. The AI profited, even from working with low-performing subjects, but humans did not delegate well. This bad delegation performance cannot be explained with algorithm aversion. On the contrary, subjects tried to follow a provided delegation strategy diligently and appeared to appreciate the AI support. However, human results suffered due to a lack of metaknowledge. They were not able to assess their own capabilities correctly, which in turn leads to poor delegation decisions. In contrast to reluctance to use AI, lacking metaknowledge is an unconscious trait. It limits fundamentally how well human decision makers can collaborate with AI and other algorithms. The results have implications for the future of work, the design of human-AI collaborative environments and education in the digital age.

Key words: Future of work, Artificial Intelligence, Deep Learning, Machine Learning

History: Submitted in August 2019

1. Introduction

Early AI that tried to mimic human decision rules was only partly successful as it suffered from what Autor (2014) calls Polanyi's paradox: The fact that humans can often not describe accurately by which decision rules they solve a problem. Modern AI like deep neural nets seems to overcome this limitation by learning flexible models from large training sets instead of relying on human-set rules (LeCun et al. 2015, Schmidhuber 2015). AI is now widely applicable and effective, and perceived as a general purpose technology (McAfee and Brynjolfsson 2017) that fuels innovation in diverse domains, such as medicine (Kononenko 2001, Esteva et al. 2017), and generic perceptional tasks, such as processing images, text, and speech (Hinton et al. 2012, Deng and Yu 2013). We agree that AI performance will likely improve further and that AI will be embedded in our day-to-day life. But, as Brynjolfsson et al. (2018) point out, not all decision making can be automated completely as some tasks remain challenging for AI.

Perhaps the best prediction we can make today is that humans will remain integral to the workplace, and they will work together with AI, algorithms, or intelligent machines. This is reflected in current IS research, having special issues on augmented intelligence in ISR and on managing AI in MISQ, both issued in 2021. Baird and Maruping (2021) propose a theoretical framework for next generation of IS research focusing on delegation from and to information systems, such as AI. They explicitly state that both the information system and the human could be the delegating unit. This mindset is in line with the concept of Human-in-the-loop AI (Zanzotto 2019), where humans remain an integral part of AI decision making.

The critical question in delegation is how firms should distribute work between humans and AI. Polanyi's work points to a facet of human decision making that is critical to the discussion, but vastly ignored until now. When humans can solve a problem, but are unable to explain their decision rules clearly, they should, nonetheless, be able to contribute complementarities to an algorithm (Autor et al. 2003). Since humans have different experiences and education, decision models and knowledge also differ between individual humans, and between humans and AI. Because we cannot articulate our idiosyncratic "decision rules" well, it is hard for learning algorithms to imitate them precisely, even from large training sets. Humans' inability to tell what they know shields their abilities from perfect digital imitation. The variety of human thought, only partially observable through their actions, creates the possibility that humans have complementary knowledge with respect to AI algorithms. The AI, on the other hand, may find a way to solve a problem no human being has thought about before. Thus, both humans and the AI potentially have complementary knowledge, and the performance of humans working with an AI system may be better than that of the AI system (or humans) working alone.

Our study focuses on the case where either an AI algorithm or a human is allowed to perform a task by itself, or to delegate that task to the other actor. On a generic level, there are three main boundary conditions that enable the combined performance to exceed the performance of the better performing actor:

1. Existence of complementarities: In order to enhance performance through delegation between two actors, complementary knowledge has to exist between the two actors, that is, a human and an AI system. We claim that this should be the case for all tasks where Polanyi's paradox applies, i.e., where humans cannot exactly specify their decision rules.

2. Recognition of complementarities: A delegating partner needs to recognize that complementarities between the two partners exist and that the tasks should be performed by the better-suited partner. While having information on the other partner's ability helps, the most important ability is to estimate one's own ability. If an actor knows that they can perform a task, it is always wise for them to complete the task themselves. On the other hand, it is wise to delegate if the actor knows that they cannot perform the task buttheir partner potentially can. In line with Lories et al. (1998) and Evans and Foster (2011), we denote this ability to assess own capabilities as "metaknowledge." Therefore, we argue that metaknowledge is a crucial resource for recognizing complementarities.

3. Execution of efficient delegation rules: Once a delegating partner recognizes complementarities, they haves to delegate tasks to the better-suited actor. While this can be easily constructed

and executed for an AI system, humans have to be willing to construct and follow such a delegation rule.

In a series of behavioral experiments, we investigate how work is delegated between humans and an AI algorithm. When designing the experiments, we aimed at making the results as generalizable as possible. Central to the paper is a delegation rule that does not make assumptions about the context in which it is used. Further, we conducted the experiments in a context where humans make good decisions naturally, and modern AI performs equally well or even better. We chose an environment where humans do not require any specific training. Contexts requiring specific training make the results less generalizable, whereas abilities in more general settings can carry over into more specialized tasks. Furthermore, expertise may not enable better decision making: Metaknowledge seems to be only minimally increased by training (Hansson et al. 2008), and trained experts could even show lower levels of metaknowledge compared to inexperienced subjects (Brezis et al. 2018).

Therefore, we chose image classification as the focal task for our experiments. While humans are naturally skilled at it, deep learning has improved AI algorithms beyond the human performance level recently (Russakovsky et al. 2015). In our experiments, humans and an AI algorithm work together on the same image classification tasks. In rest of the paper, we refer to to the image classification AI algorithm as the "AI." Humans can delegate images to the AI, and the AI can delegate images to humans in a condition called inversion.

Such collaboration between AI systems and humans is currently under-researched. A related field analyzes human decision makers' attitude toward algorithms. Several researchers have documented the reluctance of human decision makers to use algorithms (Bazerman 1985, Dawes 1979, Kleinmuntz 1990), although some recent research has challenged this notion. In a recent seminal work, Dietvorst et al. (2015) demonstrated that humans might react more strongly to errors made by machines than to errors made by humans, even if the machine performs better, and if its errors are smaller than those of human decision makers. The authors label this loss of confidence "algorithm aversion," a term picked up willingly be the popular press. But neither Dietvorst et al. (2015) nor a follow-up study (Dietvorst et al. 2018) document general human distrust towards algorithms. Logg et al. (2019) study multiple prediction tasks and find that humans are indeed willing to work with machines, a tendency they label as "algorithm appreciation."

We are not aware of research that studies delegation between AI and humans in a setting with complementary skills, and explores fundamental factors which might hinder or support collaboration. The central questions that our study answers are:

- Study 1: Can delegation between humans and AI outperform humans or AI working alone, and who can delegate better?
- Study 2: What factors limit human delegation performance? How can we overcome these limitations?

Our experiments reveal that while the AI improves considerably by delegating to humans, humans are naturally bad delegators to the AI. Even worse, and more interestingly, humans' performance only slightly improves when they are taught a good delegation rule, even when they apply it consistently and rationally. We observe little or no bias against the use of AI, in other words, our results indicate that our subjects do not exhibit general algorithm aversion. Instead, in our studies, humans try to work with the AI to best of their abilities, however, they fail despite their best intentions. Humans seem to be unable to judge their own capabilities, and/or the difficulty of the task, which in turn leads to bad delegation decisions. Thus, humans delegating to the AI do not meet the boundary condition of a sufficient level of metaknowledge to enable a successful human-AI collaboration. We also conduct additional robustness studies addressing the impact of continuous feedback and increasing task difficulty.

In the following, we summarize the theoretical underpinnings of our study and differentiate our study from the existing literature. We discuss theoretical antecedents in Section 2, describe the experimental studies and the robustness checks in Sections 3 and 4, and conclude with a discussion and an outlook on future research directions in Section 5.

2. Theory

We describe extant work and theories that inform our research questions and experimental design. In the following subsections, we discuss contributions of our work. Specifically, we shed light on human attitude towards algorithms (Section 2.1), consider the role of delegation settings (Section 2.2), the role of complementarities on the task instances level (Section 2.3), and the role of feedback (Section 2.4) to lay the foundation for our work.

2.1. Attitude towards AI

Research that studies how humans use computers for problem solving dates back decades and includes works that compare human decisions with results from mathematical models (Meehl 1954). Even during the infancy of computing environments in the 1950s, some computer models outperformed human decision makers. It was observed that in many instances, humans were reluctant to use algorithms, despite possible performance benefits (Bazerman 1985, Dawes 1979, Kleinmuntz 1990). However other studies, such as Dijkstra (1999), demonstrated general willingness of humans to use algorithms, even allowing machines to overrule their own inferior decisions. How humans make decisions in concert with algorithms has been revisited lately and has been of great interest with the surge in usage of AI techniques. Current applications look at investor usage of robo-advising services in fintech (Ge et al. 2021), reactions to AI advice in health care (Jussupow et al. 2021), or the effect of similarity to human language used by chatbots (Schanke et al. 2021).

In a seminal work on attitude towards algorithms, Dietvorst et al. (2015) demonstrate that humans react differently to errors made by humans as compared to algorithms. In their experiments, they let humans work with AI algorithms for prediction tasks. When the subjects saw the AI perform, and err, they lost confidence in it. Interestingly, this loss is much stronger than the loss in confidence in humans who made mistakes in the same task. The authors label this tendency "algorithm aversion." Logg et al. (2019) studied whether human decision makers prefer external advice from other humans or algorithms and found that decision makers show a clear tendency for preference for algorithmic advice over human advice. This preference holds for multiple prediction tasks, and (according to a survey the authors conducted) was not expected by most academics. In contrast to Dietvorst's work, Logg et al. (2019) label their results as "algorithm appreciation." While this seems like a contradiction, Dietvorst's work does not show a general aversion towards algorithms but rather a different reaction towards errors made by algorithms and human decision makers.

2.2. Delegation

We consider a scenario, where a human has to perform a task without information about the solution or performance of the AI for a specific task. We denote this as a "delegation" scenario. We allow delegation to work in both directions. The human may delegate work to the machine, and the machine may delegate work to the human. The latter approach is sometimes called inversion, or using the human as an exit option (McAfee 2013). In delegation settings, a good decision heuristic is the following: "If I am certain to know the correct answer, I should do the job. If I am uncertain I should delegate!" This rule works well, because delegating a task that the decision maker is not able to perform cannot decrease performance, independent of the other party's abilities.

When humans apply this rule, they have to rely on their metaknowledge. This is the ability to assess one's own capabilities, that is, to know what you know (Lories et al. 1998, Evans and Foster 2011). A decision maker with strong metaknowledge can delegate well, as she will know whether her answer is correct or not. If her level of metaknowledge is insufficient, she might be certain that incorrect answers are correct, and she might be uncertain about the correctness of correct answers. In such cases, the joint performance of a human and an algorithm will suffer due to inappropriate delegations.

Note that this entire idea – to delegate tasks when you are uncertain – is particularly relevant if delegation does not move an entire stack of tasks, or all the work, but when it occurs on the level of task instances. We discuss this in the next section.

2.3. Complementarity on the instance level

Occupations typically consist of bundles of tasks (Autor et al. 2003), some of these tasks are suitable for machine learning, others not (Brynjolfsson and Mitchell 2017). Because of this many experts do not expect that AI will automate entire bundles of tasks associated with a job but only specific parts of the bundle (Brynjolfsson et al. 2018). Therefore, a likely consequence of automation on the task level is that some tasks are automated, while others are not. This leads to redesigned job profiles and workflows based on the economic benefits associated with such work arrangements, e.g., Agrawal et al. (2018) discuss ways to compute the economic value of automation on the task level and present a related AI canvas.

We take this argument further and argue that structural complementarities between humans and AI may exist even on the *task instance* level. Our reasoning builds on the design principles for current AI systems. While traditional expert systems were built by humans who coded concrete decision rules, current AI algorithms discover their own decision rules based on training data. Due to structurally different decision rules, we argue that for each task there might be instances where human decision rules work better that AI decision rules, and vice versa.

Sharing work between humans and AI algorithms on the task level can leverage these complementarities, and joint performance of an AI working with humans may exceed the performance of either of the parties individually. Even if the AI performance is better than the human performance, the optimal allocation will assign some work to a human and some to the AI, and humans and AI both can provide value. Therefore, it is important to conduct research in the area of human-AI collaboration on the level of task instances. As we demonstrate in this paper, it offers significantly different implications for the future of work than the established paradigms have argued and/or demonstrated.

2.4. The role of feedback

While some extant research has used the effect of feedback on task level performance and accuracy, we decided to not include such feedback to concentrate on our main research questions. The reason for our choice is manifold. For successful human-AI collaboration, several factors have to be considered on the human side of the equation. When prior research considered environments where humans make their decisions on the basis of observed AI performance and errors, they found that receiving immediate feedback on AI performance might trigger behavioral effects that undermine the collaborative setting. Prominent examples include diminished trust in algorithms (measured by a lower adherence to algorithmic advice) when humans see the algorithm err (Dietvorst et al. 2015), or the overweighting of signals from forecasting errors (Kremer et al. 2011). In that case, errors resulting from random variation of data are misinterpreted as systematic errors.

There are many situations where AI feedback is not even available or practical, for example when predicting long-term effects, such as climate change (Logg et al. 2019). In other situations, decisions have to be made quickly, such as in autonomous driving, or very frequently, such as in digital markets. In these contexts, AI errors are either not available, it may not be economical to consider them repeatedly, or there may simply be no time to integrate AI performance in decision making.

While we do not study the effects of feedback in our main study, we do explore whether continuous feedback affects our findings in a dedicated study in the robustness check section.

In the following sections, we provide details of our experimental studies.

3. Experimental Studies

After providing a rationale on the study context, we describe the hypotheses, designs and results of two primary experimental studies with 902 subjects in total. We followed Nosek et al. (2018) and pre-registered the experiments at the Open Science Foundation (Foster and Deardorff 2017), including the recruitment and data collection process, the initial hypotheses, and the statistical analysis.

3.1. Study context

When designing the experimental studies, we aimed for a non-specialized setting, as we claim that contexts that do require specific training make results less generalizable, while findings in general tasks can carry over to more specialized tasks. We also aimed for a task, where the three boundary conditions for value-adding delegation between humans and AI are potentially met: 1) Existence of complementarities, 2) recognition of complementarities, and 3) execution of efficient delegation rules. We chose image classification as our experimental study context. Image classification is the task of assigning a focal image to a class. A class can be thought of as a content group. A classification is correct if the focal image is assigned to the right class (a focal picture with the ground truth "poodle" is assigned to the "poodle" class, not to "huskie" or "cat"). We sampled 100 focal images with known class labels from the ImageNet database. The ImageNet database consists of tens of millions of human-annotated images that are used by current image recognition challenges, such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al. 2015). We sample images used in the ILSVRC image classification task that contain everyday objects and animals, and that are assigned to one of 1,000 possible classes. The difficulty of each task may be associated with three main dimensions: 1) The image itself. This might include visibility of the object, size of the object, or whether multiple objects are present. 2) Possible classes. A main driver for difficulty of image classification tasks is the definition of possible choices, for example, fine-grained recognition between similar classes, such as breeds of dogs, is more difficult compared to more different classes, such as between zebra, lion, or tiger. 3) The annotator. Familiarity with image and with possible classes is a big driver of subjective difficulty. This could depend on training data (in case of an AI) or on personal experiences and interests (in case of a human). A difficulty of human annotators is to cope with a large number of classes, as they might not be aware of the existence of a specific class (Russakovsky et al. 2015). To avoid this effect, we chose to display ten possible classes along with the focal image. As in Russakovsky et al. (2015), we illustrated each possible answer class by name and 13 example images. One answer was correct. A central performance measure was classification accuracy, the percentage of correctly classified images.

We chose GoogLeNet Inception v3 (Szegedy et al. 2016) as AI. It is among the best AIs for image classification and was trained on the ImageNet database with 1,000 classes. GoogLeNet assigns a score to each classthat can be interpreted as the likelihood of being correct. We obtained its classification accuracy by applying it to the 100 images, and by comparing the image with the highest score to the correct answer. As the AI is trained based on outcome data, the decision rules differ from human decision rules, and complementarities between the AI and humans should exist.

We recruited human subjects via Amazon's Mechanical Turk (MTurk). We believe using MTurk to recruit subjects is particularly suitable for our study for the following reason: we are interested in assessing basic human capabilities and image classification is a natural task for humans that requires no specific training. Many tasks at MTurk relate to classification problems (Difallah et al. 2015), thus, our experiment is a natural task for MTurk workers. We provide evidence that subjects took the tasks seriously and performed them with a high degree of internal validity. They made logical delegation decisions based on their internal assessment (see Figure 6 and the subsequent discussion).

Study 1 "Delegation and Inversion" compares four different types of delegation: AI working alone, humans working alone, humans who may delegate to AI, and an AI that may delegate to humans (inversion). As expected, the AI outperforms humans. Surprisingly, the AI delegates better than humans when it follows a simple rule. Study 2 "Explaining and Enforcing a Delegation Strategy" explores the root cause for poor human delegation and analyzes the effects of teaching humans a similar strategy to that of the AI.

3.2. Study 1: Delegation and Inversion

Hypotheses. Study 1 tackles our first research question: can delegation outperform humans or AI working alone, and who can delegate better? Thus, we compare four different settings: AI working alone, humans working alone, humans who may delegate to AI, and an AI that may delegate to humans (inversion). We chose image classification as focal task, and use an AI that is expected to perform (slightly) above human performance. Our key measure is classification accuracy, that is, the percentage of correctly classified images. We formulated and preregistered four initial hypotheses considering the relation of accuracy between those options. In the following, we present three of those hypotheses and theory that motivates them. One pre-registered hypothesis claimed that a state-of-the-art AI outperforms human decision makers on average. While this is supported for our specific setting (Szegedy et al. 2015, 2016), it lacks generality, and we decided to exclude it.

The first two hypotheses motivate the value added through delegation. In the introduction, we defined three boundary conditions: Existence of complementarities between the AI and humans, recognition of complementarities, and execution of efficient delegation rules. We assume the first property, existence of complementary knowledge, to hold for all tasks that follow Polanyi's paradox, such as image classification. While we are not aware of any structural evidence regarding the recognition of complementarities and the execution of efficient delegation rules, there is ample research from the domain of humans working with decision support systems, ranging from seminal theoretical work, such as Huber (1990), to recent experimental studies as carried out in Dietvorst et al. (2018), confirming that humans can benefit from working with advanced information technologies, and that the second and third boundary conditions are at least partially given. This directly leads to our second hypothesis:

Hypothesis 1.1: Humans who can delegate tasks to the AI (after seeing the image to be classified) perform better than humans who can not.

A more difficult question is to hypothesize on the effect of providing AI the possibility to delegate to humans (inversion), given that human performance is potentially inferior to that of the AI. To be able to improve accuracy, the AI has to delegate those tasks that AI is not able execute with accuracy, but humans can potentially execute with higher accuracy. For the second boundary condition, the recognition of complementarities, it is important that the AI can assess its own certainty, that is, probability of success. Assessing its own certainty is a main feature of modern AI that has gone through appropriate level of training, and enables the AI to perform in a robust manner by using its certainty assessment to make the final choice. In our case of image classification, the AI score of the sample of 100 images estimated an average likelihood of being correct of 0.769, and classified 77 images correctly. Using AI score as an indicator for certainty and some benchmark for expected human certainty, we define an efficient delegation mechanism that the AI follows to leverage the potential of complementary knowledge. We formulate our next hypothesis:

Hypothesis 1.2: AI that can delegate image classification tasks to humans (after seeing the image) performs better than an AI that can not.

In theory, both delegation and inversion could achieve the same accuracy: If the delegating actor delegates all tasks that she or he is not able to perform, all tasks that at least one could perform would be considered correct. We denote this as ex-post optimal combination of humans and AI, where all complementarities are realized. The AI in our inversion condition applies such a strategy under uncertainty.

Whether humans or AI is better at delegating tasks might depend on the second and third boundary conditions of successful delegation: humans need to have a sufficient level of metaknowledge, that is, correctly identifying tasks where they do not perform well, and humans have to come up with an efficient delegation strategy, and be willing to follow it through. We know that the AI has a very high level of metaknowledge, and will follow an efficient delegation strategy, whereas both boundary conditions are uncertain for human delegators. This leads to our final hypothesis of Study 1:

Hypothesis 1.3: AI that can delegate image classification tasks to humans performs better than humans who can delegate to the AI.

In the following, we lay out the details of our study design before presenting our results.

Design. We compare classification accuracy between four conditions. In the “AI alone” condition (1), GoogLeNet classified alone. In the “humans alone” condition (2), subjects classified alone. Subjects in the delegation condition (3) could choose for each image to either classify alone, or to delegate the image to the AI (subjects were informed about the AI accuracy measure). In the inversion condition (4), the AI could choose for each image to classify alone or to delegate the image to humans.

For conditions (2)-(4), we ran a between-subject design with 449 subjects in August 2018. We randomly assigned subjects to the conditions humans alone (149 subjects), delegation (154 subjects), and inversion (146 subjects). The humans alone (2) and inversion (4) conditions were identical. Figure 1 shows a screenshot of the humans alone/inversion and delegation conditions.

In the delegation conditions, we added a button labeled “Delegate this question to the AI” at a random position between the answers. If a subject clicked, she did not classify the image herself, but delegated it to the AI. She would not see the AI’s answer. The AI’s answer was considered

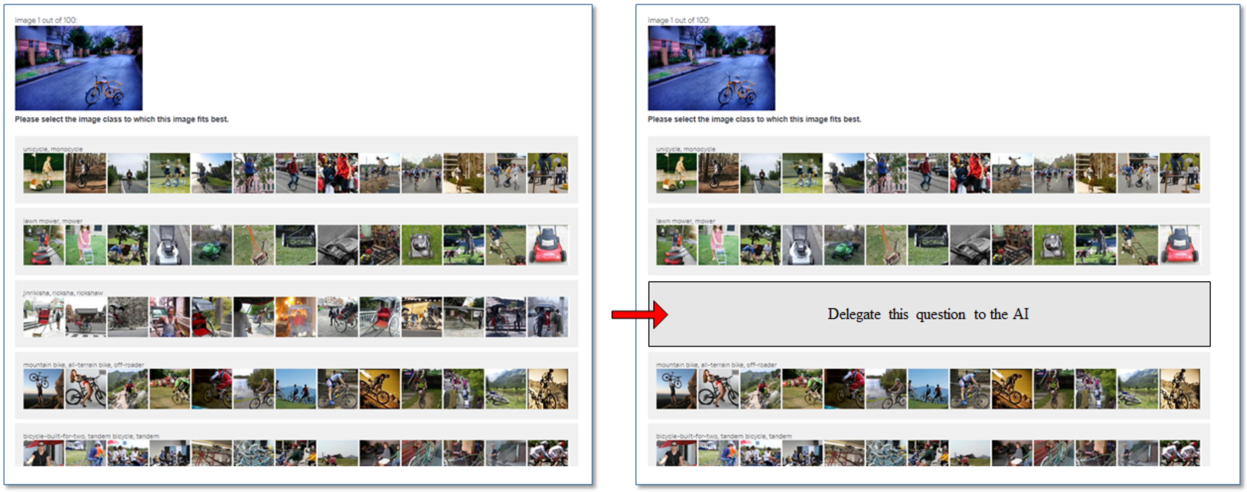

Image 1 out of 100:

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

unicycle, monocycle
lawn mower, mower
jinrikisha, ricksha, rickshaw
mountain bike, all-terrain bike, off-roader
bicycle-built-for-two, tandem bicycle, tandem

Image 1 out of 100:

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

unicycle, monocycle
lawn mower, mower
Delegate this question to the AI
mountain bike, all-terrain bike, off-roader
bicycle-built-for-two, tandem bicycle, tandem

Figure 1 Screenshots of the humans alone/inversion condition (left) and the delegation condition (right).

hers, and she received her payment accordingly. We made it clear that each correct classification earned a payment, regardless of whether the AI or the human classified the image. Subjects in the delegation condition were informed about the AI and its accuracy at the beginning of the experiment.

To ensure that the effects can be related to different delegation, and not to different human classification behavior, subjects in the inversion condition had to classify all 100 images, like subjects in the other conditions. We constructed the results for the inversion condition (4) after the experiment. The AI classifies images or delegates them to humans based on a simple rule: if the score for the best answer was below a certain threshold, then GoogleNet delegated this image to the humans. Otherwise, GoogLeNet classified the image. To simulate this mechanism we paired GoogleNet with each subject from the inversion condition. The threshold was the average accuracy of subjects in the humans alone condition (2). The AI thus delegated all images where the estimated likelihood of being correct was below average human accuracy.

All subjects received instructions, had to pass a short quiz so that we could exclude robots, and completed an example classification to ensure that they understood the task. They then had to classify the 100 images in random order. Each subject received a base fee of 50 cents, and an additional 5 cents for each correct answer. Afterwards, they were asked how many images they think they classified correctly. They could earn 1 additional dollar if this estimation did not differ from the actual number by more than five images.

Average pay was \$4.45, slightly above average pay on MTurk in general (Hara et al. 2018). The average duration of the experiment was 57.7 minutes.

Table 1 Summary statistics for accuracy (Study 1).

		Summary statistic							
Dep. Var.:	Treatment	N	Min.	Mean	Max.	St. Dev.	Pctl(25)	Median	Pctl(75)
Accuracy									
	AI alone			0.770					
	Humans alone	149	0.310	0.716	1.000	0.132	0.650	0.740	0.810
	Delegation	154	0.250	0.740	0.990	0.101	0.700	0.760	0.800
	Inversion	146	0.710	0.870	0.980	0.042	0.850	0.870	0.898

Results. Descriptive statistics (Table 1) and visual evidence (Figure 2) suggest that the ability to delegate affects classification accuracy. On average, accuracy is highest in the inversion condition (87.0%), followed by the delegation condition (74.0%) and humans alone (71.7%). By itself, AI accuracy is 77.0% (vertical dashed line in Figure 2). The standard deviation of accuracy (in number of images) is highest when humans work alone (13.2), smaller when humans can delegate (10.1) and smallest when the AI delegates to humans (4.2). We used .717 as the threshold in the inversion condition. The results for inversion are robust for different threshold values (inversion accuracy is above .840 for all thresholds between the 25th and 75th percentile of human performance, that is, .650 and .810). The ex-post optimal combination of human and AI of Treatment 1 would lead to an upper bound of 89.9% accuracy – that is, assuming that each image is classified correctly, where either the AI or the human classified the image correctly.

The variance of accuracy is significantly different across experimental conditions (Levene test, F(2, 446) = 36.752, $p < .001$; Hartleys F_{max} test, $F_{max} = 9.962 >$ critical value), and means are significantly different as well (ANOVA with heterogeneous variances, F(2, 245.05) = 178.41, $p <$ .001, $\eta^2 = .315$, which represents a large effect). Post-hoc tests with Tanhames T2 statistic for

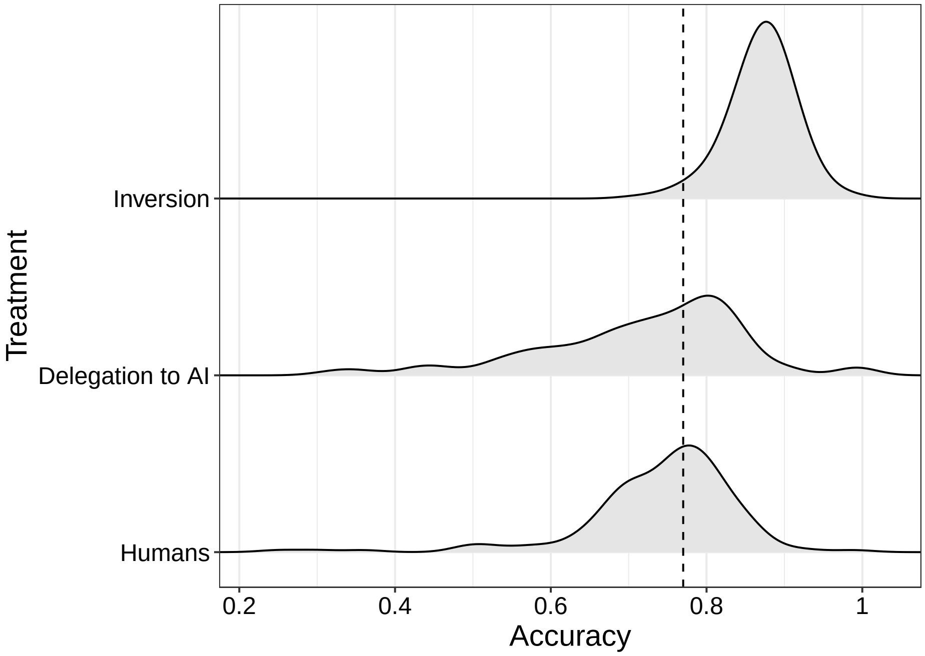

Figure 2 Distribution plots for accuracy per experimental condition (Study 1). The vertical dashed line is the AI classification accuracy of 77%.

multiple comparisons show that most pair-wise mean differences are significant (see Table 2 for a summary of pairwise comparisons). Humans in the *delegation* condition seem to outperform *humans alone*. However, this difference (2.37 percentage points) is not significant ($p = .120$) and represents a relatively small effect ($d = .2$). *Inversion* clearly outperforms *humans alone*. This difference (15.38 percentage points) is significant ($p < .001$) and represents a large effect ($d = 1.67$). *Inversion* also outperforms the *delegation* condition. This difference (13 percentage points) is significant ($p < .001$) and represents a large effect ($d = 1.56$).

Table 2 Summary of p-values for pairwise comparisons of accuracy (Study 1).

	Humans alone	Delegation	Inversion
AI	<0.001	<0.001	<0.001
Humans alone		0.120	<0.001
Delegation			<0.001

Mean accuracies for the *humans alone*, *delegation* and *inversion* conditions are significantly different from *AI alone* ($p < .001$), and except for *inversion*, are all lower than *AI alone*. Performance

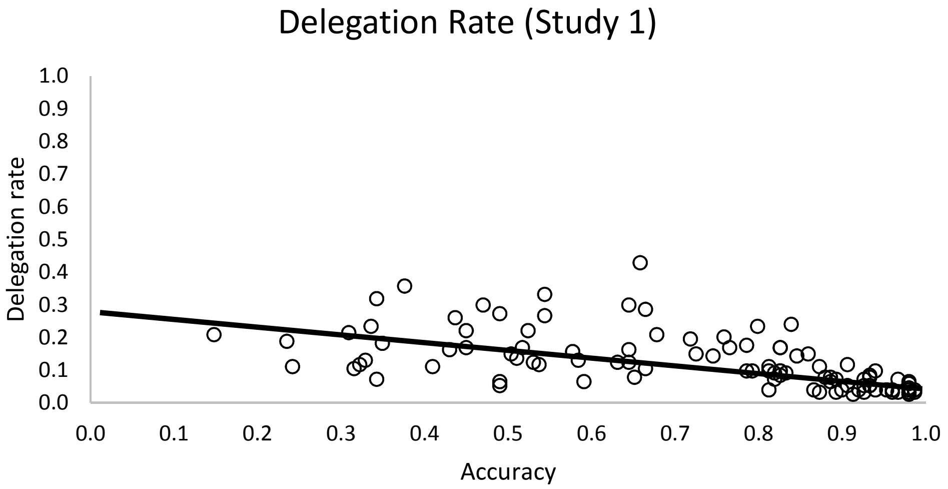

Figure 3 Scatter plot of accuracy per image (horizontal axis) against delegation rate per image (vertical axis). The regression line is estimated from all images.

in the *inversion* condition is significantly higher than when the AI is working alone, suggesting that humans can improve the performance of an AI by providing their input. Not only is *inversion* better than the other settings on average, we also notice that the AI benefits from working with almost all humans. In the inversion condition, only three of the 146 AI-human pairs had a performance smaller than the AI itself; even the 25th accuracy percentile of *inversion* (85.0%) is much larger than AI accuracy. To summarize, sharing work between humans and AI could outperform humans and AI working alone. Inversion was highly effective, but human delegation was not.

To understand inferior human performance in the delegation condition, we investigate how humans delegate. In Figures 3 and 4, image difficulty is depicted on the horizontal axis. Image difficulty is the average accuracy in the humans alone condition of the respective image. A .2 difficulty/accuracy means 20% of subjects classified the image correctly. The vertical axis in both figures shows the delegation rate, i.e., the ratio of subjects who delegated the image to the AI.

If we consider the entire data set (Figure 3), a weak trend can be detected where images with higher accuracy (lower difficulty) are delegated less often and vice versa. Thus, humans seem to “rationally” delegate those images more often, which they are less able to classify correctly. However, if we partition the data into images with less than 70% accuracy (these images are more

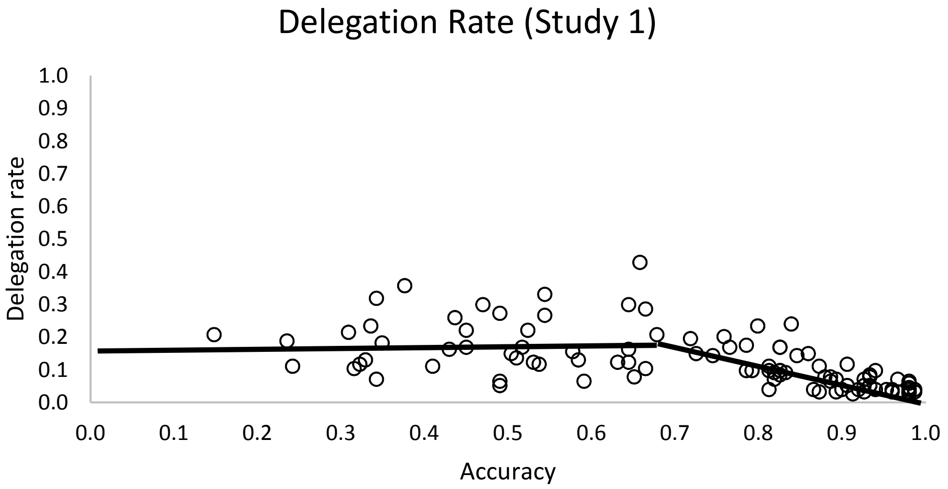

Figure 4 Scatter plot of accuracy per image (horizontal axis) against delegation rate per image (vertical axis). The two regression lines are estimated from two partitions of the data.

difficult than average human performance), and above 70% accuracy (these images are easier than average human performance), the pattern changes (Figure 4): Human delegation is not influenced by the difficulty of an image, if the image is relatively "difficult." Human delegation seems to follow a "rational" pattern if the image was relatively "easy" to classify. Note that we chose the threshold of 70% consistently over the following studies - as no image had an average accuracy between 0.7 and 0.717, setting the threshold to the precise average accuracy of 0.717 does not lead to any changes. Effect sizes and significance levels are in Table 3. Why did humans delegate randomly if images were hard?

In the next section, we present Study 2 that explores possible explanations related to the boundary conditions of successful delegation as potential mechanisms for the observed phenomenon.

3.3. Study 2: Explaining and Enforcing a Delegation Strategy

Hypotheses. This study seeks to analyze the cognitive challenges in human delegation, and aims at providing assistance to explore potential paths towards more productive delegation. In the previous study, we observed random delegation patterns for hard images. Our boundary conditions of successful delegation lead to different possible explanations that we test in this study:

1. Humans might not have a sufficient level of metaknowledge. To be able to analyze this, we asked humans to self-report their level of certainty for each image in all treatments.

Table 3 Regression results for delegation (Study 1).

	Dependent variable: delegation rate	
	$< 70\%$ Accuracy	$\geq 70\%$ Accuracy
Accuracy	0.029	-0.535^{***}
	(0.104)	(0.067)
Constant	0.169^{***}	0.557^{***}
	(0.052)	(0.059)
Observations	41	59
R^2	0.002	0.529
Adjusted R^2	-0.024	0.521
Residual Std. Error	0.090 (df = 39)	0.038 (df = 57)
F Statistic	0.080 (df = 1; 39)	64.138^{***} (df = 1; 57)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses	

2. Humans might not be able to come up with a good delegation strategy: We added a treatment where we advised humans with a strategy that imitates the inversion logic.

3. Humans might not be willing to delegate sufficiently: We added a treatment where delegation was enforced based on human certainty - this treatment applies inversion with humans.

Consequently, Study 2 contains three treatments, all allowing the option to delegate to the AI: A “baseline” condition that replicates the delegation condition of Study 1 asking for self-reported certainty for each image, a “strategy explained” condition, where we suggest a delegation strategy that imitates the delegation logic of inversion based on human certainty, and a “strategy enforced” condition, where we enforce a strategy by automatically delegating images to the AI if a human reports low certainty. As was the case in Study 1, we preregistered two initial hypotheses based on classification accuracy.

Assuming that all three explanations of inferior human delegation performance apply, we state two hypotheses. First, if humans are simply unaware of good delegation strategies, they should improve if such a delegation strategy is suggested. Thus, we pose our first hypothesis:

Hypothesis 2.1: Explaining the delegation strategy leads to a slight improvement in accuracy.

If humans are reluctant to delegate, then enforcing a good strategy should increase accuracy even stronger than just suggesting it, which leads to our second hypothesis:

Hypothesis 2.2: Enforcing the strategy leads to a stronger improvement in accuracy.

In the following, we present the detailed experimental design and the results of Study 2.

Design. In this study we compare three between-subjects conditions. We also asked humans to report their level of certainty for each image on a scale from 1 (uncertain) to 4 (certain) in all conditions. The first condition, "baseline", was set up like the delegation condition of Study 1. For the remaining two conditions, we propose a simple delegation rule similar to that the AI in the inversion condition, where (1) it assessed its classification certainty and (2) delegates to humans if the certainty score was below average human performance. Accordingly, we advised subjects in the second condition, "strategy explained," to delegate images for which they were uncertain (certainty levels between 1 and 3, average accuracy expected to be below the AI performance of 0.77). If subjects' certainty was high (certainty level 4, average accuracy expected to be above the AI performance of 0.77), we advised them to classify the image themselves. In the third condition, "strategy enforced," subjects could not delegate actively. We informed them that images will be delegated automatically if their self-reported certainty was between 1 and 3. The human answer was only considered if the reported score was a 4. This treatment represents most closely the human version of our inversion condition of our first experimental study.

We recruited 453 subjects via MTurk and randomly assigned them to experimental conditions. Average pay was $5.19, average duration was 56.2 minutes. The assignment process and experimental protocol was equivalent to that of Study 1.

Results. Table 5 shows summary statistics for accuracy and delegation rates. Accuracy improved slightly and delegation rates increased strongly when the delegation strategy was explained or enforced. Delegation rates in the strategy enforced condition look similar to the strategy explained condition. This suggests that humans indeed followed the suggested delegation rule.

This is supported by statistical analysis. A Levene test reveals no significant differences between the variances across experimental conditions (F(2, 450)=.849, $p = .429$), but means are different (ANOVA, F(2, 450)=2.97, $p = .052$, $\eta^2 = .13$ which represents a medium effect). All pairwise comparisons are summarized in Table 4. Tukey's significance test shows that humans in the *strategy* *enforced* condition outperform humans in the *baseline* condition. This difference (2.714 percentage points) is significant ($p = .048$) and represents a small to moderate effect (d=.281). Mean accuracy in the *strategy explained* condition is similar to that in the *strategy enforced* condition ($p = .761$). Also, the difference between the *strategy explained* group and the *baseline* group (1.913 percentage points) is not significant ($p = .207$). It would represent a small effect (d=.185). We have then compared the condition's accuracies with AI performance. The *baseline* condition shows a significantly lower performance ($p = .010$), but there is no significant difference between AI and the *strategy explained* ($p = .727$) or the *strategy enforced* condition ($p = .484$). In total, engaging with a good delegation strategy led to more delegation, but accuracy did not increase proportionally. We also compute the accuracy of humans per (self-reported) certainty score. From pre-tests, we expected the accuracy for images with certainty scores between 1 and 3 to be below 0.77, and for images with a certainty score of 4 to be above 0.77. Our results validate this assumption: For Treatment 1, the average accuracies for non-delegated images were 0.43 (certainty score 1), 0.52 (certainty score 2), 0.68 (certainty score 3), and 0.87 (certainty score 4).

Table 4 Summary of p-values for pairwise comparisons of accuracy (Study 2).

	Baseline	Strategy explained	Strategy enforced
AI	0.010	0.727	0.484
Baseline		0.207	0.048
Strategy explained			0.761

Figure 5 shows the delegation pattern. The horizontal axis depicts image difficulty (average human accuracy of Treatment 1, Study 1), the vertical axis delegation rates. The baseline condition replicated the results of Study 1. Further, humans delegated more when the strategy was explained or enforced. However, their behavior for difficult images was still random. The randomness just centered around a higher average than in the baseline condition. Therefore, knowing a good delegation strategy did not prohibit random delegation of difficult images. Hence, we can rule out the second explanation one from above: while providing a strategy helps to increase delegation, it

Table 5 Summary statistics for accuracy and delegation rate (Study 2).

		Summary statistic							
Dep. Var.:	Treatment	N	Min.	Mean	Max.	St. Dev.	Pctl(25)	Median	Pctl(75)
Accuracy									
	Baseline	150	0.160	0.748	0.900	0.104	0.720	0.770	0.810
	Strategy explained	157	0.240	0.767	0.880	0.103	0.750	0.800	0.825
	Strategy enforced	146	0.140	0.775	0.900	0.088	0.750	0.790	0.823
Delegation rate									
	Baseline	150	0.000	0.131	0.680	0.151	0.010	0.080	0.200
	Strategy explained	157	0.000	0.342	0.950	0.203	0.185	0.330	0.475
	Strategy enforced	146	0.010	0.335	0.960	0.183	0.190	0.315	0.463

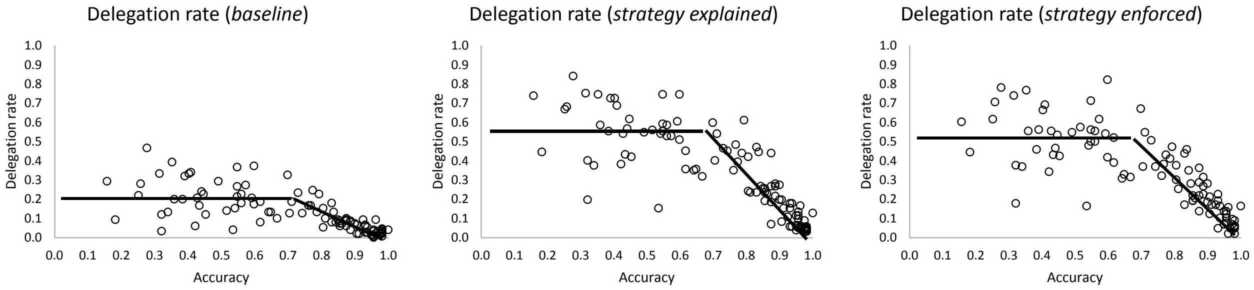

Figure 5 Scatter plots of accuracy against delegation rate per image and experimental conditions (Study 2). The two regression lines are estimated from the two partitions of the data.

Table 6 Regressions per experimental condition (Study 2). The dependent variable is the images' delegation rate. The data is partitioned into two regions.

	Experimental condition					
	Baseline		Strategy explained		Strategy enforced	
	Dependent variable: Delegation rate for images with accuracy of					
	$< 70\%$	$\geq 70\%$	$< 70\%$	$\geq 70\%$	$< 70\%$	$\geq 70\%$
Accuracy	-0.041	-0.588***	-0.230	-1.610***	-0.131	-1.486***
	(0.121)	(0.072)	(0.182)	(0.165)	(0.180)	(0.142)
Constant	0.233***	0.592***	0.650***	1.626***	0.579***	1.520***
	(0.060)	(0.064)	(0.090)	(0.146)	(0.090)	(0.126)
Observations	41	59	41	59	41	59
R^2	0.003	0.537	0.039	0.625	0.013	0.657
Adjusted R^2	-0.023	0.529	0.015	0.619	-0.012	0.651
Residual Sd. Error	0.105	0.041	0.158	0.094	0.157	0.081
F Statistic	0.117	66.08***	1.59	95.14***	0.531	109.30***

Note: * p< .1; ** p< .05; *** p< .01
Standard errors in parentheses

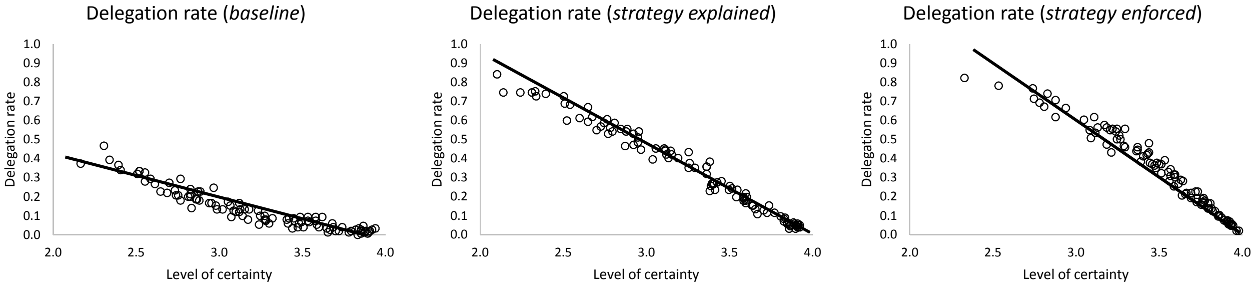

Figure 6 Scatter plots of certainty against delegation rate per image and experimental conditions (Study 2).

could not fix the random delegation pattern for difficult images. Table 6 shows the corresponding statistical results.

We now address explanation three (humans do not want to use the AI). We analyzed how delegation rates changed with perceived image difficulty (i.e., self-assessed certainty). Figure 6 plots delegation rates (vertical axis) against self-assessed certainty (horizontal axis). The figure suggests that humans delegated with great internal consistency. Images they perceived as more difficult were delegated more often. This was true, independent of whether they knew the delegation strategy or not. The subjects appeared to be aiming for a consistent delegation pattern. Once they learned a good delegation strategy, delegation rates more than doubled. Therefore, we conclude that in our experiments humans did not show reluctance towards using the AI.

In light of these findings, the first explanation seems likely. Humans might not be able to judge the difficulty of images when the images are hard. They may thus not be able to use the AI systematically for these images, a problem associated with lack of metaknowledge.

To explore this explanation, we study how well humans can assess their own ability to classify images. In Figure 7 we plot the average self-assessed certainty of an image (vertical axis) against the average accuracy of the image (horizontal axis). The visual impression and the regression results in Table 7 suggest that humans can assess their ability for relatively easy images (accuracy above 70%), but they can not assess it for difficult images. An interesting side finding can be observed for the *strategy enforced* condition. Here, the constant of the regression model is positive and significant for easy images. Thus, objective difficulty explains perceived difficulty (as in the

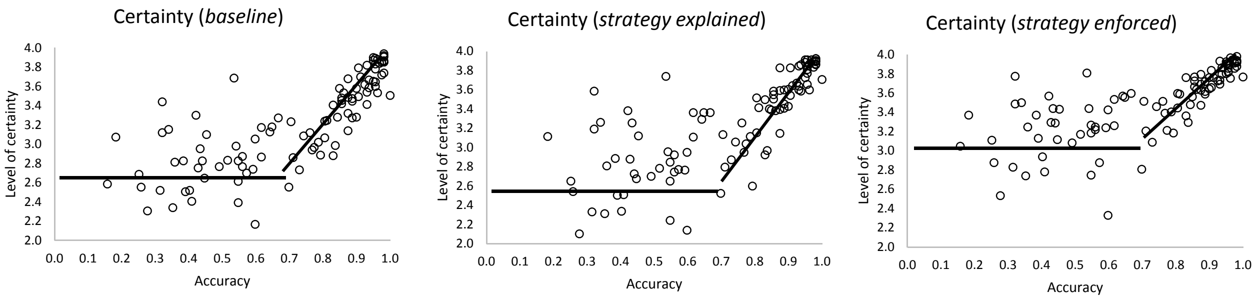

Figure 7 Scatter plots of accuracy against certainty per image and experimental conditions (Study 2).

other conditions), but subjects seem to report higher certainty values independent of actual image difficulty. A possible explanation is that subjects avoided automated delegation by (mis)reporting certainty values of four.

Table 7 Regressions per experimental condition (Study 2). The dependent variable is the subjects' certainty per image. The data is partitioned into two regions.

	Experimental condition					
	Baseline		Strategy explained		Strategy enforced	
	Dependent variable: Certainty, where images have accuracy of...					
	$< 70\%$	$\geq 70\%$	$< 70\%$	$\geq 70\%$	$< 70\%$	$\geq 70\%$
Accuracy	0.249	3.524^{***}	0.368	3.505^{***}	0.109	2.363^{***}
	(0.384)	(0.344)	(0.465)	(0.374)	(0.384)	(0.238)
Constant	2.716^{***}	0.383	2.675^{***}	0.453	3.161^{***}	1.608^{***}
	(0.191)	(0.304)	(0.231)	(0.332)	(0.191)	(0.211)
Observations	41	59	41	59	41	59
R^2	0.011	0.648	0.016	0.606	0.002	0.633
Adjusted R^2	-0.015	0.642	-0.009	0.599	-0.024	0.626
Residual Sd. Error	0.334	0.196	0.403	0.214	0.334	0.136
F Statistic	0.419	105.10^{***}	0.626	87.64^{***}	0.081	98.25^{***}

Note: * p< .1; ** p< .05; *** p< .01
Standard errors in parentheses

Therefore, while humans delegate quite rationally based on their internal assessment (Figure 6), this assessment is not precise for relatively difficult tasks (Figure 7). Put differently, although human delegation decisions are often misaligned with real problem difficulty, they are not misaligned with their perceived problem difficulty. We conclude that lack of metaknowledge seems to drive the inferior delegations. According to this explanation, humans did not know what they knew and delegated the wrong images to the AI.

4. Robustness Checks

In this section, we test robustness of our findings with two additional studies. In the first robustness check, we analyze whether continuous feedback on both human and AI performance has an impact on delegation behavior and human metaknowledge. In the second robustness check, we test whether an AI could realize complementarities with humans, even in cases, where the tasks are more difficult than those the AI was trained with. We manipulate task difficulty by scaling the images to a lower resolution, and test the effectiveness of inversion.

4.1. Study 3: The Role of Feedback

Purpose. Study 3 relaxes the assumption of receiving no feedback on task results to analyze the effects of feedback on human delegation behavior and on human metaknowledge.

At the outset, we would like to point out that the potential effects are unclear, ex-ante, and the literature provides no clear direction. We lay out possible effects in the following discussion. First, metaknowledge might be increased by providing continuous feedback because it could improve the human perception with regards to their own performance. However, we observe that even long-term experience does not seem to prevent poor metaknowledge (Brezis et al. 2018). Thus, the effect remains unclear. Second, feedback on human and AI performance might increase salience of AI superiority, and could consequently lead to higher delegation rates. On the other side, Dietvorst et al. (2015) demonstrated that humans relied less on algorithmic advice after seeing it err, even if the algorithmic performance was superior. Thus, we do not state any directional hypotheses in this study. In the following, we lay out the details of our study design before presenting our results.

Design. We compare classification accuracy and delegation rate between two conditions. The "baseline" condition (1) replicates the "delegation" condition of Study 1 and the "baseline" condition of Study 2. In the "feedback" condition (2), subjects received feedback after each classification task, consisting of their own answer, the AI answer, and the correct answer. We ran a between-subject design with 289 subjects in February 2021, and randomly assigned subjects to the baseline condition (148 subjects) and the feedback condition (141 subjects).

All subjects received instructions, had to pass a short quiz so that we could exclude robots, and completed an example classification to ensure they understood the task. They then had to classify the 100 images in random order. Each subject received a base fee of $2 , and an additional 5 cents for each correct answer. Afterwards, they were asked how many images they think they classified correctly. They could earn 1 additional dollar if this estimation did not differ from the actual number by more than five images. Average pay was $4.92, slightly above average pay on MTurk in general (Hara et al. 2018). The average duration of the experiment was 62.7 minutes.

Results. Descriptive statistics (Table 8) show little difference among the the experimental conditions, regarding both accuracy (*baseline*: 53.4% *feedback*: 54.4%) and delegation rate (*baseline*: 12.0%, *feedback*: 12.4%). Accuracy does not significantly differ in the *baseline* and the *feedback* conditions ($p = .697$), neither does the delegation rate ($p = .860$). Thus, there is no indication that continuous feedback on human and AI performance affects human delegation behavior. Note that in line with literature on MTurk performance during COVID-19, the average accuracy values are below those of our previous studies, while delegation rates remain similar.

Table 8 Summary statistics for accuracy and delegation rate (Study 3).

		Summary statistic							
Dep. Var.:	Treatment	N	Min.	Mean	Max.	St. Dev.	Pctl(25)	Median	Pctl(75)
Accuracy									
	Baseline	149	0.070	0.534	0.880	0.221	0.373	0.540	0.758
	Feedback	141	0.070	0.544	0.950	0.232	0.340	0.620	0.750
Delegation rate									
	Baseline	149	0.000	0.120	1.000	0.231	0.000	0.010	0.120
	Feedback	141	0.000	0.124	0.990	0.212	0.000	0.010	0.155

Next, we analyze the effect of continuous feedback on metaknowledge by replicating the analysis of the relationship between accuracy and certainty in Study 2. We illustrate the relationship of delegation rate and difficulty (average human accuracy of Treatment 1, Study 1) in Figure 8. We further summarize the regression results in Table 9: as in Study 2, we only see a significant influence of "difficulty" (measured by the average accuracy of each image) for images with an

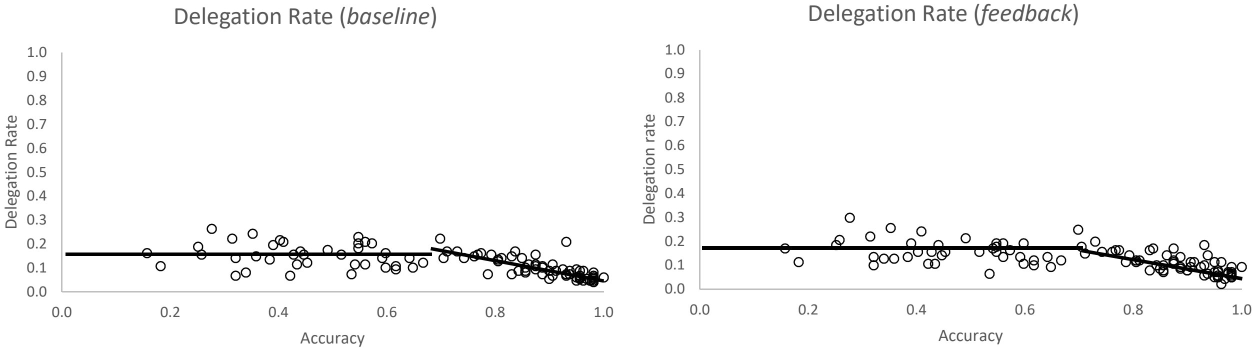

Figure 8 **Scatter plots of accuracy against certainty per image and per experimental conditions (Study 3). The two regression lines are estimated from the two partitions of the data.**

average accuracy of at least 70%. Note that we use the average accuracy of the first treatment of Study 1, where humans classified without AI delegation, to maintain a consistent definition of "easy" and "difficult" tasks.

Table 9 **Regressions per experimental condition (Study 3). The dependent variable is the subjects' certainty per image. The data is partitioned into two regions.**

	Experimental condition			
	Baseline		Feedback	
	DV: Certainty, images accuracy of...			
	$< 70\%$	$\geq 70\%$	$< 70\%$	$\geq 70\%$
Accuracy	0.099	1.527***	0.182	1.372***
	(0.179)	(0.238)	(0.134)	(0.205)
Constant	3.130***	2.067 ***	3.083***	2.148
	(0.089)	(0.210)	(0.067)	(0.205)
Observations	41	59	41	59
R^2	0.008	0.420	0.045	0.439
Adjusted R^2	-0.018	0.410	0.021	0.430
Residual Sd. Error	0.156	0.136	0.116	0.117
F Statistic	0.305	44.33***	1.09	44.67***

Note: * p< .1; ** p< .05; *** p< .01
Standard errors in parentheses

In the next section, we present a robustness check on the impact of more difficult tasks on the efficiency of delegation and inversion.

4.2. Study 4: The Role of Difficulty

Purpose. Inversion was the most effective condition in our first experimental study. The key of its success was the AI's ability of assessing its own quality. Study 4 confronts the AI with tasks that are more difficult than those it was trained with. In the case of image classification, this could relate to images with a lower resolution. Thus, we replicate Study 1 with a higher task difficulty by applying a lower resolution to all images. We aim to analyze whether the AI would still be able to delegate efficiently.

Design. We compare classification accuracy between two conditions, "humans alone" (1) and "delegation" (2). Those conditions mirror the first two conditions of Study 1. We further use condition (1) to simulate different inversion strategies. We ran a between subject design with 299 subjects in January 2021 and randomly assigned subjects to the "humans alone" condition (150 subjects) and the "delegation" condition (148 subjects).

All subjects received instructions, had to pass a short quiz so that we could exclude robots, and completed an example classification to ensure they understood the task. They then had to classify the 100 images in random order. Each subject received a base fee of $2, and an additional 5 cents for each correct answer. Afterwards, they were asked how many images they think they classified correctly. They could earn 1 additional dollar if this estimation did not differ from the actual number by more than five images. Average pay was $4.61, slightly above average pay on MTurk in general (Hara et al. 2018). The average duration of the experiment was 65.9 minutes.

Table 10 Summary statistics for accuracy (Study 4).

		Summary statistic							
Dep. Var.:	Treatment	N	Min.	Mean	Max.	St. Dev.	Pctl(25)	Median	Pctl(75)
Accuracy									
	Humans alone	150	0.120	0.481	0.790	0.168	0.360	0.480	0.633
	Delegation	148	0.140	0.512	0.790	0.153	0.420	0.510	0.630

Results. Humans slightly improve by about three percentage points with the possibility to delegate ($p = .093$), even though on average, only 7.6% of images were delegated. While no direct

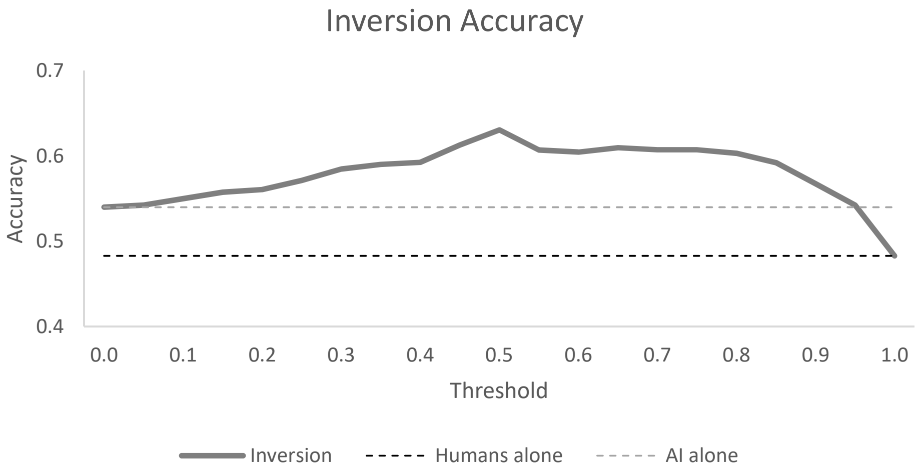

Figure 9 **Inversion accuracy by threshold value.**

comparison is possible, we see that the human performance does not seem to be strongly affected by lower resolution, as the total accuracy seems to be similar to those of Study 3. AI performance, however, was decreased significantly from 77% to 54%. Human performance still remains below AI accuracy. Next, we analyze whether the AI is still able to improve by delegating to humans, even though its own performance dropped strongly. In Figure 9, we simulate inversion accuracies based on the *humans alone* condition with varying threshold values. A threshold of 0 symbolizes always choosing the AI prediction resulting in the AI accuracy of 54%, while a threshold of 1 symbolizes always choosing the human prediction resulting in the average human accuracy of 48.1%. With every threshold value below 0.95, the inversion accuracy outperforms both human and AI accuracy, with a maximum value at a threshold of 0.50, close to the average human accuracy. Using an inversion delegation rule with a threshold of 0.50 improves the human accuracy by 15 percentage points, and the AI accuracy by nine percentage points. Thus, even in a situation where the AI performs relatively poor, inversion seems to be a powerful delegation mechanism.

5. Discussion and Directions of Future Research

Our results demonstrate that humans and AI can work together on image classification, even if there is no feedback about the AI performance and errors. In such a situation, it is beneficial to let the AI delegate work to humans in case the AI is uncertain. Humans were unable to delegate well. We claim that the reason for their wrong delegation decisions is that the human subjects were unable to assess if they know the correct class of difficult images. This result is interesting because it shows inferior human performance, but no aversion to use the algorithm. Our data supports the explanation that subjects were indeed motivated to work with the machine, and were willing to follow rational delegation strategies. They were unable to execute those due to their wrong perception of task difficulty. We interpret this as a fundamental and latent limitation rather as an act of conscious reluctance. In this regard, our results are consistent with the general view issued in Logg et al. (2019): humans do appreciate the help from AI. But we also show that they might still have problems working with it.

Further, our results challenge the assumption that an entire task should be handed over to an AI if the AI is better. We stated three three boundary conditions where delegation and a good distribution of work can outperform the assignment to one party. First, humans and AI have to have complementary skills. We claim this should be the case for tasks where decision rules are not clearly defined. We confirm this using image classification as an example: An optimal combination of the AI and humans from the inversion condition would lead to an accuracy of 89.9%, considerably more than 77% accuracy for AI alone and 71.7% for humans alone. Second, complementarities have to be recognized. We define a sufficient level of metaknowledge as a necessary condition. While the AI seems to have a good perception of own abilities, humans are not able to differentiate between tasks they are able to do and those where this is not the case, especially for difficult images. Third, an efficient delegation rule needs to be followed, where a task is moved to the actor that is better at solving it. Under perfect information, a simple rule is effective: If you are able to do the task, do it yourself; if you are not, then delegate. We demonstrate that such a rule can easily be implemented for AI, and that humans can potentially be trained to follow such a rule.

When AI delegates to humans: Inversion. If AI would be responsible to delegate to humans, several interesting things could happen. First, in our experiment the resulting performance was higher than that of the AI alone. This makes inversion economically desirable. Second, humans would do some of the work. They contribute to the superior result, without them we would not reach it. Inversion might also improve human work perspectives. Humans are more motivated when working in a stimulating environment (Pink 2011). In our example, classifying easily identifiable images is perhaps routine and boring, whereas the classification of difficult images could be an interesting challenge. Inversion might enable humans to spend less time on mundane tasks and more time on challenging tasks, thereby creating a more fulfilling workplace. Thus, receiving assignments from a machine could be interpreted not only as a delegation to humans, but also as freeing humans some valuable time. The AI would not be the humans' boss, but rather an assistant who swipes away distractions from the real work. Hoewever, inversion comes with a loss of human control. The AI decides about the delegations, it asks the human for support only if it is required. It does this without emotions, only to leverage complementarities that exist as foreseen by Polanyi's pathbreaking work.

When humans delegate to AI: Metaknowledge and the quest for good delegation. Our research points to a fundamental characteristic of human behavior that needs to be understood in order to design more effective human-AI collaborative environments: Humans did not perform well in delegating tasks to the AI. We can design and teach simple delegation rules, especially in modularized tasks. However, even when humans diligently and rationally apply a delegation rule (Figure 5) that is internally consistent with their perception of task difficulty, humans that delegate to AI do not perform as well as they should. The reason for this (Table 9) is an apparent lack of understanding what is difficult for them, and what is not. This phenomenon is not isolated to working with AI or computers. Humans tend to misjudge their certainty when dealing with high difficulty questions as compared to medium-difficulty or easier questions (Pulford and Colman 1997). In our research context, this translates into humans making more arbitrary delegation decisions when dealing with difficult tasks, which worsens their overall performance.

More generally, the phenomenon to not understand the difficulty of a task at hand relates to a lack of metaknowledge in terms of "appreciation of what we do know and what we do not know"

(Russo and Schoemaker 1992, p.8). In our formal education, we do not emphasize this higher level of learning to recognize our own strengths and shortcomings. Often, the impact of lacking metaknowledge when facing difficult issues is mitigated by team work, when other human group members point out alternatives, logical and procedural inconsistencies and/or errors. One way to interpret group discussions is to try and achieve the best compromise based on metaknowledge and primary knowledge based on facts, concepts, models, relationships and solution techniques. However, humans working with AI are unlikely to be in an environment where they can reason with AI, or more specifically where AI engages humans in a dialogue to resolve issues related to metaknowledge. If we want to produce students that will be effective in the future workforce, improving metaknowledge should be a central tenet of higher education.

Limitations. Our study informs on delegation between humans and AI. To ensure a certain degree of generalizability of results, we aimed for a generic, non-specialized task and non-specialized workers relying on image classification and MTurk workers. While we do believe that our findings should carry over to many other settings, there might be additional effects in any specialized environment, that strengthen or weaken our findings. While relying on non-specialized situations is a limitation of this study, it also creates and opportunity for future research, and to test whether our findings can be replicated in specialized environments.

There has been a lot of discussion on the suitability of MTurk workers for behavioral experiments. These discussions concentrate on three main criticisms of using MTurk in behavioral experiments: First, non-naivete, that is, subjects might be experienced in similar experiments and behave strategically (Chandler et al. 2019). Second, carelessness, that is, subjects act with a lower degree of rigor leading to noisy and partially inconsistent results (Aruguete et al. 2019). Third, representativeness, that is, the MTurk population does not reflect the composition of society in general. However, it should be noted that MTurk subjects are better representatives of the general population compared to typical student subjects used in a large number of academic studies (Chandler et al. 2019).

How could we safeguard against these issues? We contend that non-naivete does not apply in this instance since our study is unique in character, compared to, for example, potentially hundreds of studies looking at newsvendor problems or dictator games. A potential solution could be to rely on new MTurk workers (Robinson et al. 2019), or to exclude workers who participated in similar studies if this information is available. We excluded all subjects who participated in our own related studies or pre-tests. In terms of carelessness of MTurk workers, Aruguete et al. (2019) show some evidence of carelessness in terms of a higher spread in data quality in MTurk samples compared to traditional student samples and recommend measures to ensure validity of results. Following these suggestions, we decided to: a) restrict our subject pool to subjects with a positive track record and at least 90% positive reviews; and b) included an attention check and a classification exercise that had to be passed without errors in order to participate in the study. Please note that for our set of robustness experiments (Study 3 and 4), we had to conduct the experiments during the COVID-19 pandemic. This led to an increased level of subjects' carelessness and lower performance compared to the other studies. This finding is in line with the literature (Arechar and Rand 2021), and we refrain from direct comparisons of the specific results between the first set of experiments and the robustness checks. We admit that our study does not claim to represent a general population. Thus, we do not make any claims regarding absolute results of our study, such as "we expect humans to delegate 13% of tasks to an AI," rather we compare differences in behavior between conditions. Replicating several studies from different subject samples with MTurk samples, Coppock (2019) conclude that MTurk samples can be compared to other national samples. Many other studies validate the appropriateness of MTurk samples for experimental studies in social sciences, such as Buhrmester et al. (2016), Horton et al. (2011), or Lee et al. (2018).

Future research. As laid out above, a potentially relevant limitation of our sample lies in an expected low performance, especially for the samples drawn during the COVID-19 pandemic. In concert with focusing on non-specialized tasks, this limits the generalizability of the results regarding the absolute performance of our experiment. While we do not expect that those limitations have affected our main findings regarding different configurations of delegation schemes, or the mechanisms we observed, we believe that analyzing similar settings with high-impact decisions and dedicated workers is a fruitful avenue for future studies.

In addition to addressing potential limitations of our study, a key research area should be about making humans better delegators in order to develop effective human-AI collaborative environments. This requires research on three fronts:

a) Research on human-computer dialogue and decision authority. How should an AI engine communicate and adapt when working with humans that have different levels of metaknowledge, and how should it develop an appropriate framework for decision making in these environments? For example, an AI engine can delegate decision authority to individuals with high levels of metaknowledge, whereas it may simply receive inputs from highly competent individuals lacking metaknowledge.

b) Research on system feedback to increase metaknowledge. Prior research (Pulford and Colman 1997) has shown that feedback may not affect metaknowledge, especially when the task is difficult. No concerted effort has been made to design feedback environments that lead to improved individual metaknowledge when other options are available.

c) Research on improving metaknowledge. Laboratory studies have shown that experience only partially impacts metaknowledge (Hansson et al. 2008), and our robustness check in Study 3 showed no effect of providing continuous feedback on metaknowledge. We still cannot rule out that long-term debriefing, for example as it is common with airline pilots (Kikkawa and Mavin 2017), might improve metaknowledge by providing a better understanding of our own strengths, weaknesses and boundaries. This may lead to better appreciation of alternative sources that can help in decisions. In human-only environments, providing long-term feedback and intensive debriefing is costly. Human feedback may show internal consistency problems and may be intrusive at the task level. However, modern technology, including realistic simulations (see Ketter et al. (2016) as an example), can potentially provide innovative solutions that help improve our metaknowledge.

Acknowledgments

References

Agrawal A, Gans J, Goldfarb A (2018) *Prediction machines: The simple economics of artificial intelligence* (Boston, Massachusetts: Harvard Business Review Press), ISBN 9781633695672.

Arechar AA, Rand DG (2021) Turking in the time of covid. *Behavior Research Methods* 1–5.

Aruguete MS, Huynh H, Browne BL, Jurs B, Flint E, McCutcheon LE (2019) How serious is the 'carelessness' problem on mechanical turk? *International Journal of Social Research Methodology* 22(5):441–449.

Autor DH (2014) Polanyis' paradox and the shape of employment growth. *Proceedings of the Federal Reserve Bank of Kansas City - Jackson Hole Economic Policy Symposium* (August).

Autor DH, Levy F, Murnane RJ (2003) The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics* 118(4):1279–1333, URL `http://dx.doi.org/10.1162/003355303322552801`.

Baird A, Maruping LM (2021) The next generation of research on is use: A theoretical framework of delegation to and from agentic is artifacts. *MIS Quarterly* 45(1):315–341.

Bazerman MH (1985) Norms of distributive justice in interest arbitration. *Industrial and Labor Relations Review* 38(4):558, ISSN 00197939, URL `http://dx.doi.org/10.2307/2523991`.

Brezis M, Orkin-Bedolach Y, Fink D, Kiderman A (2018) Does physician's training induce overconfidence that hampers disclosing errors? *Journal of Patient Safety* forthcoming.

Brynjolfsson E, Mitchell T, Rock D (2018) What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings* 108:43–47, ISSN 2574-0768, URL `http://dx.doi.org/10.1257/pandp.20181019`.

Brynjolfsson E, Mitchell TM (2017) What can machine learning do? workforce implications. *Science* 358(6370):1530–1534, URL `http://dx.doi.org/10.1126/science.aap8062`.

Buhrmester M, Kwang T, Gosling SD (2016) Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? *American Psychological Association* .

Chandler J, Rosenzweig C, Moss AJ, Robinson J, Litman L (2019) Online panels in social science research: Expanding sampling methods beyond mechanical turk. *Behavior Research Methods* 51(5):2022–2038.

Coppock A (2019) Generalizing from survey experiments conducted on mechanical turk: A replication approach. *Political Science Research and Methods* 7(3):613–628.

Dawes RM (1979) The robust beauty of improper linear models in decision making. *American Psychologist* 34(7):571–582, URL `http://dx.doi.org/10.1037/0003-066X.34.7.571`.

Deng L, Yu D (2013) Deep learning: Methods and applications. *Foundations and Trends in Signal Processing* 7(3-4):197–387, URL `http://dx.doi.org/10.1561/2000000039`.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of experimental psychology. General* 144(1):114–126, URL `http://dx.doi.org/10.1037/xge0000033`.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170, ISSN 0025-1909, URL `http://dx.doi.org/10.1287/mnsc.2016.2643`.

Difallah DE, Catasta M, Demartini G, Ipeirotis PG, Cudré-Mauroux P (2015) The dynamics of micro-task crowdsourcing: The case of amazon mturk. *Proceedings of the 24th international conference on world wide web*, 238–247.

Dijkstra JJ (1999) User agreement with incorrect expert system advice. *Behaviour & Information Technology* 18(6):399–411, URL `http://dx.doi.org/10.1080/014492999118832`.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115–118, URL `http://dx.doi.org/10.1038/nature21056`.

Evans JA, Foster JG (2011) Metaknowledge. *Science* 331(6018):721–725, URL `http://dx.doi.org/10.1126/science.1201765`.

Foster ED, Deardorff A (2017) Open science framework (osf). *Journal of the Medical Library Association* 105(2), URL `http://dx.doi.org/10.5195/jmla.2017.88`.

Ge R, Zheng Z, Tian X, Liao L (2021) Human–robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Information Systems Research* Forthcoming.

Hansson P, Juslin P, Winman A (2008) The role of short-term memory capacity and task experience for overconfidence in judgment under uncertainty. *Journal of Experimental Psychology: General* 34(5):1027–1042, ISSN 1939-2222.

Hara K, Adams A, Milland K, Savage S, Callison-Burch C, Bigham JP (2018) A data-driven analysis of workers' earnings on amazon mechanical turk. Mandryk R, Hancock M, Perry M, Cox A, eds., *Proceed-*

ings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, 1–14 (New York, New York, USA: ACM Press), ISBN 9781450356206, URL `http://dx.doi.org/10.1145/3173574.3174023`.

Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Kingsbury B, Sainath T (2012) Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine* 29:82–97, URL `https://www.microsoft.com/en-us/research/publication/deep-neural-networks-for-acoustic-modeling-in-speech-recognition/`.

Horton JJ, Rand DG, Zeckhauser RJ (2011) The online laboratory: Conducting experiments in a real labor market. *Experimental economics* 14(3):399–425.

Huber GP (1990) A theory of the effects of advanced information technologies on organizational design, intelligence, and decision making. *Academy of Management Review* 15(1):47–71.

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* Forthcoming.

Ketter W, Peters M, Collins J, Gupta A (2016) A multiagent competitive gaming platform to address societal challenges. *MIS Quarterly* 40(2):447–460, ISSN 02767783, URL `http://dx.doi.org/10.25300/MISQ/2016/40.2.09`.

Kikkawa Y, Mavin TJ (2017) A review of debriefing practices: Toward a framework for airline pilot debriefing. *Aviation Psychology and Applied Human Factors* 7(1):42.

Kleinmuntz B (1990) Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological bulletin* 107(3):296–310, ISSN 0033-2909, URL `http://dx.doi.org/10.1037/0033-2909.107.3.296`.

Kononenko I (2001) Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine* 23(1):89–109, ISSN 09333657.

Kremer M, Moritz B, Siemsen E (2011) Demand forecasting behavior: System neglect and change detection. *Management Science* 57(10):1827–1843.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.

Lee YS, Seo YW, Siemsen E (2018) Running behavioral operations experiments using amazon's mechanical turk. *Production and Operations Management* 27(5):973–989.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103, ISSN 07495978, URL http://dx.doi.org/10.1016/j.obhdp.2018.12.005.

Lories G, Dardenne B, Yzerbyt V (1998) From social cognition to metacognition. Yzerbyt V, Lories G, Dardenne B, eds., *Metacognition: Cognitive and Social Dimensions*, 1–15 (SAGE Publications Ltd), ISBN 9780761952596, URL http://dx.doi.org/10.4135/9781446279212.n1.

McAfee A (2013) Big data's biggest challenge? convincing people not to trust their judgement. *Harvard Business Review* URL https://hbr.org/2013/12/big-datas-biggest-challenge-convincing-people-not-to-trust-their-judgment.

McAfee A, Brynjolfsson E (2017) *Machine, platform, crowd: Harnessing our digital future* (New York and London: W.W. Norton & Company), first edition edition, ISBN 978-0-393-25429-7.

Meehl PE (1954) *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* (University of Minnesota Press).

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT (2018) The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America* 115(11):2600–2606, ISSN 0027-8424, URL http://dx.doi.org/10.1073/pnas.1708274114.

Pink DH (2011) *Drive: The surprising truth about what motivates us* (New York: Riverhead Books), first paperback edition edition, ISBN 1594484805.

Pulford BD, Colman AM (1997) Overconfidence: Feedback and item difficulty effects. *Personal and Individual Differences* 23(1).

Robinson J, Rosenzweig C, Moss AJ, Litman L (2019) Tapped out or barely tapped? recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. *PloS One* 14(12):e0226394.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252, URL `http://arxiv.org/pdf/1409.0575v3`.

Russo JE, Schoemaker PJH (1992) Managing overconfidence. *Sloan Management Review* (Winter):7–17.

Schanke S, Burtch G, Ray G (2021) Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research* Forthcoming.

Schmidhuber J (2015) Deep learning in neural networks: An overview. *Neural networks : the official journal of the International Neural Network Society* 61:85–117.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9 (IEEE), ISBN 978-1-4673-6964-0, URL `http://dx.doi.org/10.1109/CVPR.2015.7298594`.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Zanzotto FM (2019) Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research* 64:243–252.