Roles of AI in Collaboration with Humans: Automation, Augmentation and the Future of Work

Andreas Fügener, Dominik D. Walzner, Alok Gupta

Appendix A: Mathematical proofs

Proof of Equation (11)

Automation achieves better task performance than augmentation if the relative AI performance is greater than the ratio of ignoring incorrect advice and ignoring correct advice:

$$p_t^{AI} > p_t^{AI}\left(p_t^H + r^{AI}(1-p_t^H)\right) + (1-p_t^{AI})\left(p_t^H - r^{\overline{AI}}p_t^H\right)$$

$$\frac{p_t^{AI}}{1-p_t^{AI}} > \frac{p_t^{AI}}{1-p_t^{AI}}\left(p_t + r^{AI}(1-p_t^H)\right) + \left(p_t - r^{\overline{AI}}p_t\right)$$

$$\frac{p_t^{AI}}{1-p_t^{AI}}\left(1 - \left(p_t^H + r^{AI}(1-p_t^H)\right)\right) > \left(p_t^H - r^{\overline{AI}}p_t\right)$$

$$\frac{p_t^{AI}}{1-p_t^{AI}}(1-p_t^H)(1-r^{AI}) > p_t^H\left(1-r^{\overline{AI}}\right)$$

$$\frac{\frac{p_t^{AI}}{1-p_t^{AI}}}{\frac{p_t^H}{1-p_t^H}} > \frac{1-r^{\overline{AI}}}{1-r^{AI}} \Leftrightarrow \widehat{p_t^{AI}} > \frac{1-r^{\overline{AI}}}{1-r^{AI}}.$$

Proof of Equation (13)

Augmentation achieves better task performance than humans if the relative AI performance is greater than the ratio of incorrect and correct advice.

$$p_t^{AI}\left(p_t + r^{correct}(1-p_t)\right) + (1-p_t^{AI})\left(p_t - r^{incorrect}p_t\right) > p_t$$

$$p_t^{AI}\left(r^{correct}(1-p_t)\right) + (1-p_t^{AI})\left(-r^{incorrect}p_t\right) > 0$$

$$\frac{p_t^{AI}}{1-p_t^{AI}}\left(r^{correct}(1-p_t)\right) > \left(r^{incorrect}p_t\right)$$

$$\frac{p_t^{AI}}{1-p_t^{AI}} > \frac{r^{incorrect}}{r^{correct}}\frac{p_t}{1-p_t}$$

$$\frac{\frac{p_t^{AI}}{1-p_t^{AI}}}{\frac{p_t}{1-p_t}} > \frac{r^{incorrect}}{r^{correct}} \Leftrightarrow \widehat{p_t^{AI}} > \frac{r^{incorrect}}{r^{correct}}$$

Appendix B: Effect of reallocation benefit on the distribution of work

In Figure 1, we replicate our analysis of the dominant work constellation shown in Section 3.5 in the main text. This time, we incorporate potential reallocation benefit, which means that releasing a human with automation enables this human to provide some value on another task. In this analysis, we treat the reallocation benefit as exogeneous and analyze the resulting work constellation for different combinations of between-task complementarity, within-task complementarity, and reallocation benefit. For low between-task complementarity, we set $a = 0.1$ and $b = 0.8$, that is, $p_t^H = 0.1 + 0.8p_t^{AI}$, and for high between-task complementarity, we set $a = 0.4$ and $b = 0.2$, that is, $p_t^H = 0.4 + 0.2p_t^{AI}$. As in the main analysis, we set $r^{AI} + r^{\overline{AI}} = 1$, and $\hat{r} = \frac{r^{AI}}{r^{\overline{AI}}} = \frac{r^{AI}}{1-r^{AI}}$.

Considering reallocation benefit increases the proportions of tasks where automation is most beneficial. The proportions of tasks in which augmentation is most beneficial are reduced considerably as a consequence: As automation is already chosen in cases of higher levels of AI performance, augmentation can outperform humans only if the relative weight of correct advice is high. Interestingly, a high reallocation benefit particularly affects humans with low between-task complementarity, as they are prone to being replaced by AI if reallocation is beneficial, and humans perform relatively better for fewer tasks. In the case of a reallocation benefit of 0.1, there is no task in which the human performance exceeds the AI performance by more than 0.1, thus, all tasks are automated. Augmentation might be relevant for cases with very high relative weight of correct advice (in this example, the first augmentation case is dominant for $p_t^{AI} = 0.3$ and $\hat{r}$=2.15).

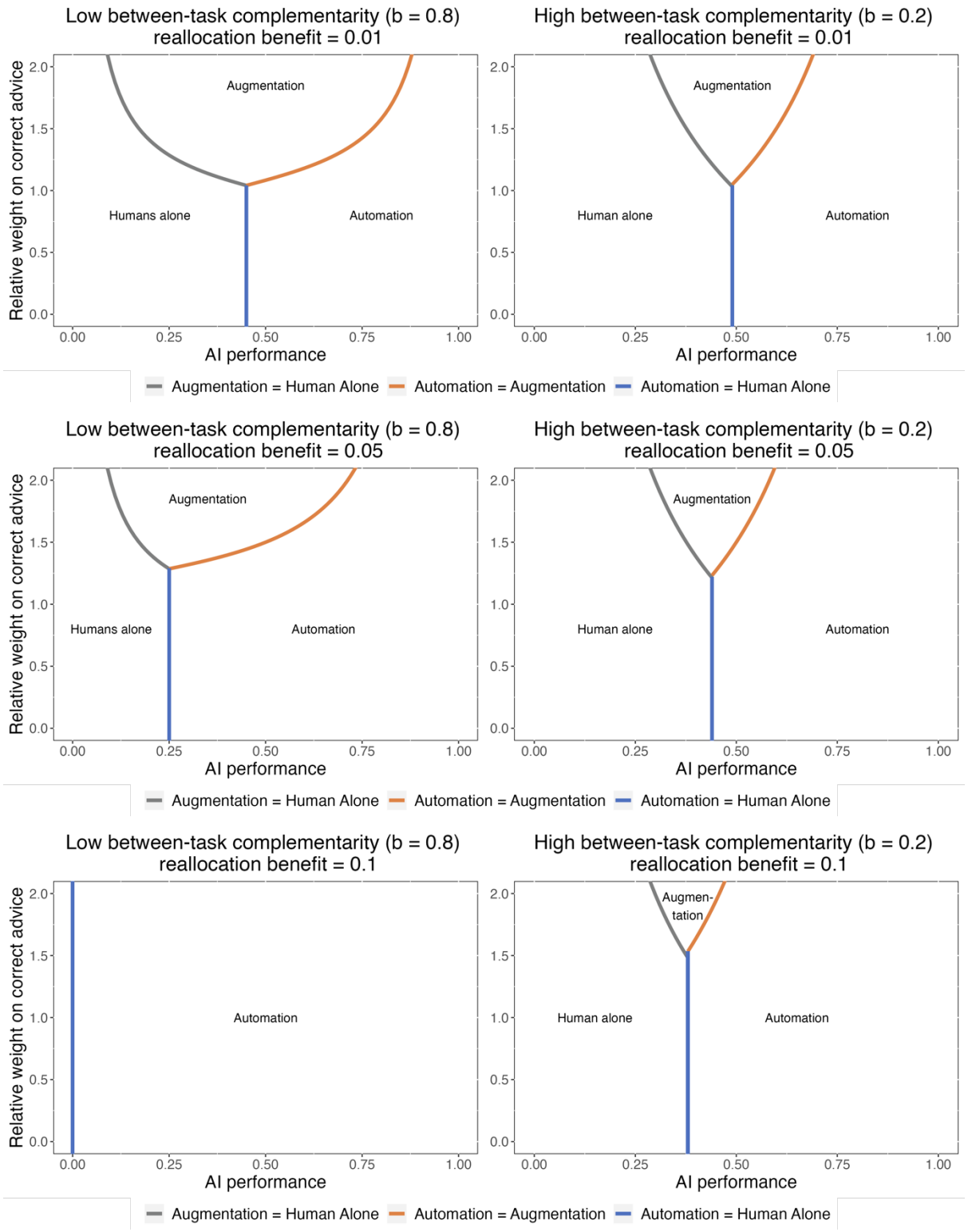

Figure 1: Dominant work constellation depending on AI performance, within-task complementarity (relative weight of correct advice), between-task complementarity and reallocation benefit

Appendix C: Experimental details (Fügener et al. 2021)

In this appendix, we report the details regarding the collected experimental data from Fügener et al. (2021) that we used to simulate the performance of our task allocation framework.

Task and treatment description

In the experiment, the subjects were given the task of assigning a focal image (such as an image of a small black dog) to one of ten possible image classes. To facilitate this process, the class name for each of the ten categories was provided, such as "Swiss mountain dog", along with 13 images that belonged to that specific class. The subjects chose an image class by clicking on the box that contained the class name along with the 13 images, similar to the experiments in Russakovsky et al. (2015). The images varied in difficulty, meaning that some images were easier to classify for humans than other images were. All subjects classified the same images.

The experiment included three treatments, however, for the purpose of our analysis, only the first two treatments were used in our main study. We used the third treatment for a robustness check, which is shown in Appendix F. In Treatment 1, the subjects classified images alone without any AI support. In Treatment 2, the subjects received advice in the form of an image class recommendation from an AI, that is, GoogLeNet Inception v3 (Szegedy et al. 2016). In Treatment 3, the subjects additionally received information about AI certainty to help them differentiate between correct and incorrect advice. For each image, AI calculated a likelihood score for each of the possible classes. This score represents the probability that a particular class is the correct class for the given image and thus indicates the expected AI performance for this task. In the experiment, AI recommended the image class with the highest likelihood score to the subjects. In the set of 100 images used, AI classified 77 out of the 100 images correctly.

At the beginning of the experiment, the subjects received introductory information about the task and underwent an attendance check. Subsequently, subjects were randomly assigned to one of the experimental treatments and given instructions that differed only in terms of the advice provided. In the main task, the subjects classified 100 randomly ordered images, with the possible classes for each image presented in random order as well. After the classification task, the subjects were asked to estimate the number of images they had correctly classified. They were also requested to report how they made their decisions. The subjects in Treatments 2 and 3 were additionally required to answer a questionnaire on human-computer trust (adapted from Madsen and Gregor 2000). The experiment concluded with a brief demographic questionnaire, after which the subjects received feedback on their performance.

Study Protocol

The data was collected on August 8, 2019. In total, 458 subjects were recruited through Amazon Mechanical Turk (MTurk). Only subjects from the United States were considered who had a positive rating of at least 90% on MTurk, who had not previously participated in any

related studies, who correctly answered an attention check, and who met the necessary technical requirements. Through random assignment, 146 subjects were allocated to Treatment 1, 160 to Treatment 2 and 152 to Treatment 3. Information about the subjects' gender, age, education level, and income class can be found in Table 1. Each participant received $1 for participating and an additional $1 if they correctly estimated the number of images they classified (with a margin of +/- five images) after completing the 100 classifications. Additionally, the subjects earned $0.05 for each image they correctly classified. In Treatments 2 and 3, the subjects received an extra $0.50 for completing a survey regarding their trust in the AI used for classification. The total payment for the subjects varied between $1 and $7 for Treatment 1 and between $1.5 and $7.5 for Treatments 2 and 3. On average, the duration of the experiment was 57.4 minutes, and the average payment was $5.77.

	Treatment 1	**Treatment 2**	**Treatment 3**
	$n = 146$	$n = 160$	$n = 152$
Gender			
Female	53%	38%	46%
Male	47%	61%	52%
Other	0%	1%	2%
Age			
18-24 years	7%	11%	11%
25-34 years	43%	51%	41%
35-44 years	25%	21%	28%
45-54 years	14%	7%	12%
55-64 years	10%	8%	7%
65-74 years	1%	3%	1%
Education			
None	0%	2%	1%
Highschool	15%	13%	13%
College	27%	26%	24%
Bachelor	40%	47%	45%
Master	15%	11%	14%
Professional	3%	1%	1%
Doctorate	1%	1%	1%
Income			
<$20k	19%	20%	15%
$20k-$40k	27%	32%	22%
$40k-$60k	24%	18%	30%
$60k-$80k	18%	16%	17%

$80k-$100k	5%	8%	7%
>$100k	7%	6%	9%

Table 1: Demographic information of the subjects from the image classification experiment (Fügener et al. 2021)

Appendix D: Regressions for the simulations

We report the results for the two sets of regressions that were required for the simulations in our empirical study. The first set of regressions captures the relationship between the expected performance of AI l^{AI} and that of humans P_n^H on the set of images that were used in the experiment. The second set of regressions captures the relationship between the expected performance of AI l^{AI} and that of augmentation P_n^{HAI}. We create linear regressions for each potential number of humans allocated to a task n so that we can estimate the performance for different crowd sizes of humans and of humans working with AI based on the ex-ante performance of AI. We create the two sets of regressions for the main study (which we discuss in Section 5.2) and for each of the six subgroups (which we discuss in Section 5.3). The respective regression coefficients are shown in Tables 3 to 9.

	$P_n^H \sim a_n + b_n \cdot l^{AI}$		$P_n^{HAI} \sim a_n + b_n \cdot l^{AI}$	
n	a_n	b_n	a_n	b_n
1	0.44	0.31	0.24	0.74
2	0.45	0.30	0.24	0.74
3	0.53	0.30	0.23	0.79
4	0.60	0.26	0.23	0.79
5	0.64	0.24	0.22	0.80
6	0.66	0.22	0.22	0.80
7	0.69	0.20	0.21	0.81
8	0.72	0.17	0.21	0.82
9	0.73	0.16	0.21	0.82
10	0.75	0.15	0.20	0.83

Table 2: Coefficient values for both sets of regressions (main study)

	$P_n^H \sim a_n + b_n \cdot l^{AI}$		$P_n^{HAI} \sim a_n + b_n \cdot l^{AI}$	
n	a_n	b_n	a_n	b_n
1	0.01	0.92	-0.03	1.07
2	0.03	0.90	-0.03	1.07
3	0.07	0.95	-0.06	1.13
4	0.13	0.90	-0.06	1.13
5	0.16	0.89	-0.08	1.15
6	0.19	0.86	-0.08	1.15
7	0.23	0.82	-0.09	1.16
8	0.26	0.79	-0.10	1.17
9	0.27	0.78	-0.10	1.17
10	0.30	0.75	-0.10	1.17

Table 3: Coefficient values for both sets of regressions (low between-task complementarity)

	$P_n^H \sim a_n + b_n \cdot l^{AI}$		$P_n^{HAI} \sim a_n + b_n \cdot l^{AI}$	
n	a_n	b_n	a_n	b_n
1	0.67	-0.14	0.35	0.55
2	0.67	-0.14	0.35	0.55
3	0.77	-0.17	0.33	0.64
4	0.84	-0.20	0.34	0.65
5	0.88	-0.21	0.33	0.66
6	0.90	-0.22	0.33	0.67
7	0.92	-0.23	0.32	0.68
8	0.94	-0.25	0.32	0.68
9	0.95	-0.25	0.31	0.69
10	0.97	-0.25	0.30	0.70

Table 4: Coefficient values for both sets of regressions (high between-task complementarity)

	$P_n^H \sim a_n + b_n \cdot l^{AI}$		$P_n^{HAI} \sim a_n + b_n \cdot l^{AI}$	
n	a_n	b_n	a_n	b_n
1	0.44	0.30	0.21	0.72
2	0.45	0.29	0.22	0.72
3	0.55	0.25	0.18	0.80
4	0.63	0.19	0.18	0.81
5	0.67	0.16	0.16	0.84
6	0.70	0.14	0.15	0.85
7	0.73	0.11	0.14	0.87
8	0.77	0.08	0.13	0.87
9	0.78	0.07	0.12	0.88
10	0.80	0.05	0.11	0.90

Table 5: Coefficient values for both sets of regressions (low within-task complementarity)

	$P_n^H \sim a_n + b_n \cdot l^{AI}$		$P_n^{HAI} \sim a_n + b_n \cdot l^{AI}$	
n	a_n	b_n	a_n	b_n
1	0.45	0.32	0.26	0.74
2	0.45	0.32	0.26	0.74
3	0.51	0.36	0.28	0.76
4	0.56	0.33	0.30	0.75
5	0.59	0.32	0.31	0.74
6	0.62	0.30	0.31	0.74
7	0.64	0.29	0.31	0.74
8	0.66	0.27	0.31	0.74
9	0.67	0.27	0.32	0.74

10	0.69	0.25	0.32	0.73

Table 6: Coefficient values for both sets of regressions (high within-task complementarity)

	$P_n^H \sim a_n + b_n \cdot l^{AI}$		$P_n^{HAI} \sim a_n + b_n \cdot l^{AI}$	
n	a_n	b_n	a_n	b_n
1	0.46	0.48	0.32	0.71
2	0.47	0.47	0.33	0.71
3	0.57	0.44	0.34	0.73
4	0.65	0.37	0.35	0.72
5	0.70	0.32	0.34	0.72
6	0.72	0.30	0.34	0.72
7	0.76	0.27	0.34	0.73
8	0.78	0.24	0.34	0.73
9	0.80	0.22	0.33	0.74
10	0.81	0.21	0.33	0.74

Table 7: Coefficient values for both sets of regressions (inferior AI)

	$P_n^H \sim a_n + b_n \cdot l^{AI}$		$P_n^{HAI} \sim a_n + b_n \cdot l^{AI}$	
n	a_n	b_n	a_n	b_n
1	-0.24	0.95	-0.57	1.58
2	-0.21	0.92	-0.57	1.58
3	-0.29	1.10	-0.66	1.74
4	-0.23	1.07	-0.69	1.78
5	-0.22	1.09	-0.71	1.81
6	-0.18	1.05	-0.73	1.83
7	-0.15	1.04	-0.74	1.84
8	-0.10	0.99	-0.75	1.86
9	-0.08	0.98	-0.77	1.87
10	-0.04	0.94	-0.77	1.88

Table 8: Coefficient values for both sets of regressions (superior AI)

Appendix E: Benefits of AI

In this appendix, we provide the results regarding the benefits that AI provides through automation and augmentation based on our empirical simulation. We report the results for our main analysis as well as for our subgroup analysis, in which we vary the levels of between-task complementarity, within-task complementarity, and relative performance. Here, benefit is defined as the improvement over the baseline case, in which each task is performed by one human.

The mean benefits of AI are displayed as percentage points in Table 9 (with standard errors in brackets). The first two columns show the benefits of the full-automation and full-augmentation benchmarks. The third column shows the benefits of our framework, where AI is used for both automation and augmentation. The remaining three columns decompose the framework benefit into substitution benefit, augmentation benefit, and reallocation benefit.

	Auto-mation	Augmen-tation	Framework: Total	Framework: Substi-tution	Framework: Augmen-tation	Framework: Reallo-cation
Main	8.89 (0.04)	11.71 (0.05)	19.51 (0.05)	16.92 (0.03)	-0.39 (0.02)	2.97 (0.03)
Low between	9.36 (0.05)	8.40 (0.06)	9.06 (0.06)	1.69 (0.03)	7.19 (0.04)	0.19 (0.02)
High between	8.34 (0.06)	15.12 (0.08)	24.01 (0.07)	22.91 (0.05)	-2.16 (0.03)	3.29 (0.04)
Low within	9.45 (0.06)	9.29 (0.07)	17.52 (0.07)	15.16 (0.05)	-1.13 (0.02)	3.48 (0.04)
High within	8.32 (0.06)	14.26 (0.07)	18.56 (0.07)	10.61 (0.04)	6.09 (0.04)	1.87 (0.03)
Inferior AI	-8.20 (0.05)	1.94 (0.07)	6.82 (0.07)	2.55 (0.02)	2.53 (0.03)	1.80 (0.03)
Superior AI	25.94 (0.06)	21.51 (0.07)	25.55 (0.06)	17.37 (0.05)	4.85 (0.03)	3.34 (0.04)

Table 9: Mean benefits of AI for the main analysis and the subgroup analysis (standard errors in brackets)

Appendix F: Robustness check for the empirical study

We replicate the simulation that we describe in Section 4.2 with subjects from Treatment 1 and Treatment 3. The difference from the original empirical study is that the subjects who work together with AI (augmentation) receive information about the certainty of the AI. This means that subjects can get an intuition about the task difficulty for the AI and the quality of the AI advice. The idea is that, in this way, humans have the opportunity to develop better within-task complementarity.

Data preparation

We start by determining the performance of crowds with sizes n varying from 1 to 10. We simulate the expected performance of humans by sampling from Treatment 1, and we simulate the expected performance of augmentation by sampling from Treatment 3. We then create two sets of linear regressions that capture the relationship between AI's expected performance l^{AI} and the performance of humans P^{H} (performance of augmentation P^{HAI}). The regression coefficients for the robustness check are shown in Table 11. We use the expected performance estimates of humans and augmentation based on the AI's expected performance as inputs for our task allocation framework and determine the optimal task allocation based on Equation (3). Based on the optimal task allocation, we then simulate the performance of our framework with 10,000 simulation runs. If a task is allocated to AI (automation), we use the realized AI performance for that task. If a task is allocated to n humans working alone, we sample n humans from Treatment 1 and determine their accuracy. If a task has been allocated to n humans working with AI (augmentation), we sample n humans from Treatment 3 and determine their accuracy. We also determine the performance of the full-automation and full-augmentation benchmarks. Furthermore, we decompose the benefit of our framework into substitution, augmentation and reallocation benefit.

	$P_n^H \sim a_n + b_n \cdot l^{AI}$		$P_n^{HAI} \sim a_n + b_n \cdot l^{AI}$	
n	a_n	b_n	a_n	b_n
1	0.44	0.31	0.27	0.69
2	0.45	0.30	0.27	0.69
3	0.53	0.30	0.30	0.71
4	0.60	0.26	0.33	0.68
5	0.64	0.24	0.34	0.68
6	0.66	0.22	0.35	0.67
7	0.69	0.20	0.35	0.67
8	0.72	0.17	0.35	0.66
9	0.73	0.16	0.36	0.66
10	0.75	0.15	0.36	0.66

Table 10: Coefficient values for both sets of regressions (robustness check)

Results

Since we also use subjects from Treatment 1 in this robustness check to simulate humans, the results for between-task complementarity are the same as those from the original study. To assess the effect on within-task complementarity, we compare the performance of subjects from Treatments 2 and 3 as well as their respective weights of correct advice r^{AI} and incorrect advice $r^{\overline{AI}}$.

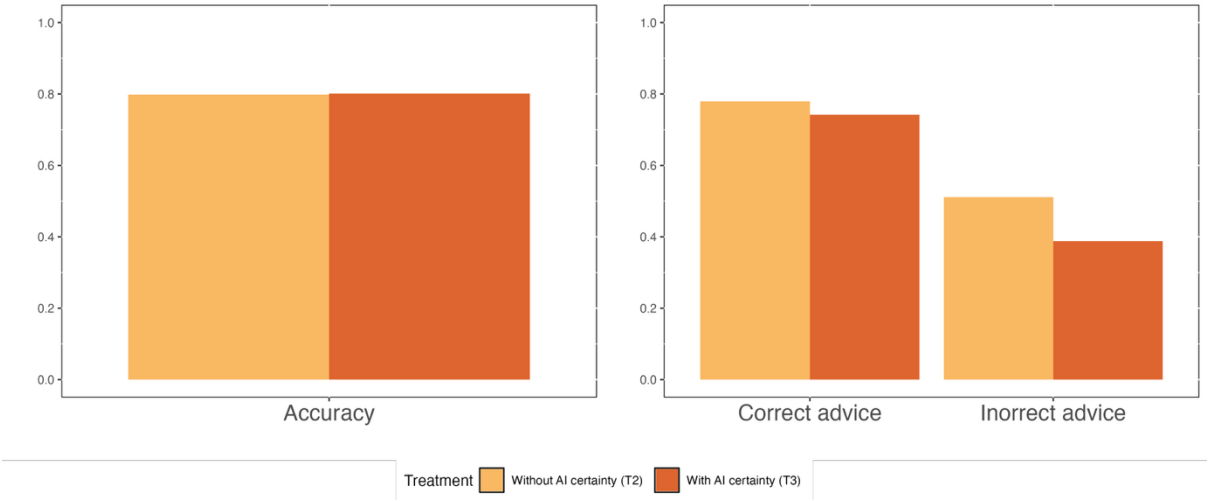

Figure 2: Comparison of subjects from Treatments 2 and 3. Left: accuracy. Right: weights of advice

Figure 2 shows the comparison of subjects from Treatments 2 and 3. The left plot shows the average classification accuracy of augmentation based on Treatments 2 and 3. While the augmentation based on Treatment 3 also achieved greater accuracy than did humans and automation, the difference to augmentation based on Treatment 2 is almost nonexistent. The right plot shows the weights that subjects from Treatments 2 and 3 placed on correct and incorrect advice. While the provision of AI certainty information helped subjects in Treatment 3 place less weight of incorrect advice, it also led to slightly less weight on correct advice. This finding shows that providing AI certainty seems to have no significant effect on the level of within-task complementarity between humans and AI.

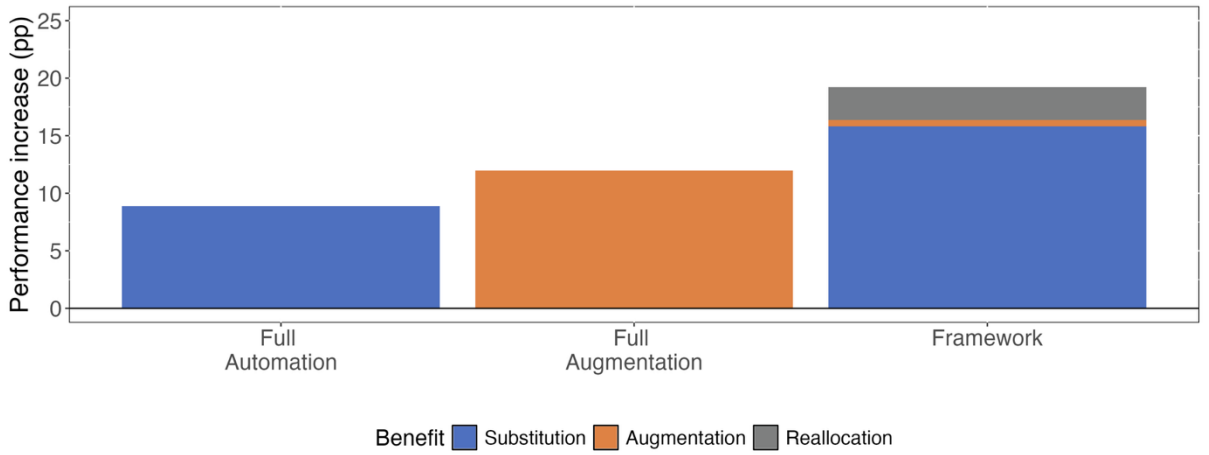

Figure 3: Benefits of automation and augmentation in the robustness check

Figure 3 illustrates the benefits that AI can provide through automation and augmentation for all 100 images. Again, the baseline case is one human working without AI on each task. The overall benefit does not differ from that in our main study. If anything, we observe a marginally higher, yet very small, benefit from augmentation. In general, we interpret these results as adding robustness to our main findings but do not see meaningful improvements from communicating AI certainty in human–AI augmentation with AI-advised decision making.

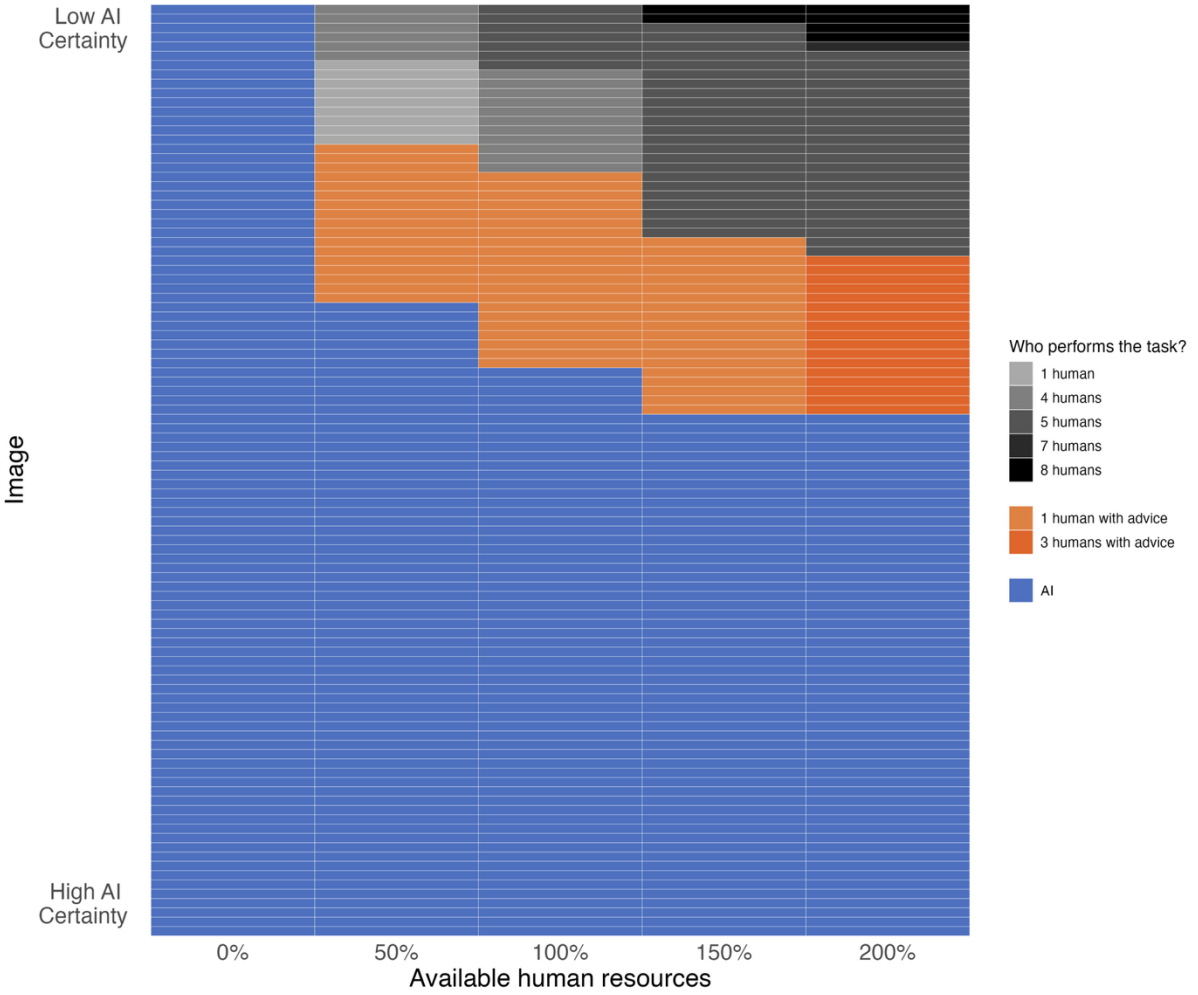

Figure 4: Distribution of work between humans and AI for different numbers of available human resources in the robustness check

Figure 4 shows the distribution of work for all 100 images sorted in ascending order based on AI certainty on the vertical axis (first row represents the most difficult image, and the last row represents the easiest image). On the horizontal axis, we report five different levels of human resources (0% to 200%). When comparing the results of the robustness check with the results from the original study, we observe only minimal changes in the distribution. For example, the results of the robustness check for the 100% resource level show that the 61 easiest images are performed by automation (original: 64), each of the next 21 images is performed by one human with AI advice (original: 16), and each of the remaining 18 images is performed by four to five humans (original: 20).