Collaboration and Delegation Between Humans and AI: An Experimental Investigation of the Future of Work

Andreas Fügener

University of Cologne, andreas.fuegener@wiso.uni-koeln.de

Jörn Grahl

University of Cologne, grahl@wiso.uni-koeln.de

Alok Gupta

University of Minnesota, gupta037@umn.edu

https://carlsonschool.umn.edu/faculty/alok-gupta

Wolfgang Ketter

University of Cologne, Erasmus University Rotterdam,

ketter@wiso.uni-koeln.de, wketter@rsm.nl

April 8, 2019

Abstract

A defining question of our age is how AI will influence the workplace of the future and, thereby, the human condition. The dominant perspective is that the competition between AI and humans will be won by either humans or machines. We argue that the future workplace may not belong exclusively to humans or machines. Instead, it is better to use AI together with humans by combining their unique characteristics and abilities.

In three experimental studies, we let humans and a state of the art AI classify images alone and together. As expected, the AI outperforms humans. Humans could improve by delegating to the AI, but this combined effort still does not outperform AI itself. The most effective scenario was inversion, where the AI delegated to a human when it was uncertain. Humans could in theory outperform all other configurations if they delegated effectively to the AI, but they did not. Human delegation suffered from wrong self-assessment and lack of strategy. We show that humans are even bad at delegating if they put effort in delegating well; the reason being that despite their best intentions, their perception of task difficulty is often not aligned with the real task difficulty if the image is hard. Humans did not know what they did not know. Because of this, they do not delegate the right images to the AI. This result is novel and important for human-AI collaboration at the workplace. We believe it has broad implications for the future of work, the design of decision support systems, and management education in the age of AI.

Keywords: Future of work, Artificial Intelligence, Augmented Decision Environments, Deep Learning, Human-AI Collaboration, Machine Learning, Intelligent Software Agents

1 Introduction and prior work

There are huge expectations, and there is a lot of uncertainty, about how AI will change the workplace. Thought leaders, scientists and policy makers have come to see it as a general purpose technology. Andrew Ng, co-founder of Coursera, calls AI a new type of electricity (Knowledge@Wharton 2017) that fuels innovation on a broad scale and in diverse domains, including medicine (Kononenko 2001, Peek et al. 2015, Esteva et al. 2017, Hosny et al. 2018), transportation (Chen et al. 2015, Kahlen et al. 2018), trading markets (Bichler et al. 2010), problem-solving, games and cognition (Schölkopf 2015, Silver et al. 2016, Moravčík et al. 2017), and perceptional tasks, such as processing images, text, and speech (Hinton et al. 2012, Deng and Yu 2013). The broad applicability of modern techniques like deep learning (LeCun et al. 2015, Schmidhuber 2015, Goodfellow et al. 2016) seems to imply a slow but sure redundancy of human input in the future. In fact, the superiority of algorithmic results over human decision making is not novel in certain fields (Grove et al. 2000, Russakovsky et al. 2014). With AI becoming mainstream, a continual discussion is being held in the public sphere about how AI will influence the future of work, and thereby, the human condition (Frey and Osborne 2017, Williams 2017, Marlin 2018, Gray 2018).

We, of course, agree with the general sentiment that AI will be embedded in day-to-day life in the future, and that the performance of AI will improve further. This has an obvious and important implication: AI will outperform humans in even more cognitive and perceptional tasks. Therefore, research efforts that only consider direct comparisons between humans and AI will rarely provide surprising answers. Put differently, if we approach questions about employment with an "us versus them" mindset and ask questions like: "will the AI outperform humans in this task, or will humans outperform the AI?", the answer will ever more often be that the AI outperforms humans. This line of arguments supports a grim outlook on employment, and it emphasizes the frictions that can arise from accepting and adopting AI.

Contrary to this, computer scientists have recognized early on that humans and AI can benefit from each other. They have coupled AI and humans in various ways, as parts of technical systems that outperform the AI alone, solve fuzzier problems or process unstructured data (Maes 1994, Joshi et al. 2009, Branson et al. 2010, Nagar and Malone 2011, Holzinger 2013, Attenberg et al. 2015, Russakovsky et al. 2015, Wang et al. 2017). For example, image classification algorithms can put "humans in the loop" by asking them questions about the content of an image and modeling on the basis of their answers (Branson et al. 2010). Other approaches ask humans for labels if instances are considered hard (Joshi et al. 2009), or use human input to identify extraordinary cases that the computational model gets wrong consistently (Attenberg et al. 2015). The Re-Captcha project digitizes books using human input to the AI, and it operates at scale since 2007.

Researchers without formal training in AI or machine learning, who are not trying to build technical systems and who do not design algorithms, are more hesitant to study human-AI collaboration. Of course there are important exceptions, like the works by Dietvorst et al. (2016) and Logg et al. (2018), who study the human decision makers' attitude toward algorithms (both the terms algorithm aversion and algorithm appreciation have been proposed). But our overall impression is that human-AI collaboration is currently under-researched in Information Systems, the social sciences and management. This is rather unfortunate. As AI is becoming a mainstream technology, we need research that studies human-AI collaboration from a management perspective. Consistent with the call for research raised in Bichler et al. (2010), we conclude that learning to collaborate with AI is the real challenge for the medium-term future of work.

We explore fundamental mechanisms for collaboration and delegation between humans and AI. The focal task is image classification. We have chosen image classification because humans were traditionally highly capable at this and have only recently been outperformed by deep learning-based AI (Szegedy et al. 2015). Image classification is not purely mathematical so that the limits of human computing power do not play a major role (this is also the case for most jobs that humans do at present). Image classification has important applications ranging from the development of autonomous vehicles (Floreano and Wood 2015) to the detection of skin cancer (Esteva et al. 2017). Central to our arguments is a rule for delegating work to the human or the AI. This rule is independent of the image classification context. We thus believe that our insights are generalizable.

Our central research questions are:

- Can delegation between humans and AI outperform humans or AI alone?
- Who delegates better, and why?
- How can humans learn to delegate to an AI? . . .

. . . and maybe most importantly:

- What does all this mean for the future of work?

We tackle these questions with three experimental studies. In the first study, we compare the performance that arises from different forms of delegation: No delegation (the humans or the AI performs all of the work alone), delegation (humans may delegate to AI), and inversion (the AI may delegate to humans, as suggested in McAfee 2013). There are two main outcomes of Study 1. (1) Inversion achieves by far the best performance. This might not be completely novel (some computer science systems are using a similar approach to reach high performance). It is nonetheless important for our arguments because the hierarchical shift induced by inversion (the AI tells humans what to do, and when), may be undesirable in certain applications. In these cases an alternative configuration is needed that performs good enough to be sustainable. One such alternative can be humans delegating to the AI, because (2) humans who delegate to the AI could in theory outperform inversion. However, they do not! It turns out that human performance is inferior, and they delegate badly due to unrealistic self-evaluation and lack of strategy.

In Study 2, we provide direct feedback to the subjects, so that they can learn what they did wrong and right, and thereby improve their self-evaluation. This is a simple and intuitive approach. However, it neither induced more delegation, nor did it lead to better performance.

Because of this, we supported the human decision makers even more strongly in Study 3. Here, we provided them with a good delegation strategy: They should delegate the image if they were uncertain, and they should classify the image themselves if they were certain about the correct result. We also let the humans report their certainty for every decision they made. This intervention lead to considerably more delegation and to better performance. The humans were now on par with the AI, but they sill could not beat it.

To understand this result, we study human delegation. Human delegation can be described using two regimes of image difficulty; a high-error region containing images with above-average difficulty, and a low-error region containing images with below-average difficulty. Humans delegate images in the low-error region less often as they become easier. This is rational and leads to high performance. However, we find no significant influence of image difficulty on delegation rate in the high-error region. In other words, humans delegate randomly if images are hard.

What might be even more interesting is that this separation into two regimes occurs on the level of "objective difficutly", but not on the level of "perceived difficutly". Human delegations are consistent with their personal perception of image difficulty, and they are reasonable as well. Humans reliably delegated images that they *considered* hard, and they classified images themselves that they *considered* easy. Unfortunately their perception of difficulty, and the images' true difficulty were often misaligned. Humans refused to delegate many hard images and classified them wrongly, because they thought the image was easy. In other words, humans did not know what they did not know.

We believe humans put real effort into delegating well, as they delegate consistently on basis of their perceived difficulty. They eventually did not reach good performance due to their misjudgment of task difficulty.

In the following, Section 2 describes the experimental designs and the results. Section 3 discusses the implications of the results for the future of work.

2 Experimental Studies

This section describes experimental designs and results of three experimental studies with a total of 1,506 subjects. Study 1 "Delegation and Inversion" compares different possibilities of delegation between humans and AI: AI alone, humans alone, humans who may delegate to AI (delegation), and an AI that may delegate to humans (inversion). Study 2 "Self-evaluation and Delegation" explores the self-evaluation of humans and its effect on delegation. This study employs a 2x2 factorial design with the dimensions delegation (yes/no) and feedback (yes/no). Study 3 "Explaining and Enforcing a Delegation Strategy" analyzes the effect of providing humans with a strategy on how to delegate. We compare three conditions: a baseline condition similar to the delegation condition in Study 1, a condition where a strategy based on self-evaluation is proposed, and a condition where this strategy is being enforced.

Please note that we followed the guidelines of Nosek et al. (2018) and pre-registered all experiments at the Open Science Foundation (this included the recruitment and data collection process, the initial hypotheses, and the statistical analysis).

2.1 Study 1: Delegation and Inversion

2.1.1 Experimental Design.

In all experiments, humans and/or AI classified images. 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 (for example, a focal picture with the ground truth of a poodle is assigned to the "poodle" class, not to the "huskie" or "cat" class). We selected 100 focal images from the ImageNet database (Russakovsky et al. 2014). The correct classes for the images are provided with the dataset. The performance measure was classification accuracy: the percentage of correctly classified images.

We compare classification accuracy between four conditions. In the *AI alone* condition (1), we used the GoogLeNet neural network (Szegedy et al. 2016). GoogLeNet is currently among the best AIs for image classification. We obtained its classification accuracy by applying GoogLeNet to the 100 images and by comparing its predictions to the actual classes. In the *humans alone* condition (2), human subjects recruited via Amazon's Mechanical Turk (MTurk) classified the images without any support. Subjects in the *delegation* condition (3) could choose for each image to either classify it themselves, or to delegate it to GoogLeNet. In an *inversion* condition (4), the AI could choose for each image to classify it itself or to delegate it to the humans.

We now describe the conditions in detail. In the *AI alone* condition (1) we used GoogLeNet Inception v3 (Szegedy et al. 2016). Inception v3 was trained on the ImageNet database with 1,000 classes. GoogLeNet assigns a score to each class and the class with the highest score is chosen as the answer.

For conditions (2)-(4), we conducted between-treatment experiments with 449 subjects, via MTurk, in August 2018. Each subject received a base fee of 50 cents, an additional 5 cents for each correctly classified image, and a bonus of 1 dollar if she succeeded in estimating her own accuracy after the experiment. Overall, subjects could get a maximum payment of \$6.50. Average pay was \$4.45, which was slightly above average pay on MTurk in general (Hara et al. 2018). The average duration of the experiment was 57.7 minutes.

We randomly assigned subjects to one of the conditions *humans alone* (149 subjects), *delegation* (154 subjects), and *inversion* (146 subjects). They received instructions, had to pass a short quiz so that we could exclude robots, and they completed an example classification to ensure that they understood the task. They then had to classify the 100 images in randomized order. Afterward, they were asked how many images they think they classified correctly (they could earn 1 dollar if this estimation did not differ from the actual number by more than five images), and they answered a short questionnaire.

In all cases, subjects were presented a focal image and ten classes as possible answers. Only one class was correct, the correct class being the ground truth from the ImageNet database. Like Russakovsky et al. (2014) we illustrated each answer class by its name and by showing 13 example images.[1]

The *humans alone* and *inversion* conditions were identical. In the *delegation* condition, we added a button labeled “Delegate this question to the AI”. The button was placed randomly between the answer categories. If a subject clicked on it, she would not classify the image herself, but delegate it to the AI. She would not see the AI’s answer, but the AI’s choice would be considered hers, and she would receive her payment accordingly. Thus, it was made clear that a subject is paid for each correct classification regardless of whether the AI or the human performed the task. Subjects in the *delegation* condition were informed about the AI and its accuracy at the beginning of the experiment. Figure 1 shows a screenshot of the *humans alone/inversion* and *delegation* conditions.

[1] Example: The focal picture contained a dog. The correct image class was “poodle”. We showed 10 possible answers, including poodle, German shepherds, bulldogs, boxers, Siberian huskies, and others. For each of the 10 classes, we showed 13 examples; 13 pictures of poodles, 13 pictures of boxers, 13 pictures of huskies and so on. The subject had to find out that the focal picture was a poodle. She had to click somewhere on the 13 poodle pictures. She could then click on a button labeled “Next image”. The answer was recorded, the screen refreshed and the next image was shown.

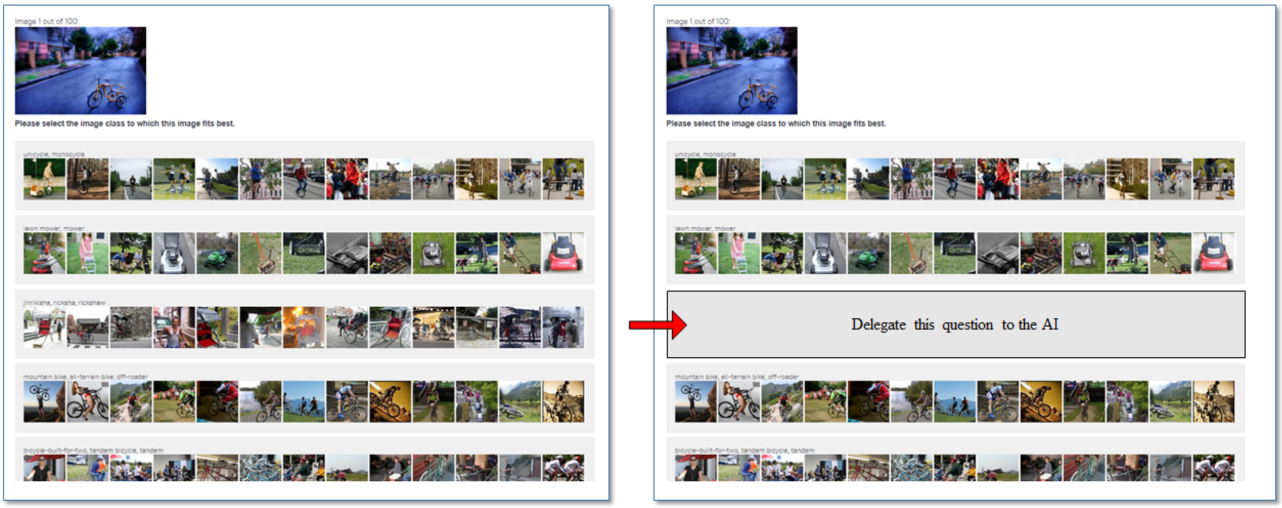

Image 1 out of 100:

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: Screenshot of the *humans alone/inversion* condition (left) and the *delegation* condition (right)

2.1.2 Results.

We constructed the results for the *inversion* condition (4) after the experiment. In this condition, the AI would classify images or delegate them to humans based on a simple rule: if the score of the potential answer was below a certain threshold, then the AI delegated this image to the humans. Otherwise, GoogLeNet classified by itself. To simulate this mechanism, we paired the AI with all subjects from the *inversion* condition, which lead to 146 pairs. We used 0.717 as the threshold, which was the average human accuracy from the *humans alone* condition. The rationale was that the AI should delegate images where the likelihood of being correct was below the average accuracy of human workers, i.e., humans could know better. We did this for every image and thereby generated classification accuracies for each of the 146 AI-human pairs. Please note that to avoid biases, all subjects in the *inversion* condition had to classify the same 100 images as the subjects in the other conditions.

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 the *delegation* condition (74.0%) and *humans alone* (71.7%). By itself, the accuracy of the AI is 77% (the vertical dashed line in Figure 2.[2] The standard

[2] We report Top1 accuracy values, which is the accuracy for choosing the top class correctly. In the AI literature, the Top5 error rate is sometimes reported, which is how often the true class is not within the top five classes. In our case, the Top1 error rate of the AI is 23%, the Top5 error rate is 6%. This is in line with

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									
	Delegation	154	0.25	0.74	0.99	0.10	0.70	0.76	0.80
	Humans	149	0.31	0.72	1.00	0.13	0.65	0.74	0.81
	Inversion	146	0.71	0.87	0.98	0.04	0.85	0.87	0.90

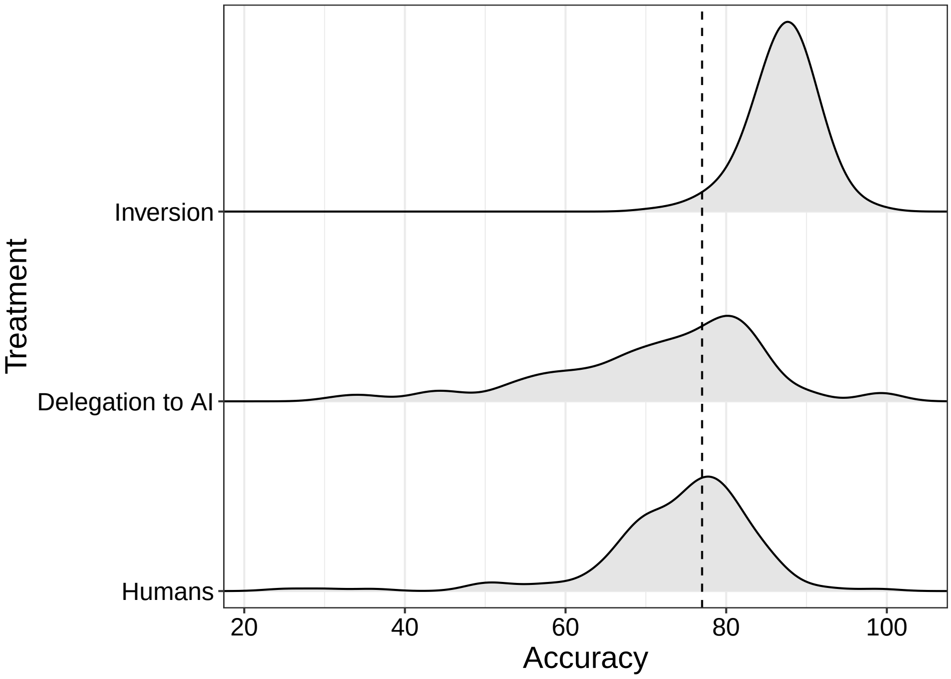

Figure 2: Distribution plots for accuracy per experimental condition in study 1. The vertical dashed line is the AI classification accuracy of 77%.

deviation of accuracy (in number of images) is highest when humans are working alone (13.2), smaller when humans can delegate (10.1) and smallest when the AI delegates to humans (4.2).

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 the 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 multiple comparisons show that most pair-wise mean differences are significant. 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 ($d = .20$). *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 ($d = 1.56$).

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*. In the *inversion* condition, the performance is significantly higher than that of AI working alone, suggesting that human workers can significantly improve the performance of an AI by providing their input. Not only is *inversion* on average better than the other settings, we also notice that the AI profits from almost all humans. Only three out of the 146 AI-human pairs generated a performance below AI accuracy (77%), and the 25th accuracy percentile of *inversion* (85%) lies already far above AI accuracy. This suggests that the AI can even profit from collaborations with low-performing human workers. Collaboration and delegation between humans and AI can thus produce results that outperform humans or AI alone. Importantly, inversion was highly effective while delegation was not.

To further understand what may be driving the inferior human performance in the *delegation* condition, we investigate the pattern of delegation by humans. In Figure 3 and Figure 4 the difficulty of an image is depicted on the horizontal axis. Image difficulty is indicated by accuracy in the *humans alone* experimental condition, e.g., an .2 difficulty/ accuracy means that 20% of the subjects classified the image correctly. The vertical axis in both figures shows the delegation rate, i.e., the ratio of subjects who have decided to not classify the image but to delegate it to the AI.

If we consider the entire dataset (Figure 3), a weak overall trend can be detected where

the literature.

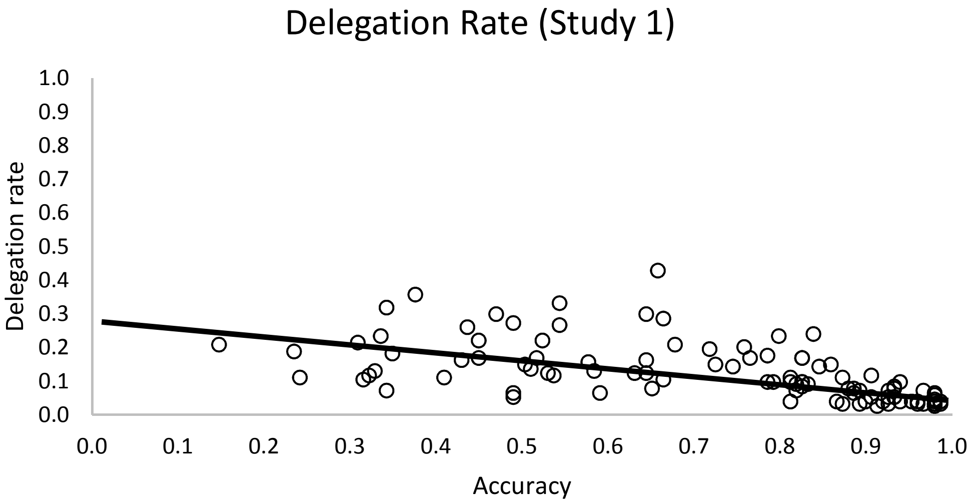

Figure 3: Scatter plot of accuracy per image (horizontal axis) against delegation rate per image (vertical axis). The regression line is estimated from all images.

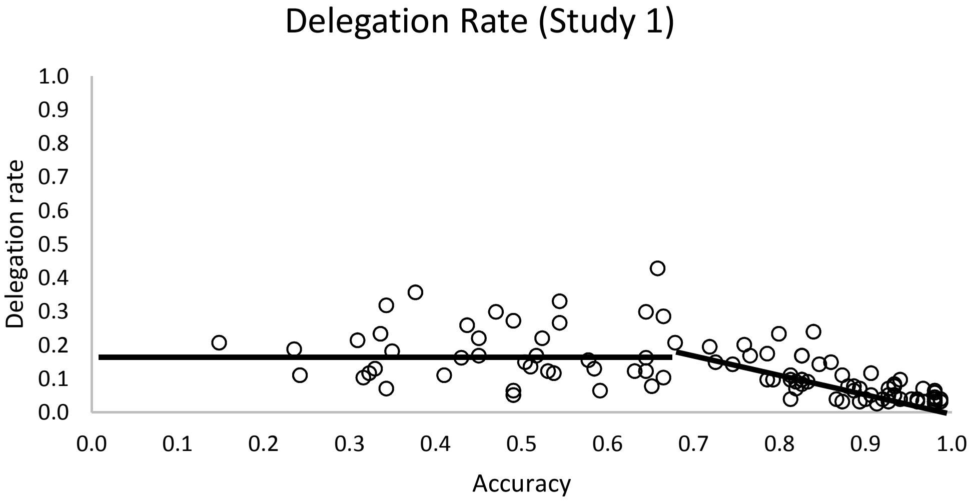

Figure 4: Scatter plot of accuracy per image (horizontal axis) against delegation rate per image (vertical axis). The two regression lines are estimated from two partitions of the data (only significant parameters considered). A high-error region contains images with above-average difficulty (accuracy per image $<$ 70%). A low-error region contains images with below-average difficulty (accuracy per image $\geq$ 70%.)

images with higher accuracy (lower difficulty) are delegated less often and vice versa. This is encouraging because it shows that humans were acting quite rationally by delegating more difficult images more often. However, partitioning the data into images that have less than 70% accuracy (these images are harder than average, recall that overall human accuracy is around 70%), and above 70% accuracy (these images are easier than average), the true trend can be detected. As Figure 4 depicts, human delegation is quite random when accuracy is low (the images are hard) and it follows a more rational delegation pattern beyond the threshold (when images are easy).

Table 2: Regression results for delegation (Study 1).

	Dependent variable: delegation rate	
	< 70% Accuracy	≥ 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	

This can be verified statistically by running two simple regressions on the partitioned data. We show the effect sizes and significance levels in Table 2. It becomes clear that there is no significant relationship between accuracy and delegation rate (dependent variable) below the 70% accuracy threshold (insignificant coefficient *Accuracy*), but there exists a strongly negative and significant relationship above 70% accuracy.

While humans seem have difficulties in delegation when classifying difficult images, they delegate quite rationally when dealing with images that are easy to classify (i.e., they delegate more when dealing with relatively difficult images and vice versa in this regime).

As discussed earlier, the collaboration between humans and AI has potential to significantly improve classification performance compared to either of them alone. However, while the AI is able to delegate effectively applying a simple rule, humans are not able to perform as well. This can be due to several reasons. First, it is well known that humans are reluctant to delegate decisions, possibly due to a form of distrust towards machines (Dietvorst et al. 2015). Second, given the results from the partitioned data above, we wondered whether it is possible for humans to come up with an effective delegation strategy at all. What hinders them to delegate well? We conducted the following two experiments to see whether we can instruct individuals on how to delegate, and thereby improve their performance.

2.2 Study 2: Self-evaluation and Delegation

2.2.1 Experimental Design.

In this set of experiments, our goal was to educate humans about their errors and thereby help them make better delegation decisions. The simple idea was to take the subjects through the same set of images that were used in the first experiment, but for the first 50 images provide them feedback on whether they classified the images correctly or not. The experimental design builds on Study 1. We split the 100 images into two sets of 50 images with comparable human and AI accuracies.

We implemented a 2x2 factorial design by manipulating two factors. First, we did (or did not) provide *feedback* during the first 50 images. If subjects received feedback, then they could see whether they were correct or not after each image; we also showed them the right answer. Second, we allowed (or did not allow) *delegation* to the AI during the second 50 images.

The resulting four experimental conditions were run using a between-subjects design with 604 subjects recruited via MTurk. Each subject received a base fee of 1 dollar, an additional 5 cents for each correctly classified image, and a bonus of 1 dollar each if she succeeded in estimation her accuracy for both sets of images after classification. Average pay was $5.43, the average duration was 58.0 minutes. The assignment process and experimental protocol was equivalent to Study 1.

2.2.2 Results.

In the following we consider accuracy and delegation rate for the second 50 images. Table 3 shows summary statistics for the accuracy and delegation rates. From the table, the results seem similar to the previous experiment. While *delegation* improves performance, *feedback* does not seem to have a significant effect.

Table 3: Summary statistics for *accuracy* and *delegation rate* (Study 2).

	Treatment		Summary statistic							
Dep. Var.:	Delegation	Feedback	N	Min.	Mean	Max.	St. Dev.	Pctl(25)	Median	Pctl(75)
Accuracy										
	No	No	159	0.08	0.71	0.90	0.12	0.66	0.74	0.78
	No	Yes	151	0.08	0.71	0.92	0.12	0.66	0.74	0.78
	Yes	No	150	0.42	0.76	0.92	0.09	0.74	0.78	0.82
	Yes	Yes	144	0.34	0.77	0.94	0.08	0.74	0.78	0.82
Delegation rate										
	No	No	159	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	No	Yes	151	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Yes	No	150	0.00	0.19	1.00	0.21	0.06	0.12	0.24
	Yes	Yes	144	0.00	0.20	1.00	0.23	0.04	0.14	0.28

Table 4: Regression results (Study 2)

	Dependent variable:	
	Accuracy	Delegation rate
Delegation	0.054***	
	(0.012)	
Feedback	0.002	0.015
	(0.012)	(0.026)
Delegation × Feedback	0.004	
	(0.017)	
Constant	0.711***	0.187***
	(0.008)	(0.018)
Observations	604	294
R^2	0.071	0.001
Adjusted R^2	0.066	-0.002
Residual Std. Error	0.102 (df = 600)	0.222 (df = 292)
F Statistic	15.213*** (df = 3; 600)	0.325 (df = 1; 292)
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses	

To analyze the causal effects of the experimental conditions on accuracy and delegation rates, we estimate linear models. The dependent variable is accuracy of the second 50 images per subject. Independent variables are two dummy variables modeling whether the subject had the ability to delegate (variable *Delegation* =1: yes, 0: no), and whether we provided feedback when she classified the first 50 images (variable *Feedback* =1: yes, 0: no). The regression results in Table 4 confirm that the performance feedback provided did not increase accuracy, and it did not make the subjects delegate more images (insignificant coefficients *Feedback* in both models). However, the results show that our proposition from Study 1 apparently holds: having the ability to delegate to an AI has a significant positive influence on performance.

It is somewhat disappointing that providing feedback - a common and most intuitive approach – did not create significantly better delegation. But it may be that simply providing feedback is not enough. Instead, we may need to provide active guidance on the delegation strategy.

To test this proposition, we designed another set of experiments. Recall that for the inversion condition in our first study, the machine follows two steps: it first assesses its own certainty about the classification, and then it delegates to humans if its certainty is below average human performance. There may be other strategies that yield the same, or even slightly better, performance but that is not the point. We do not try to maximize accuracy. Instead, we explore whether humans can detect and deploy an effective delegation strategy at all. Further, we want to explore what impediments exist, such that humans may eventually improve. Therefore, in a third set of experiments, we actively provide delegation strategy guidance to the human subjects.

2.3 Study 3: Explaining and Enforcing a Delegation Strategy

2.3.1 Experimental Design.

In this experiment we studied if teaching a delegation strategy, or enforcing it, helped human decision makers to better realize the potential of delegating to the AI. The experimental design was a between-subjects group comparison with three experimental conditions based on the *delegation* condition of Study 1. The first condition, *baseline*, was set up in analogy to the *delegation* condition of Study 1. Here, human decision makers classified 100 images and could delegate to the AI. We only made a small addition: the subjects had to report their level of certainty for each image on a scale from 1 (uncertain) to 4 (certain). The second condition, *strategy explained*, added a short text that describes and recommends a delegation strategy.

We told the subjects that they should delegate all images for which they were not certain (all images with certainty levels between 1 and 3). If certainty was high (certainty level 4), we advised them not to delegate. In the third condition, *strategy enforced*, subjects could not delegate actively. However, we informed them before the classification task that images will be delegated automatically if their self-reported certainty score was between 1 and 3. The human answer was only used if the reported certainty score was 4.

We recruited 453 subjects via MTurk, and randomly assigned them to experimental conditions. Each subject received a base fee of 1 dollar, an additional 5 cents for each correctly classified image, and a bonus of 1 dollar if she succeeded in estimating her own accuracy after the experiment. The average pay was $5.19, and the average duration of the experiment was 56.2 minutes. The assignment process and experimental protocol was equivalent to Study 1.

2.3.2 Results.

Table 5 shows summary statistics for accuracy and delegation rates. While the accuracy rates improved marginally, the delegation rates increased strongly when the strategy was explained or enforced. In fact, the delegation rate with *strategy enforced* looks similar to that of *strategy explained*, where humans made the delegation decisions themselves. Thus, humans follow the guidelines for delegation quite successfully. The standard deviations of accuracy were similar in the three groups (.09 for strategy enforcement and .1 for explaining the strategy and for not considering it all).

Table 5: 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	150	0.16	0.75	0.90	0.10	0.72	0.77	0.81
	Strategy explained	157	0.24	0.77	0.88	0.10	0.75	0.80	0.82
	Strategy enforced	146	0.14	0.78	0.90	0.09	0.75	0.79	0.82
Delegation rate									
	Baseline	150	0.00	0.13	0.68	0.15	0.01	0.08	0.20
	Strategy explained	157	0.00	0.34	0.95	0.20	0.19	0.33	0.47
	Strategy enforced	146	0.01	0.33	0.96	0.18	0.19	0.32	0.46

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 the means are different (ANOVA, F(2, 450)=2.97, $p = .052$, $\eta^2 = .13$ which represents a medium effect). 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 ($p = 0.048$) 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 (two-sided T-test, $p = .010$), but there is no significant difference between AI, the *strategy explained* ($p = .727$), and the *strategy enforced* condition ($p = .484$).

The results suggest that explaining the strategy, and enforcing it, can improve performance when delegating to an AI. Delegation alone did not create performance values that could compete with the AI. Explaining a good delegation strategy or enforcing it can however create delegation behavior that results in performance that is en par with AI. Still, humans could still not outperform it.

We now study the causal effect of the experimental variation on the delegation rate. Descriptive statistics suggest that explaining the strategy produced slightly higher average delegation rates than enforcing it (34.2% versus 32.5%). Subjects in the *baseline* condition delegated considerably fewer images (13.1%). A Levene test for homogeneity of variances shows that the variances are different across experimental conditions (F(2, 450)=7.5972, $p < .001$). According to an ANOVA with heterogeneous variances, means are also different (F(2, 296.14)=77.772, $p < .001$, $\eta^2 = .227$, which represents a medium to large effect).

Additional post-hoc tests with Tanhames T2 statistic for multiple comparisons show that explaining the strategy increases the delegation rate significantly (21.11 percentage points increase, $p < .001$, which represents a large effect (d=1.175)). Enforcing the strategy does not lead to a significantly different delegation rate from explaining it ($p = .750$). Enforcing the strategy leads to a significantly higher number of delegations than not considering the strategy at all (a 20.35 percentage point increase, $p < .001$, which represents a large effect (d=1.213)).

We felt perplexed that despite so much more delegation, the accuracy did not go up in similar amounts. Therefore, we explored the nature of delegation further. Figures 5, 6 and 7 present the delegation trends in Study 3. Like in Section 2.2.2, the horizontal axis depicts image difficulty (i.e., average accuracy of image classification by humans) and the vertical axis depicts delegation rates. Figure 5 considers the *baseline* condition, Figures 6 and 7 consider the *strategy explained* and *strategy enforced* conditions. It is interesting to see that consistent with the aforementioned tests, humans delegate more when strategy is explained. However, the delegation trend for difficult images is still random – the randomness just centers around a higher average delegation rate compared to the *baseline* condition.

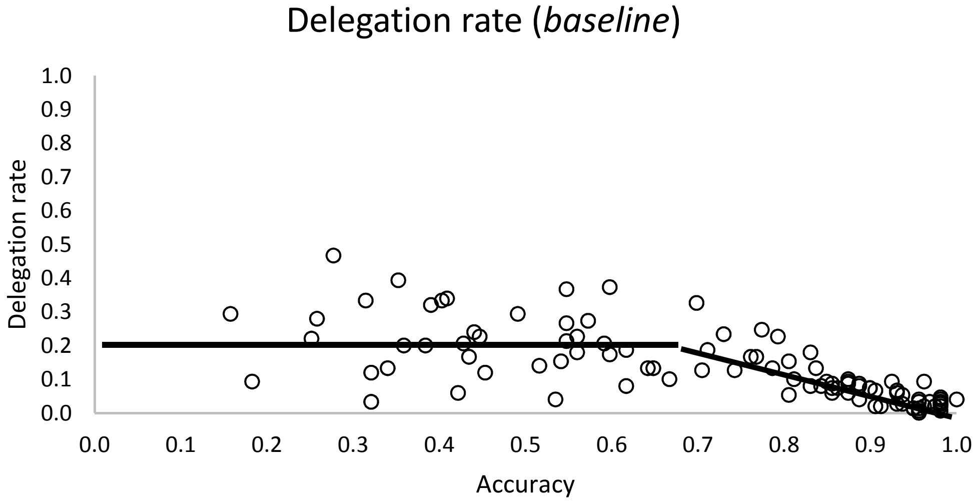

Figure 5: Scatter plot of accuracy per image (horizontal axis) against delegation rate per image (vertical axis) for the *baseline* condition in study 3. The two regression lines are estimated from the two partitions of the data (only significant parameters considered).

We ran regressions on partitioned data. Table 6 shows the statistical results which support the initial visual impression. It appears that the modest gains in performance come from the fact that human delegation, while it increases significantly on average with explaining or enforcing a strategy, is still is prone to somewhat random fluctuations for difficult images.

To explore this perplexing outcome further, we study whether humans are able to assess their own ability to classify images. Remember that subjects evaluated their certainty of their choices on a scale between 1 (uncertain) and 4 (certain). We ran a regression where the level of certainty was the dependent variable, and accuracy the independent variable. This model shows us whether the subjective assessment depends on the objective difficulty of an image. The regression results in Table 7 show that humans are good at assessing their own ability, i.e., how difficult the images are to classify, for simpler images (where accuracy of image classification was above 70%), but they are not able to do this for difficult images. The values seem to differ slightly for the *strategy enforced* condition. An explanation is that humans reported

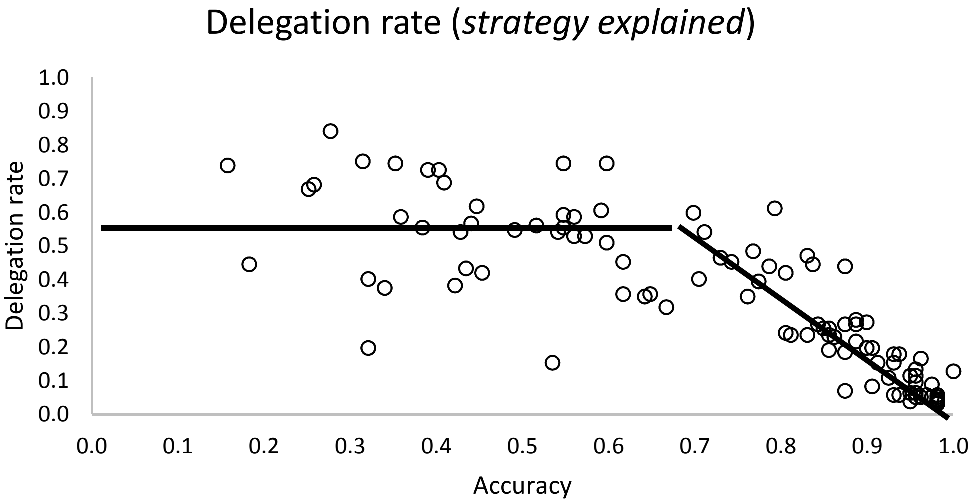

Figure 6: Scatter plot of accuracy per image (horizontal axis) against delegation rate per image (vertical axis) for the *strategy explained* condition in study 3. The two regression lines are estimated from the two partitions of the data (only significant parameters considered).

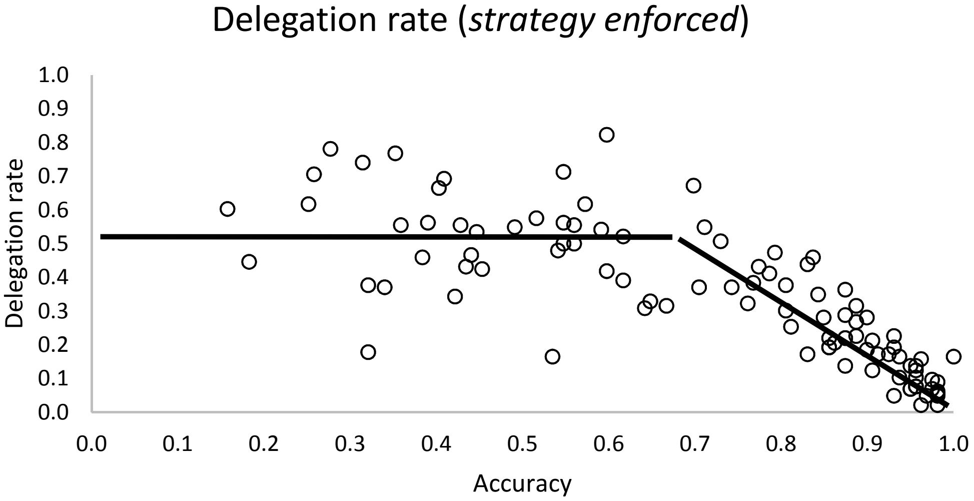

Figure 7: Scatter plot of accuracy per image (horizontal axis) against delegation rate per image (vertical axis) for the *strategy enforced* condition in study 3. The two regression lines are estimated from the two partitions of the data (only significant parameters considered).

Table 6: Regressions per experimental condition (Study 3). 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	-0.113	-0.631***	-0.337	-1.710***	-0.211	-1.535***
	(0.121)	(0.055)	(0.178)	(0.124)	(0.180)	(0.110)
Constant	0.268***	0.636***	0.698***	1.731***	0.617***	1.578***
	(0.058)	(0.049)	(0.086)	(0.110)	(0.086)	(0.098)
Observations	40	60	40	60	40	60
R2	0.023	0.695	0.086	0.768	0.035	0.771
Adjusted R2	-0.003	0.689	0.062	0.764	0.010	0.767
Residual Sd. Error	0.104	0.033	0.154	0.075	0.155	0.066
F Statistic	0.885	131.912***	3.586*	191.564***	1.377	195.449***

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

higher certainty levels to avoid delegation to the AI (remember that any image with a reported certainty level of less than 4 is automatically delegated).

Therefore, it appears that while humans delegate quite rationally once the delegation process is explained, their own assessment of what they can do exhibits a high amount of uncertainty for relatively difficult tasks. It also suggests that although human decisions are often misaligned with real problem difficulty, they are not misaligned with perceived problem difficulty. This becomes clear from the raw data plots in Figure 8 (please find a summary of all plots of Study 3 in the Appendix). We believe that the subjects put real efforts into delegating well. But their cognitive limitations hindered them in delegating the right images.

In the following, we discuss the implications of our finding for the future of human work with AI, algorithms and other intelligent machines.

Table 7: 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		Strategy explained		Strategy enforced	
	Dependent variable: Certainty, where images have accuracy of...					
	$< 70\%$	$\geq 70\%$	$< 70\%$	$\geq 70\%$	$< 70\%$	$\geq 70\%$
Accuracy	0.402	3.612***	0.666	3.750***	0.329	2.438***
	(0.381)	(0.271)	(0.461)	(0.277)	(0.384)	(0.184)
Constant	2.640***	0.273	2.538***	0.201	3.054***	1.519***
	(0.183)	(0.243)	(0.222)	(0.247)	(0.185)	(0.165)
Observations	40	60	40	60	40	60
R^2	0.029	0.753	0.052	0.760	0.019	0.751
Adjusted R^2	0.003	0.749	0.027	0.756	-0.007	0.747
Residual Sd. Error	0.329	0.164	0.399	0.167	0.332	0.111
F Statistic	1.116	177.148***	2.085	183.367***	0.732	174.959***

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

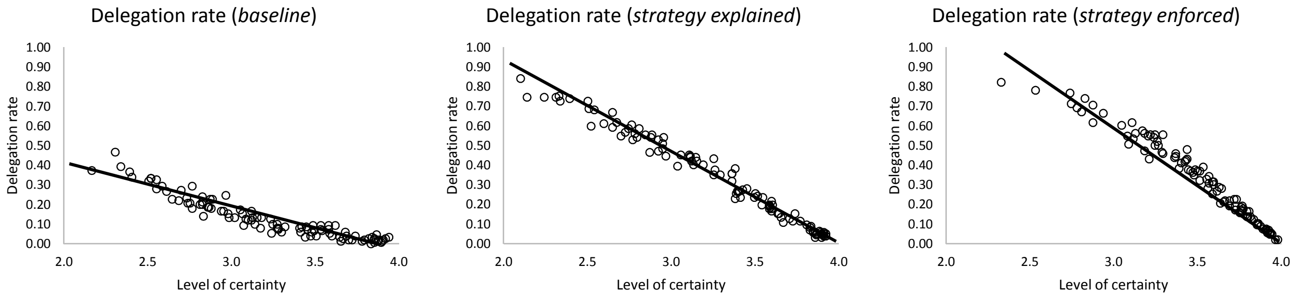

Figure 8: Scatter plots of *certainty* against delegation rate per image and per experimental conditions in study 3.

3 Discussion

Collaboration and delegation between humans and AI can produce results that outperform humans or AI alone. Humans and AI were working on the same task and they were shifting the workload through delegation. The best-performing scenario is *inversion*, where the AI is responsible, but delegates when it is uncertain. In this scenario, both the AI and the humans are contributing to the result. This sets our study apart from works where AI outperforms humans, and the complete substitution of human workers by AI is seen as the logical consequence. Humans could increase their performance through delegation, but the gains were not big enough to outperform even the AI alone.

The dominance of inversion and the inability of humans to work effectively with AI raise important questions that we try to address in the following.

3.1 Implications for Augmented Decision Environments

Inversion is the most productive scenario in our study, but it is not what academia and practice consider decision support. Most published research and the available software tools model the interaction between humans and AI exactly the other way around: the human is the user. She is responsible for answering a question or for solving a problem, and she may use a computer, a tool, an algorithm, an AI. Our results corroborate previous findings that this can improve human performance, but it may not improve it beyond the performance of the AI. Put differently: using an AI for decision support might not save humans their jobs.

If research would follow the inversion paradigm, the literature on decision support systems (DSS) would have to adapt. There are multiple ways in which inversion can change the design principles for a DSS: humans could deliver input for the algorithm, humans could be asked to complete a partial task, or they could be asked to work on a task and submit their result. In all these cases, the AI would be the boss. How should we design such systems algorithmically and formally, how should we evaluate them, and how do they perform? Rethinking DSS for inversion could also lead to the integration of new goal functions, new constraints, or new solution algorithms.

3.2 Inversion and the Future of Work

Inversion is not a completely novel scenario, but a discussion of its impact on the future of work is still needed. The implications of implementing inversion are easily explained when considering a situation where performance comparisons between human workers and an AI are made to justify decisions about employment and task allocation. Assume we had to assign the task of image classification to a group of humans or to an AI. The AI clearly outperforms humans on average. Consequently, one would allocate this task to the AI; the humans would be *unemployed*. This is in concordance with expectations. Image recognition is an area in which humans were traditionally successful, and have recently been outperformed by deep learning-based AI. Our results, however, question whether the assignment of image recognition to an AI is actually the best choice. If the workload could be split between an AI and humans, and if the AI would be the responsible unit which would delegate to humans, then several interesting things would happen. First, the resulting performance would be higher than that of the AI alone. This makes collaboration and delegation economically desirable. Second, because the AI delegates to humans, humans would do some of the work. They contribute to the superior result; without them, we would not reach it. Put differently, the human would not be unemployed. However, the result is bittersweet as the inversion scenario comes with a loss of control. The AI decides about the delegations, it asks the human for support only if it is required. This means that the company using inversion has to agree with an AI organizing human work; and humans have to accept as well that they are below the AI in the hierarchy. Are humans then buying themselves employment by moving down the hierarchy?

Actually, the answer is not so clear and the situation not so dire. Let us propose an optimistic statement: *inversion can improve human work perspectives.* From research in human motivation, we know that humans are more motivated when working in a stimulating environment (Pink 2009). The classification of easily identifiable images is perhaps a routine and boring task, whereas the classification of difficult images could be seen as a 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 as a delegation to humans, but more as freeing humans from boring tasks. The AI would not be the humans' boss, but rather an assistant who swipes away distractions from the real work. On top of this, if inversion creates free time, workers could use it to educate themselves in tasks that are not prone to automation. In this manner, inversion could create a win-win situation.

3.3 Delegation to AI is a skill that should be taught

Splitting the workload between humans and AI by delegation can outperform humans or AI working alone if two requirements are met. First, humans and AI have to have complementary skills for the task at hand. This was clearly the case in our experiment. An optimal combination of the AI and the human workers from the *inversion* condition would lead to an accuracy of 89.9%. This is considerably higher than the accuracy levels of 77.0% for *AI alone* and 71.7% for *humans alone*. Thus, there are images that the AI can classify and humans get wrong, and vice versa.

Second, delegation has to move a task to the party that is better at it. Under perfect information, a simple rule always leads to this result: if the other actor is able to do task and you are not, then delegate. Otherwise, do it yourself. Obviously, knowing what you can and cannot do is the tricky part – both for humans and the AI. We made the AI work with such a rule without emotions. Images were delegated if the likelihood that the AI's choice is expected to be correct was below average human accuracy. Humans experienced severe problems applying such a rule and generally could not outperform the AI when working with it.

The AI seems to be much better at estimating its own accuracy. Modeling the likelihood of class membership from a large database of examples is the core technical challenge that AI engineers face when they are building an AI for classification. The success of the AI rests largely on how well it can estimate such likelihoods. In our example, the average likelihood of the top class being the correct one was 0.766, and 77 out of the 100 images were correctly classified. As a consequence, the AI was very good at deciding which images to delegate and which images to classify by itself. The average accuracy for non-delegated images was 98.6%. Humans, by contrast, overestimated the number of correctly classified images by 8.0% with a mean absolute percentage error of 21.3%. The average accuracy of images that they classified themselves was just 74.9%. Thus, humans classified many images themselves that they should have delegated to the machine.

In other words, humans did not delegate correctly. Given the results from prior studies that demonstrate aversion to working with algorithms, we tested whether the lack of delegation is due to an aversion to working with machines or due to not being aware of how best to use an algorithm or, in other words, not being aware of when to delegate. Our second and third set of experiments reveal that humans do not necessarily have an aversion to using algorithms in economically motivated environments. Providing feedback as well as consciously assessing the difficulty of the classification task can motivate humans to delegate at a much higher rate, which indicates the willingness to work with the machine.

However, even with greater delegation rates, humans do not automatically achieve the performance that the AI can achieve in the *inversion* condition. An exploration of the causal effect for this unfortunate outcome revealed that the inferior performance is due to humans not being able to judge the complexity of relatively difficult tasks as well as that of relatively easy tasks (see Table 7).

We believe that this is not covered entirely by our bounded rationality. As the concept of bounded rationality (Simon 1955) argues, humans tend to make decisions that satisfice rather than decisions that are optimal. However, bounded rationality does not address whether humans have the ability to judge how difficult a task is, and whether the decision to satisfice is therefore engaged appropriately. Our rule for satisficing in collaboration with a machine, i.e., "if the other actor is able to do task and you are not, then delegate; otherwise, do it yourself" works well for the machine. It does not work well for humans because they are not good at judging their own abilities. This is a significant finding that, to our knowledge, is novel in the area of AI-human collaboration. It provides interesting insights for designing effective work arrangements, the future of work and workforce education.

In the short term we see significant economic benefits and potential performance enhancement due to AI. This is why we consider it likely that humans will often be involved in inversion-like conditions. Computer algorithms will allocate work that the algorithm does not have a strong confidence in, so that humans are engaged in tasks that leverage human abilities and that are hard to codify.

However, there may be situations where inversion is currently not an option, for example in critical medical decisions that for cultural and ethical reasons require a human to be responsible. In this case, the human will likely be *supported* by AI, but the AI will not make the decision with the help of a human. These contexts will profit from better delegations.

Even in contexts were inversion is currently not an option, researchers might have an obligation to figure out how humans make better delegations. Cultural biases towards human decision makers may not prevail forever. If humans delegate badly, but AI continues to become better and better at the focal task, then the human might eventually be attacked for his poor results and the work arrangement might eventually be "inverted". In an age of AI coworkers it might be just essential to delegate well. Researching collaboration and delegation becomes an important committment to the survival of humans in the workforce.

Finally, as educators we share a responsibility to teach future students how to develop the ability to self-assess their abilities honestly and effectively. Delegation is a leadership trait that humans have excelled at. The invention of machines themselves is a prime example. Leadership is an essential skill that management schools purport to teach, however, we simply do not see the challenges arising from leading machines as part of our curriculum, yet. We strongly believe that delegation is a skill that can and should be taught. Doing so may not just improve our students' carreers, but it may also enhance their ability to learn and grow.

3.4 Limitations and future research

It may appear that an obvious area of future research will be to replicate our experiments in different contexts. However, in our opinion, given the general logic-based rule for delegation that we used in inversion and tried to teach the humans, a more fruitful area of research would be to understand various directional complementaries that exist between humans and machines in tasks that are in danger of being automated by AI.

We see this work as a first step in the analysis of more complex human-machine relationships. It opens up two complementary research streams: Improving human capabilities in delegating tasks to a machine, and creating fruitful working environments using inversion. We believe that humans could be strong in delegation - but what needs to be done to unleash this ability? How do systems need to be set up, how do humans need to be trained, how do interfaces between systems and humans have to be designed?

Regarding the second question: how would an inversion scenario have to be set up? Can people improve in more challenging tasks if an AI relieves them of boring work? How should the communication between AI and humans be designed? The central challenge will be to understand in which situations which style of collaboration between humans and AI should be striven for. It is also unclear what inversion will do to innovation. Can firms with and without inversion coexist in a market?

We believe that all these are fascinating issues that researchers need to focus on, rather than focusing simply on the planned obsolescence of humans at the workplace.

References

Attenberg J, Ipeirotis P, Provost F (2015) Beat the machine. *Journal of Data and Information Quality* 6(1):1–17, ISSN 19361955.

Bichler M, Gupta A, Ketter W (2010) Research commentary—designing smart markets. *Information Systems Research* 21(4):688–699.

Branson S, Wah C, Schroff F, Babenko B, Welinder P, Perona P, Belongie SJ (2010) Visual recognition with humans in the loop. Daniilidis K, Maragos P, Paragios N, ed., *European Conference on Computer Vision*, 438–451, Lecture notes in computer science (Berlin, Heidelberg: Springer).

Chen C, Seff A, Kornhauser A, Xiao J (2015) DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2722–2730 (IEEE), ISBN 978-1-4673-8391-2.

Deng L, Yu D (2013) Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing* 7(3-4):197–387.

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.

Dietvorst BJ, Simmons JP, Massey C (2016) Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64(3):1155–1170.

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.

Floreano D, Wood R (2015) Science, technology and the future of small autonomous drones. *Nature* 521(7553):460.

Frey CB, Osborne MA (2017) The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change* 114:254–280.

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (MIT Press).

Gray D (2018) Robot revolution: AI and the future of work: Will the rise of artificial intelligence make you more or less likely to find your dream job? *The Economist* URL `http://shapingthefuture.economist.com/robot-revolution-ai-and-the-future-of-work/`.

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical versus mechanical prediction: A meta-analysis. *Psychological assessment* 12(1):19–30.

Hara K, Adams A, Milland K, Savage S, Callison-Burch C, Bigham J (2018) A data-driven analysis of workers' earnings on amazon mechanical turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 449 (ACM).

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-mod`

Holzinger A (2013) Human-computer interaction and knowledge discovery (hci-kdd): What is the benefit of bringing those two fields to work together? Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu R C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Vardi MY, Weikum G, Cuzzocrea A, Kittl C, Simos DE, Weippl E, Xu L, eds., *Availability, Reliability, and Security in Information Systems and HCI*, volume 8127 of *Lecture notes in computer science*, 319–328 (Berlin, Heidelberg: Springer Berlin Heidelberg), ISBN 978-3-642-40510-5.

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts, H J W L (2018) Artificial intelligence in radiology. *Nature Reviews Cancer* ISSN 1474-1768.

Joshi AJ, Porikli F, Papanikolopoulos N (2009) Multi-class active learning for image classification. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2372–2379 (IEEE), ISBN 978-1-4244-3992-8, URL `http://dx.doi.org/10.1109/CVPR.2009.5206627`.

Kahlen M, Ketter W, van Dalen J (2018) Electric vehicle virtual power plant dilemma: Grid balancing versus customer mobility. *Production and Operations Management* 1–17.

Knowledge@Wharton (2017) Technology: Why AI is the new electricity. URL `http://knowledge.wharton.upenn.edu/article/ai-new-electricity`.

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.

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

Logg J, Minson J, Moore DA (2018) Algorithm Appreciation: People Prefer Algorithmic To Human Judgment. URL `https://ssrn.com/abstract=2941774`.

Maes P (1994) Agents that reduce work and information overload. *Communications of the ACM* 37(7):30–40, URL `http://dx.doi.org/10.1145/176789.176792`.

Marlin D (2018) Millennials, This Is How Artificial Intelligence Will Impact Your Job For Better And Worse. *Forbes* URL `https://www.forbes.com/sites/danielmarlin/2018/01/16/millennials-this-is-how-artificial-in`

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-j`

Moravčík M, Schmid M, Burch N, Lisý V, Morrill D, Bard N, Davis T, Waugh K, Johanson M, Bowling M (2017) DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356(6337):508–513.

Nagar Y, Malone TW (2011) Making business predictions by combining human and machine intelligence in prediction markets. *Proceedings of the International Conference on Information Systems ICIS 2011.*

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.

Peek N, Combi C, Marin R, Bellazzi R (2015) Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. *Artificial Intelligence in Medicine* 65(1):61–73.

Pink D (2009) *Drive: The Surprising Truth About What Motivates Us* (New York, NY: Riverhead Books).

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 (2014) ImageNet Large Scale Visual Recognition Challenge. URL `http://arxiv.org/pdf/1409.0575v3`.

Russakovsky O, Li LJ, Fei-Fei L (2015) Best of both worlds: Human-machine collaboration for object annotation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2121–2131 (IEEE), ISBN 978-1-4673-6964-0.

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

Schölkopf B (2015) Learning to see and act. *Nature* 518:486–487.

Silver D, Huang A, Maddison CJ, 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, Hassabis D (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529:484–489.

Simon HA (1955) A behavioral model of rational choice. *The Quarterly Journal of Economics* 69(1):99–118.

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.

Wang J, Ipeirotis PG, Provost F (2017) Cost-effective quality assurance in crowd labeling. *Information Systems Research* 28(1):137–158, ISSN 1047-7047.

Williams A (2017) Robot-Proofing Your Child's Future. *The New York Times* D1, URL `https://www.nytimes.com/2017/12/11/style/robots-jobs-children.html`.

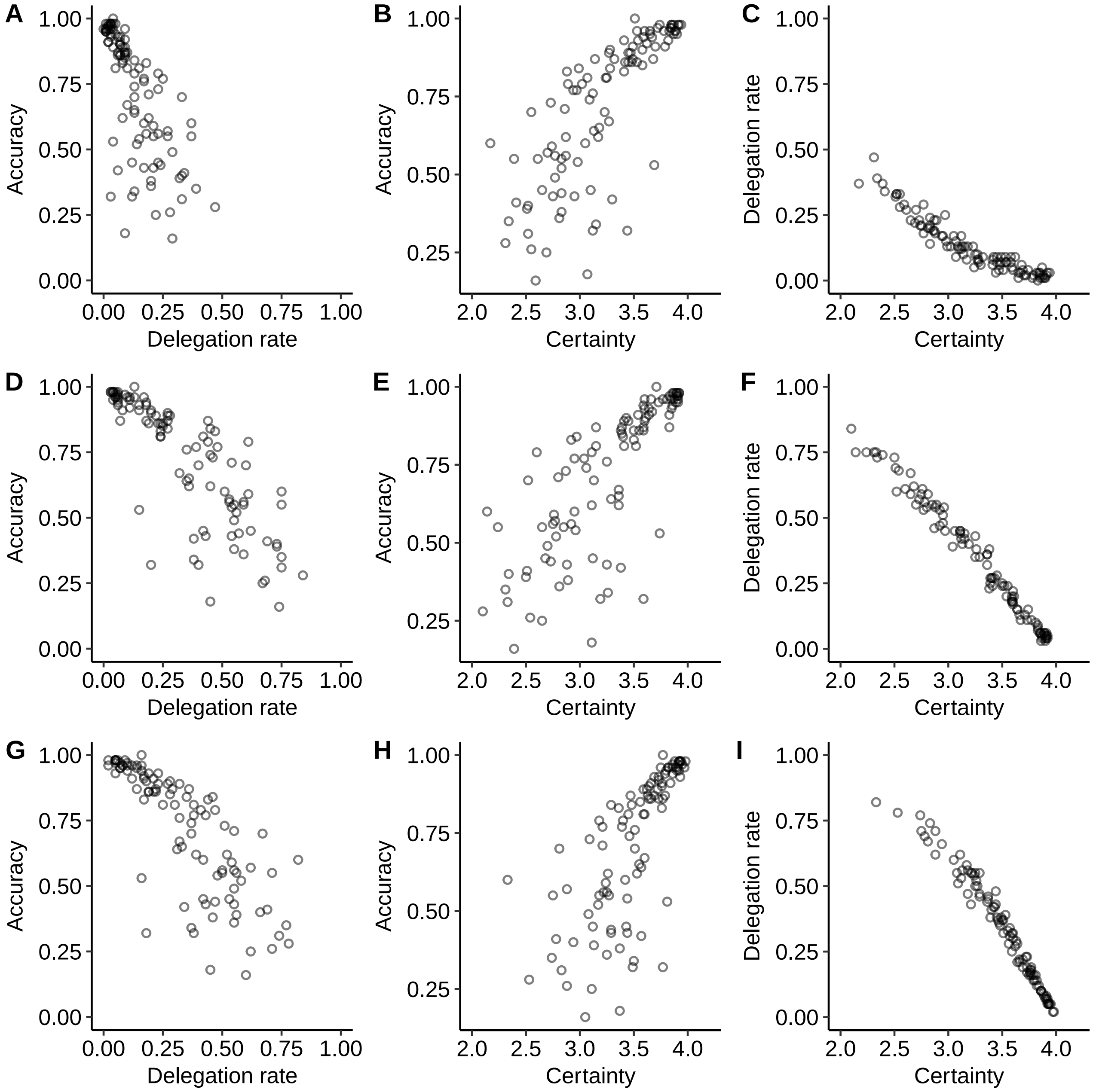

Figure 9: Scatter plots for *delegation rate*, *accuracy* and *certainty* of Study 3. The figure shows all three experimental treatments. Top row (images A, B, C): no strategy explained. Middle row (images D, E, F): strategy explained. Bottom row (images G, H, I): strategy enforced.