SUPPLEMENTARY INFORMATION

S1 Approach

S1.1 When is the ratio of means a useful measure of effect size?

The basic requirement for a ratio of means to be appropriate is that the X's be measured on a ratio scale [1] with a "real 0" that signifies the absence of whatever is being measured [2]. For instance, a measure of 0 for length, weight, or speed signifies the absence of those qualities, and an object that is 2 meters long is twice as long as one that is 1 meter long. But an object that has a temperature of 100^oF on the Fahrenheit scale is not, in any useful sense, twice as hot as one that is 50^oF. (This would be true, however, if the temperatures were measured on the Kelvin (K) scale which has an absolute 0.)

However, there are cases that are not strictly ratio scales within the Stevens taxonomy [1], but where a ratio of means measurement can still be useful. For example, proportions (including percentages and probabilities) don't satisfy Stevens's strict criteria for a ratio scale [3, 4]. One cannot, for instance, double a percentage without distorting its meaning. And when a student gets a 0 on a test, that doesn't necessarily mean that the student has no knowledge whatsoever of the subject to which the test pertains. But it is still often useful to analyze ratios of means for proportions, and we do so here.

In fact, it may even be useful to analyze ordinal data (such as Likert scales) with a ratio of means. For instance, Stevens himself, who formulated the widely used taxonomy of scale types, said, "As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety, the ordinary statistics involving means and standard deviations ought not to be used with these scales... On the other hand, ... there can be invoked a kind of pragmatic sanction: in numerous instances it leads to fruitful results" [5, 4]. Examples of treating data from ordinal scales as ratio scales in this way include [6] and [7], and we do so here.

Another factor to consider in judging whether analyzing the ratio of means is appropriate in a given situation is whether the underlying process being measured can be usefully modeled as a multiplicative, rather than additive, process. For example, in Sec. S3.6 below, we provide a simple and very general model of performance as a multiplicative function of factors such as the ability of subjects and the difficulty of tasks.

S1.2 Metric transformations

S1.2.1 Desirable lower bound

The full version of the desirable lower bound transformation is

$$f(X) = \frac{1}{X - X_{min}}$$

where the lower bound of X_{min} is usually 0. In cases where the lower bound is not 0, this transformation automatically rescales the original metric to have a lower bound of 0. And the transformed metric has a lower bound of 0 as X approaches ∞.

As noted in the main text, this transformation is useful to make metrics where lower values are more desirable comparable to ones where higher values are better. It also has a side effect of emphasizing small improvements as the lower bound is approached (see Sec. S1.2.3 below).

S1.2.2 Desirable upper bound

To derive the desirable upper bound transformation, we begin by rescaling X so that its transformed upper bound is 1:

$$X' = f(X) = \frac{X}{X_{max}}$$

where X_{max} is the original upper bound. Then we further transform the resulting value with an equation that is symmetrical to the one for the desirable lower bound transformation. In other words, we take the reciprocal of the distance to the bound:

$$f'(X') = \frac{1}{1 - X'} \ .$$

But this transformation has the disadvantage that its minimum when X' = 0 is 1, which means that the transformed values would not have a "real 0." So to correct for this, we use a different transformation which subtracts 1:

$$\begin{aligned} f''(X') &= \tfrac{1}{1-X'} - 1 \\ &= \tfrac{1}{1-X'} - \tfrac{1-X'}{1-X'} \\ &= \tfrac{X'}{1-X'}. \end{aligned}$$

Interestingly, this transformation is equivalent to the calculation of odds used in calculating an *odds ratio*. That means that the ratio of means effect size measurement with this transformation is equivalent to an effect size measurement with an odds ratio [8].

Effect of the desirable upper bound transformation. Another interesting feature of the desirable upper bound transformation is that it always makes the ρ and $\hat{\rho}$ ratios more extreme, that is, further from 1.

To see why analytically, first note that the transformation does not change which element in the denominator of $\hat{\rho}$ will be the maximum because the transformation does not change the inequality relationship between any two variables a and b.

To see this, let

$$a' = \tfrac{a}{1-a}$$
$$b' = \tfrac{b}{1-b}$$

Then, if $a > b$, the numerator of a' is greater than the numerator of b', and the denominator of a' is less than the denominator of b', so $a' > b'$. Similar reasoning applies for $a < b$, and $a = b$.

Now to see that the transformation increases the distance of a ρ ratio from 1, let

$$\rho = \tfrac{x}{y} = \textit{ratio of the untransformed values}$$

$$\rho' = \tfrac{\frac{x}{1-x}}{\frac{y}{1-y}} = \textit{ratio of the transformed values}$$

$$R = \tfrac{\rho'}{\rho} = \tfrac{\frac{1}{1-x}}{\frac{1}{1-y}} = \tfrac{1-y}{1-x}$$

Thus, if $x > y$, then $R > 1$, and the ratio of the transformed values is greater than the ratio of the untransformed ones (i.e., further above 1). The opposite is true when $x < y$. And if $x = y$, then $R = 1$, and the ratio of the transformed values equals the ratio of the untransformed ones.

S1.2.3 Emphasizing small improvements near a bound

A potential benefit of both the upper and lower bound transformations is that they *emphasize small improvements as a metric approaches its desirable limit*. This can be especially useful when approaching a limit is both increasingly difficult and increasingly desirable.

For example, if X represents the proportion of lives saved by a medical treatment, the fraction of airplane accidents prevented, or the number of costly failures in critical infrastructures (such as computer servers), then it can be both extremely difficult and extremely valuable to approach the limits in these situations.

In other situations, however, small changes near a limit may have no particular value. For instance, in producing a product with limited demand (such as a customized report), the benefit of reducing the cost from $0.01 to $0.001 may be irrelevant even though it produces a factor of 10 improvement according to the transformed metric.

S2 Study 1

S2.1 Systematic review and analysis

S2.1.1 Search strategy

To select the studies for our analysis, we followed the procedures of a systematic review [9, 10]. As such, we began by developing the search string and distilling the facets of studies that evaluated the performance of a human-computer system. We required the following: (1) a human component, (2) a computer component, (3) a collaboration component, and (4) an experiment component. Given the multidisciplinary nature of our research question, papers published in different venues tended to refer to these components under various names [11, 12]. To ensure comprehensive coverage, we compiled a list of synonyms and abbreviations for each component and then combined these terms with Boolean operations, resulting in the following search string:

1. Human agent: "human" OR "expert" OR "participants"
2. AI agent: "AI" OR "artificial intelligence" OR "ML", or "machine learning"
3. Collaboration: "collaboration" OR "teamwork" OR "combination" OR "complement" OR "synergy" OR "assist" OR "aid" OR "hybrid" OR "loop"
4. Experiment: "experiment" OR "user study" OR "user evaluation" OR "empirical study"

Search term: (1) AND (2) AND (3) AND (4)

We performed this search in the Association for Computing Machinery Digital Library (ACM DL), Institute of Electrical and Electronics Engineers Xplore Digital Library (IEEE Xplore), Association for Information Systems eLibrary (AISeL), Web of Science, and arXiv to encompass the major computer science, engineering, information systems, and social science conferences and journals. To focus on current forms of artificial intelligence, we limited the search to studies published in 2021.

S2.1.2 Selection criteria

Next, we applied the following criteria to select studies that fit our research questions. Specifically, the paper needed to include an original experiment that evaluated some instance in which a computer helped a human perform a task, and it needed to report the performance of (1) the human alone, (2) the computer alone, and (3) the human-computer team according to some quantitative measure(s). As such, we excluded meta and literature reviews, commentaries, opinions, and simulations. Furthermore, we required the study to be written in English.

Through these selection criteria, we identified 20 relevant papers from the initial search. Among this set, we performed a forward and backward search, which found 5 new studies to include in our final review. Notably, most of these papers reported the results of multiple experiments, so our final list included 25 unique papers and 79 unique experimental measurements.

S2.1.3 Data extraction

To address our research questions, we extracted the performance measure(s) for each experiment in each study and recorded the performance of the human alone (X_H), the computer alone (X_C), and the human-computer team (X_{HC}) according to the measure(s) reported in the paper. Note that if the authors assessed performance in a single experiment according to multiple metrics (e.g., specificity and sensitivity), then we included all the different metrics as separate entries in our data. For each metric, we also calculated $\hat{\rho}$, which included the desirable lower bound transformation wherever applicable, and $\hat{\rho}'$, which included both the desirable lower bound and upper bound transformations wherever they were applicable. Table A1 summarizes our results.

Study	Task	Measure	X_H	X_C	X_{HC}	$\hat{\rho}$	$\hat{\rho}'$
[13]	Answer questions	Accuracy	57%	50%	78%	1.36	2.59
[13]	Answer questions	Accuracy	57%	50%	76%	1.32	2.32
[13]	Answer questions	Accuracy	57%	50%	75%	1.31	2.21
[13]	Answer questions	Accuracy	57%	50%	71%	1.25	1.86
[13]	Answer questions	Accuracy	57%	50%	70%	1.23	1.78
[13]	Answer questions	Accuracy	57%	50%	68%	1.19	1.60
[14]	Teach students math	Accuracy	35%	51%	60%	1.18	1.44

[15]	Detect objects in an image	Recall	77%	64%	89%	1.17	2.55
[16]	Label visible car scratches	Recall	51%	43%	59%	1.16	1.40
[16]	Label visible car scratches	Precision	53%	53%	61%	1.15	1.40
[15]	Detect objects in an image	Recall	77%	64%	86%	1.12	1.82
[17]	Detect deepfakes	Accuracy	66%	65%	73%	1.11	1.39
[18]	Label data	Accuracy	72%	44%	79%	1.10	1.46
[18]	Label data	Accuracy	72%	59%	78%	1.08	1.38
[19]	Classify conservation images	Accuracy	58%	50%	62%	1.06	1.16
[15]	Detect objects in an image	Recall	77%	64%	80%	1.05	1.24
[20]	Classify images	Accuracy	68%	77%	80%	1.04	1.20
[20]	Classify images	Accuracy	68%	77%	80%	1.04	1.19
[19]	Classify conservation images	Accuracy	58%	50%	60%	1.03	1.09
[21]	Classify endoscopic images	Sensitivity	93%	89%	96%	1.03	1.75
[21]	Classify endoscopic images	Sensitivity	93%	89%	95%	1.02	1.40
[22]	Decide if passenger survived	Accuracy	72%	75%	77%	1.02	1.09
[22]	Decide if passenger survived	Accuracy	72%	75%	77%	1.02	1.09
[22]	Decide if passenger survived	Accuracy	72%	75%	76%	1.02	1.07
[22]	Play kitchen game	Steps	37.82	34	38.37	1.01	0.89
[22]	Decide if passenger survived	Accuracy	72%	75%	76%	1.01	1.06
[23]	Classify programming problems	F1	54%	86%	87%	1.01	1.09
[24]	Answer open questions	Quality	3.74	3.68	3.77	1.01	1.03
[25]	Play kitchen game	Steps	23.86	20	24.04	1.01	0.83
[20]	Classify images	Accuracy	72%	77%	78%	1.01	1.03
[26]	Predict violation of pretrial	Accuracy	56%	55%	56%	1.01	1.01
[24]	Answer open questions	Quality	3.74	3.87	3.89	1.01	1.02
[26]	Predict recidivism	Accuracy	55%	56%	56%	1.00	1.01
[26]	Predict recidivism	Accuracy	55%	56%	56%	1.00	1.00
[27]	Segment heart images	Dice	0.92	0.92	0.92	1.00	1.00
[20]	Classify images	Accuracy	72%	77%	77%	1.00	0.98
[26]	Predict violation of pretrial	Accuracy	56%	55%	56%	0.99	0.99
[15]	Detect objects in an image	Precision	98%	86%	97%	0.99	0.78
[15]	Detect objects in an image	Precision	98%	86%	97%	0.99	0.70
[19]	Classify conservation images	Accuracy	58%	50%	58%	0.99	0.98
[28]	Annotate image	Accuracy	95%	88%	94%	0.99	0.79
[29]	Find true statements	AUC	52%	75%	74%	0.99	0.95
[28]	Annotate image	Accuracy	95%	88%	94%	0.99	0.76
[25]	Play kitchen game	Steps	37.82	34	37.14	0.98	0.92
[15]	Detect objects in an image	Precision	98%	86%	96%	0.98	0.54
[21]	Classify endoscopic images	Specificity	85%	91%	89%	0.98	0.82
[29]	Find true statements	Accuracy	50%	69%	67%	0.97	0.91
[30]	Annotate clinical text	Recall	76%	83%	80%	0.96	0.82
[25]	Play kitchen game	Steps	23.86	20	22.99	0.96	0.87
[20]	Classify images	Accuracy	72%	77%	74%	0.96	0.85
[26]	Predict person's profession	Accuracy	68%	77%	74%	0.96	0.85
[26]	Predict person's profession	Accuracy	68%	77%	73%	0.95	0.82
[31]	Predict risk of fraud	Accuracy	62%	57%	59%	0.95	0.87
[32]	Mouselab-MDP	Score	28.41	39.692	36.8	0.93	0.09
[24]	Answer open questions	Quality	3.74	3.1	3.45	0.92	0.75
[30]	Annotate clinical text	Recall	76%	83%	76%	0.92	0.65
[31]	Predict risk of fraud	Accuracy	62%	57%	57%	0.91	0.80
[26]	Predict recidivism	Accuracy	60%	66%	59%	0.91	0.77
[26]	Predict violation of pretrial	Accuracy	61%	68%	62%	0.90	0.74
[24]	Answer open questions	Quality	3.74	2.84	3.35	0.90	0.68
[26]	Predict recidivism	Accuracy	60%	66%	59%	0.89	0.75
[21]	Classify endoscopic images	Specificity	65%	91%	81%	0.89	0.47
[26]	Predict violation of pretrial	Accuracy	61%	68%	61%	0.89	0.71
[33]	Diagnose dermatologic conditions	Accuracy	46%	66%	58%	0.88	0.71
[34]	Classify recyclables	Accuracy	86	94.50	82.5	0.87	1.00
[26]	Predict person's profession	Accuracy	64%	84%	73%	0.87	0.52
[26]	Predict person's profession	Accuracy	64%	84%	72%	0.86	0.50
[32]	Mouselab-MDP	Score	4.17	6.97	5.95	0.85	0.59
[32]	Mouselab-MDP	Score	22.85	29.77	24.54	0.82	0.03
[31]	Predict risk of fraud	Accuracy	62%	57%	51%	0.82	0.64
[34]	Classify recyclables	Accuracy	86	55.00	68	0.79	1.00
[29]	Find true statements	Accuracy	50%	69%	51%	0.74	0.47

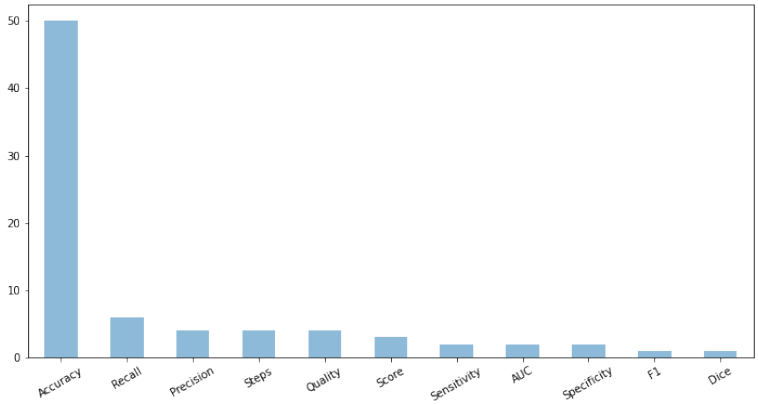


Figure S1: The distribution of performance measures for the experimental results in our review.

[29]	Find true statements	AUC	52%	75%	49%	0.65	0.32
[35]	Prescribe drug	Accuracy	36%	67%	37%	0.55	0.29
[35]	Prescribe drug	Accuracy	36%	67%	36%	0.54	0.28
[35]	Prescribe drug	Accuracy	36%	67%	35%	0.53	0.27
[35]	Prescribe drug	Accuracy	36%	67%	34%	0.51	0.25
[36]	Edit food ingredients	Accuracy	17%	75%	35%	0.47	0.18
[36]	Edit food ingredients	Accuracy	17%	75%	33%	0.44	0.16

Table S1: Summary of studies in review.

S2.1.4. Is this a meta-analysis?

The goal of a standard meta-analysis is to estimate the "true" value of an underlying parameter (such as the beneficial effect of a particular drug) by calculating a weighted average of different measurements of that parameter from different studies [37]. In that sense, our study is not a standard meta-analysis because we did not attempt to construct a weighted average of the results from different studies. Instead, we did a very simple kind of meta-analysis in which we summarized the ratios from all the different studies by displaying their distribution and by reporting their median and (unweighted) mean.

We did not attempt to compute weightings of the different studies because the studies we analyzed covered many different combinations of tasks, subjects, and technologies, and each combination could be expected to have a different "true" value for ρ. Thus any attempt to estimate a single underlying parameter for them all would be roughly analogous to trying to estimate the average weight of all the animals on earth—from tiny micro-organisms to large mammals. Even if one attempted to do this, it's not clear whether one should weight different measurements based on (a) the statistical uncertainties in each measurement (as is common in standard meta-analyses), (b) the estimated number of animals of each type on earth, or (c) some other weightings. Furthermore, even if it were appropriate to estimate a single underlying value using the statistical techniques common in meta-analyses, most of the studies we analyzed did not include enough data to calculate such weighting statistics.

We do still believe, however, that this simple way of displaying and summarizing the ratios from recent studies provides novel and informative insights that set the stage for many kinds of future research.

S2.1.5 Robustness checks

Effect of including multiple experimental measurements per paper. In our main results, we include all experimental results from every study in our analysis. Here, we also perform the analysis on only one experimental result per study, selecting the result that leads to the highest ratio (Figure S2). By construction, among this subset of experiments, a larger proportion report findings of human-computer synergy according to the definition of a performance ratio greater than one (56% versus 38%), and the average and median ratios are slightly higher (0.98 versus 0.96 and 1.01 versus 0.99, respectively). As such, our main findings still hold – a surprisingly small number of studies find evidence of human-computer synergy, and those that do find surprisingly small performance improvements.

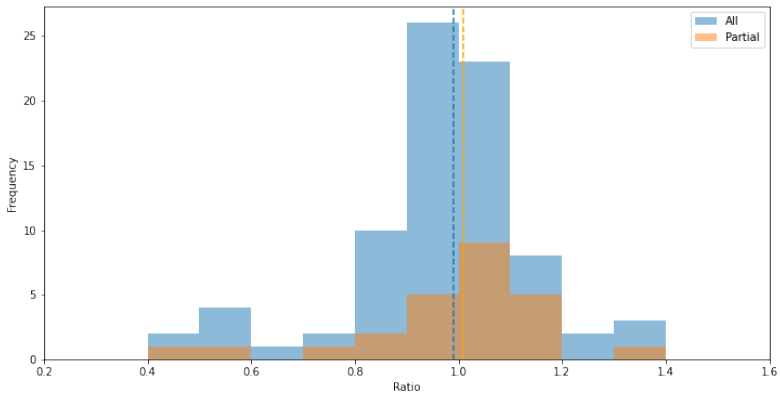
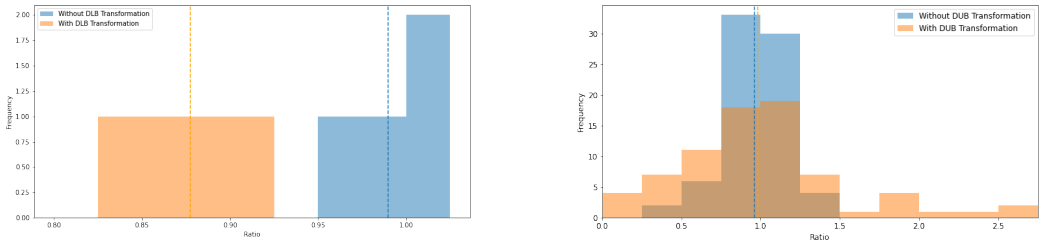


Figure S2: The distribution of performance ratio values from all experimental results ($n = 79$) and the top result per study ($n = 25$). The dotted lines correspond to the mean of the ratios.

Effect of metric transformations. To evaluate the effect of the metric transformations, we separately looked at the distribution of ratios for measures in which smaller values indicate better performance (desirable lower bound) and the distribution of ratios for measures in which larger values indicate better performance (desirable upper bound).

Only four experimental results in our review involve a metric with a desirable lower bound, namely the number of steps required to complete a task (see Figure S3a). In this case, the transformation serves three functions: (1) it inverts the measure such that larger values correspond to better performance, (2) it changes which of the baselines in the denominator (human only or computer only) is selected for comparison, and (3) it rescales the measure so that small differences near the lower bound of 0 have a larger effect. Since the computer outperforms the human-computer team in all these experiments, the transformed ratios all fall below one, and the changes in magnitude from the untransformed values are relatively small.

The remaining experimental results all pertain to metrics with desirable upper bounds. As noted in Sec. S1.2, the effect of this transformation is to rescale the ratios to be further from 1, and this effect is visible in Figure S3b. As such, among this set, the percentage of experiments that exhibit human-computer synergy as defined by a ratio greater than one remains unchanged. While the transformation widens the distribution of rho values, the median stays relatively the same (0.92 with versus 0.96 without the transformations) as does the mean (0.98 with versus 0.99 without the transformations).



(a) The distribution of ratios from experimental results with desirable lower bounds ($n = 4$).

(b) The distribution of performance ratios from experimental results with desirable upper bounds ($n = 75$).

Figure S3: The distribution of performance ratios from experimental results with and without transformations. The dotted lines correspond to the mean of the ratios.

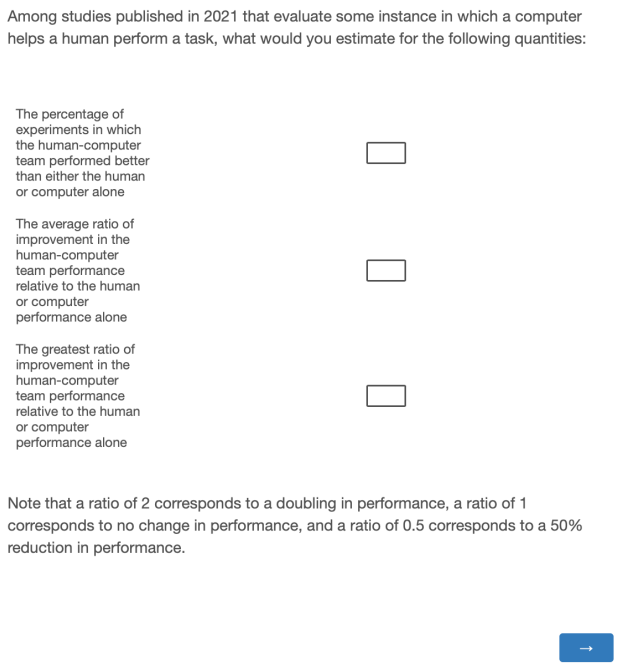


Figure S4: Screenshot of the survey page with questions for the participants.

S2.2 Survey

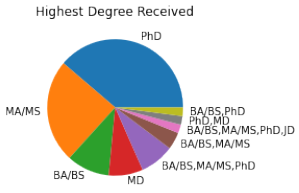
To contextualize the findings from our review, we conducted an online survey with researchers actively pursuing empirical studies of human-computer teams. In this section, we detail the survey design and evaluation, which follow the guidelines presented in [38, 39].

S2.2.1 Participant selection

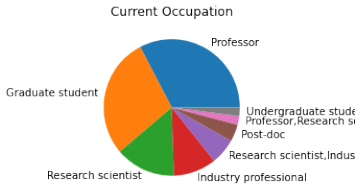
To target researchers with deep knowledge of human-computer teams, we sent the survey to the authors of the papers included in our review as well as authors from 2 additional papers that were later removed from our analysis because they did not include performance data for computers alone. Since this set only includes studies published in the past year (2021), we expect these people to possess knowledge of the current state of the field. Note that other studies such as [40] and [41] perform similar strategies to elicit expectations from practitioners and researchers. We sent emails to each author ($n = 133$), but 11 addresses were no longer active, so we only delivered 122 invitations to complete the survey. Additionally, we sent a friendly reminder about the survey after one week and two weeks. After three weeks, we closed the survey. We incentivized participation with a raffle for a $200 Amazon gift card.

S2.2.2 Survey design

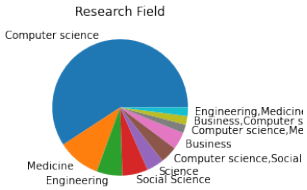
The first block of the survey presented a consent form with information about the context of the study, expected completion time, and related matters. In the next block, subjects answered the questions shown in Figure S4. We asked for their responses via text boxes to mitigate potential anchoring effects [42], and we only allowed them to submit numeric answers. The final block of questions collected demographic information such as level of education, occupation, and general research discipline. We made all questions optional, and we implemented the survey using Qualtrics.



(a) Participant highest level of education.

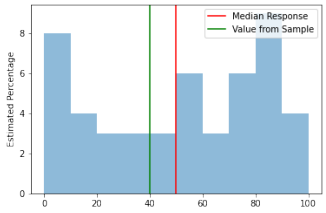


(b) Participant occupation.

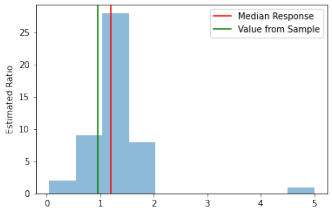


(c) Participant research area.

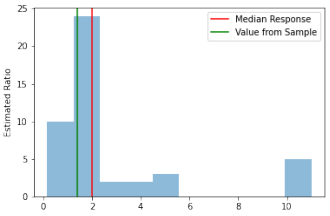
Figure S5: Participant demographic information.



(a) The percentage of experiments in which the human-computer team performed better than either the human or computer alone.



(b) The average ratio of improvement in the human-computer team performance relative to the human or computer performance alone.



(c) The greatest ratio of improvement in the human-computer team performance relative to the human or computer performance alone.

Figure S6: The distribution of participant responses to the questions in our survey. Note that for readability purposes, (b) does not display one participant's estimate of 20, and (c) does not display two participants' estimates of 50 and 100.

S2.2.3 Participant demographics

A total of 49 people completed the survey for a response rate of 40%. Figure S5 provides an overview of the demographics of the participants. Most respondents were currently pursuing or had already received a graduate degree, and most specialized in computer science.

S2.2.4 Survey results

Figure S6 illustrates the distribution of the responses to the questions in the survey. Given the number of extreme outliers in the responses, we highlight the median and interquartile ranges (IQRs) as opposed to the mean and standard deviations of the estimates for each question [43]. When asked to consider studies published in 2021 that evaluate some instance in which a computer helps a human perform a task, the median estimate of the percentage of experiments in which the human-computer team performed better than either the human or computer alone was 50% (20 - 80% IQR). The median estimate of the average ratio of improvement was 1.2x (1.1 - 1.5x IQR) versus our finding of 0.96, and the median estimate of the maximum ratio of improvement was 2.0x (1.5 - 4.0x IQR) versus our finding of 1.4.

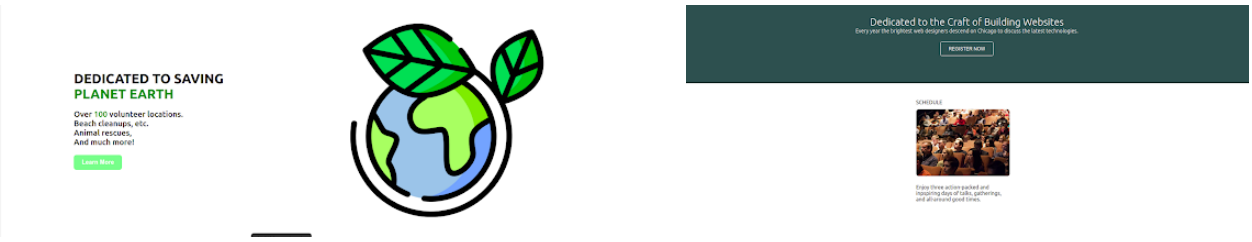
S3 Studies 2 and 3

S3.1 Web pages

To represent the space of possible web pages, we used two sample pages that varied in difficulty:

1. "Saving Planet Earth" (Fig. S7a) was designed by us and contains 27 components used in scoring the correctness of the code generated.
2. "Conference Website" (Fig. S7b) was adapted and simplified from a task given towards the end of an HTML course called Learn to Code (`https://learn.shayhowe.com/practice/adding-media/index.html`) and contains only 13 components.

In addition to the visual images shown in the figures, each image also included pop-up messages to specify the details of active elements such as buttons and links. Subjects could see these messages by rolling their mouse over the images.



(a) Saving Planet Earth, with 27 items.

(b) Conference Website, with 13 items.

Figure S7: Webpages used for experiment.

S3.2 GPT-3

S3.2.1 GPT-3 description

GPT-3 stands for "Generative Pre-trained Transformer 3" [44]. It is a massive AI system that uses autoregressive machine learning techniques to generate many kinds of text. With over 175 billion parameters, it was extensively trained on a vast corpus of text including all of Wikipedia, many books, and much more material from the Internet. It has outperformed existing state-of-the-art models in many benchmarks. Particularly impressive and relevant is its performance in the few-shot domain, where the model is conditioned to a specific task by providing only a few examples. As noted in the main text, there is now a newer version of the GPT system, GPT-4, which was released after our study was completed [45].

S3.2.2 Use of GPT-3 in these studies

We developed special software for our subjects to use as an interface to GPT-3. This software uses the OpenAI API to interface with GPT-3, and it provides two primary additional features:

1. *Parameters and prompts.* To perform specific tasks, GPT-3 needs some detailed parameters and (usually) some examples to indicate what kinds of text outputs the system should generate in response to various kinds of inputs. See Sec. S3.2.3 for details of the parameters and examples we provided.
2. *User interface.* The user interface for GPT-3 that was used by subjects in both studies is shown in Figure 3. It works as follows: First, the human subject describes in English text a component needed in the web page (e.g., "Green button that says 'Learn More'" as shown in the upper left corner of Figure 3). Then the GPT-3 system automatically generates the HTML code needed to produce that component (e.g., the code shown in the lower left corner of Figure 3). And, simultaneously, the interface we created adds that component to the replica of the web page shown on the right side of the screen (e.g., the green button shown in the middle of Figure 3). If the human thinks the displayed component is correct, the human can then move that component to its proper location in the replica of the web page. If not, the human can (a) delete the component and try again to describe in English what the component should be, or (b) modify the HTML code directly (if the human knows enough to do so).

S3.2.3. GPT-3 Parameters and Prompts

For all our experiments we used the Open-AI API with the Davinci engine, a temperature of 0.3 and a max_token length of 512.

Conditioning GPT-3 is conditioned by prompting where each prompt includes a natural language description that the user would type followed by the HTML that would be returned by GPT-3:

- "Hello world button"
  <button style="padding: 12px 24px; border: 1px solid lightgray; background-color: whitesmoke; border-radius: 4px; font-size: 14px; color: black; font-weight: normal;">Hello World</button>
- "pink button that says Banana with white text color"
  <button style="padding: 12px 24px; border: 1px solid lightgray; background-color: pink; border-radius: 4px; font-size: 12px; color: white; font-weight: normal;">Banana</button>
- "Large Google Pixel button with no background"
  <button style="padding: 12px 24px; border: 1px solid lightgray; background-color: transparent; border-radius: 4px; font-size: 24px; color: black; font-weight: normal;">Google Pixel</button>

- “red square”

- “orange rectangle”

- “paragraph that says I like apples”
  I like apples
- “large text that says Black Circle in blue”
  Black Circle
- “Air-Conditioner in brown color”
  Air-Conditioner
- “My birthday is coming up soon!”
  My birthday is coming up soon!
- “text that says Google and links to https://www.google.com/”
  Google
- “FACEBOOK and links to https://www.facebook.com/”
  FACEBOOK
- “image from https://picsum.photos/id/237/200/300”
  <img src="https://picsum.photos/id/237/200/300" style="width: 200px; object-fit: contain;"></img>
- “ ”

- “<input type="submit" value="Submit!"></input>”
  <input type="submit" value="Submit!"></input>
- “table that says Fruits, Apple, Lime, Blueberries in the first column and Color, Red, Green, Blue in the second column”
  <table style="width: 250px; border-collapse: collapse; border: 1px solid black; text-align: center;"><tr><th>Fruits</th><th>Color</th></tr><tr><td>Apple</td><td>Red</td></tr><tr><td>Lime</td><td>Green</td></tr><tr><td>Blueberries</td><td>Blue</td></tr> </table>

S3.3 Subjects

S3.3.1 Recruiting

Subjects were recruited from the online labor market UpWork.com. For the posting requesting programmers (“coders”), 100 subjects were recruited, and for the one requesting non-programmers (“non-coders”), 50 were recruited.

The following postings were used to recruit subjects:

Non-Coders Needed For Web Experiment (Open to all with no coding background)

> “We are MIT researchers doing experiments on tools that help create websites. We are looking for non-coders that can help us test our interface. If you are a coder, look for our other job here `https://www.upwork.com/ab/applicants/1425936000678547456/job-details`.
>
> Your job will be to replicate a couple of website mockups. We estimate this would take 45 mins - 1.5hrs of your time for the whole job. We might also want to discuss your experience during the task to improve it.
>
> If you perform well on this job, we will prioritize you for similar future projects.
>
> To be a good fit for this project you should have no experience with coding (If you code, you should look and will do better by doing the “Website Mockup Replication in HTML (For Coders)” experiment).
>
> You will receive a base pay of $10 dollars and will receive a bonus of $10 dollars for each problem you get correctly amounting to up to $30 dollars for the two problems
>
> You will complete this job using the special interface we provide. The link to the interface will be sent to you once we have accepted you for the job.
>
> If you are interested in this project, please submit a proposal.”

Frontend Web Developer Needed For Experiment

> "We are MIT researchers doing experiments on tools that help create websites. We are looking for coders with experience in HTML and CSS to test our interface.
>
> Your job will be to replicate a couple of website mockups. We estimate this would take 45 mins - 1.5 hrs of your time for the whole job. We may also want to discuss your experience with the task to improve it.
>
> If you perform well on this job, we will prioritize you for similar future projects.
>
> To be a good fit for this project you should have experience coding in HTML and CSS.
>
> You will receive a base pay of $10 dollars and will receive a bonus of $10 dollars for each problem you get correctly amounting to up to $30 dollars
>
> You will complete this job using the special interface we provide. The link to the interface will be sent to you once we have accepted you for the job.
>
> If you are interested in this project, please submit a proposal."

S3.3.2 Screening

The status of subjects as programmers or non-programmers was confirmed by using text message conversations with the subjects and by reviewing their resumes and biographical information on UpWork. The software interface used in the experiment also asked the subjects to self-classify as a programmer ("coder") or non-programmer ("non-coder"). Subjects who did not self-classify for the group they were recruited for were not included in the study.

S3.3.3 Instructions and incentives

All subjects were provided with the following instructions:

"In this experiment, you are asked to solve two different problems. The main goal of each problem is to try and replicate a mockup website as accurately as possible. You can receive up to $30 for participating in this experiment.

- The base amount will be $10 for completing both problems.
- You will receive an additional bonus of up to $10 dollars per problem, conditional on getting it right and depending on how fast you do it. Before starting the actual problem you will have some time with a practice mockup to get familiar with the interface, so that you don't lose your time in the actual problem when it counts towards the amount of bonus you get (the faster you do it the bigger the bonus)".

Training Session. Subjects were encouraged to practice using a simple example web page they could try to replicate with unlimited time. The performance on this training session was not evaluated. Subjects were additionally provided with videos of how to interact with the interface.

Payment. Each subject received a base payment of $10 for participating in the experiment. They received no bonus for incorrect submissions. For correct submissions, their bonus was determined according to the following rule:

- $10.00 if done more than 10 minutes faster than the average
- $7.50 if done between the average and 10 minutes faster than the average
- $5.00 if done in a time between the average and 10 minutes slower than the average.
- $2.50 if done more than 10 minutes slower than the average

Note that this detailed payment rule was not provided to the subjects. They received only the instructions specified above.

S3.4 Scoring quality of code generated

S3.4.1 Evaluating the correctness of code generated

The correctness of the code generated for each page was scored based on the total number of components in the page. A component was defined as one individual HTML element (e.g., a button or a link), and each component was worth a maximum of 4 points, one point each for:

- *position* relative to the overall frame of the webpage (0.5 points if the position is slightly off).
- *content*, text or graphical (0.5 points if mostly correct, e.g., if the correct text says "Over 100 locations" and the subject said "100 locations").
- *color*, that is, text color, background color, button color, etc.
- *functionality*, that is, what the component does (i.e., text just displays text, links should link to external pages).

Since each component was worth a maximum of 4 points, the image in Fig. S7a was worth 27x4 = 108 points, and the one in Fig. S7b was worth 13x4 = 52 points. We consider a submission "correct" if it receives at least 90% of the maximum number of points available for it.

All the code generated was evaluated by the same person for consistency. A second person rated 10% of the code submissions to measure inter-rater reliability. As expected, following the rubric, both raters had an almost perfect overlap never differing by more than one point out of the 52 and 108 total points per web page.

S3.4.2 Quality threshold

For accuracy, we calculate the proportion of submissions which are correct as defined above, and we expect at least 80% of submissions in a condition to be "correct."

S3.4.3 Level of effort

For all analyses, we excluded any subjects who didn't put in effort, defined as spending less than 10 minutes and using less than 7 calls to GPT-3 for the "human-computer" ("HC") condition; and as spending less than 10 minutes and generating less than 15 lines of code for the "human-only" ("H") condition. These are reasonable numbers below which the task cannot be performed.

These criteria led to excluding 3 out of 100 subjects in Study 2 (2 in condition HC, 1 in condition C), and 2 out of 50 subjects in Study 3.

S3.5 Robustness Checks

S3.5.1 Various methods of calculating confidence intervals for ratios of means

[46] give several methods for calculating confidence intervals for ratios of means. In all cases, we report in the main text the results of using the method Bonett and Price recommend most, but as Table S2 shows, the results of using all these different methods with our data are very similar.

Table S2: Alternative methods of calculating 95% confidence intervals for the ratio of means.

Study	Ratio	$\hat{\rho}$	Method		
			Fieller	Delta	**Recommended**
Study 2					
Speed for programmers with and without GPT-3 (Paired-Samples)	$\frac{X_{HC}}{X_H}$	1.17	[1.04, 1.32]	[1.03, 1.31]	**[1.04, 1.32]**
Study 3					
Speed for programmers and non-programmers (Independent-Samples)	$\frac{X_{HC}}{X_{HC'}}$	1.07	[0.67, 1.50]	[0.92, 1.22]	**[0.93, 1.24]**
Cost for programmers with and without GPT-3 (Paired-Samples)	$\frac{C_{HC}}{C_H}$	0.97	[0.88, 1.12]	[0.88, 1.12]	**[0.87, 1.11]**
Cost for programmers and non-programmers (Independent-Samples)	$\frac{C_{HC}}{C_{HC'}}$	1.20	[0.52, 1.94]	[0.96, 1.44]	**[0.98, 1.46]**

S3.5.2 Regression for only programmers that passed

For robustness, we run the same regression only for the population of programmers that were successful, defined as obtaining at least 90% of the total possible points on the task. This regression has 169 observations. Table S3 shows the results. The exponential of the condition coefficient remained similar improving to 1.33 with a confidence interval of [1.14, 1.55].

S3.5.3 Removing Subject Random Effects

We analyze a regression without random effects for the population of programmers using a standard ordinary least squares in Table S4. The regression has 197 observations from 100 programmers. We see that the exponential of the

Table S3: Study 2 Robustness Regression. Only for observations that had a score of 90%+.

Effect	Estimate	SE	p	95% CI LL	95% CI UL	$e^{Estimate}$
Intercept	3.632	0.068	0.000	3.498	3.765	37.788
Condition	0.287	0.078	0.000	0.134	0.439	1.332
Task	-0.030	0.069	0.667	-0.166	0.106	0.970
Order	-0.347	0.060	0.000	-0.465	-0.230	0.707
Subject RE	0.038	0.073				

condition coefficient remains almost the same at 1.26 still showing significance but with a slightly bigger variance and a confidence interval of [1.06, 1.51].

Table S4: Study 2 Robustness Regression. OLS Regression without random effects.

Effect	Estimate	SE	p	95% CI LL	95% CI UL	$e^{Estimate}$
Intercept	3.697	0.074	0.000	3.551	3.845	40.326
Condition	0.238	0.090	0.009	0.060	0.416	1.269
Task	-0.100	0.078	0.203	-0.254	0.054	0.904
Order	-0.324	0.069	0.000	-0.461	-0.188	0.723

S3.6 Regression Derivation

To estimate ρ using a regression, we derive a generalized mixed-effects linear regression model for the special case (which occurs in our experiments) where there are two conditions and two tasks which can occur in two orders. We begin with the following multiplicative model of performance:

$$s_{ij} = \beta \frac{a_i}{d_j} f_{ij} g_{ij} e_{ij}$$

where

s_{ij} = speed of subject i on task j

a_i = ability of subject i

d_j = difficulty of task j

f_{ij} = facilitation effect on performance of the condition in which subject i does task j

g_{ij} = effect on performance of the order (first or second) in which subject i does task j

e_{ij} = error for subject i on task j

The model says that relative to some baseline speed β, speed is increased by greater ability *a* of the subject and by greater facilitation *f* of the condition, and speed is decreased by greater task difficulty *d*. We also assume that there is some effect *g* based on whether the task occurs first or second. Finally, we assume that error is multiplicative, not additive. That is, errors are modeled as percentage increases or decreases not absolute increases or decreases.

With these assumptions, the performance of subject i on task j can be modeled as:

$$s_{ij} = \beta \left(\prod_{i=1}^{n} a_i^{q_i} \right) \times \left(\prod_{j=1}^{2} d_j^{-T_j} \right) \times \left(\prod_{i=1}^{n} \prod_{j=1}^{2} f_{ij}^{C_{ij}} g_{ij}^{O_{ij}} \right) \times e_{ij}$$

where

n is the number of subjects,

$$q_i = \begin{cases} 1 & \text{for subject } i \\ 0 & \text{otherwise} \end{cases}$$

$$T_j = \begin{cases} 1 & \text{for task } j \\ 0 & \text{otherwise} \end{cases}$$

$$C_{ij} = \begin{cases} 1 & \text{for the condition in which subject } i \text{ does task } j \\ 0 & \text{otherwise} \end{cases}$$

$$O_{ij} = \begin{cases} 1 & \text{for the order in which subject } i \text{ does task } j \\ 0 & \text{otherwise} \end{cases}$$

This equation can be transformed into a linear regression by treating one option for each indicator variable as part of β and then taking natural logarithms on both sides:

$$ln\, s_{ij} = ln\, \beta + \sum_{i=1}^{n-1} q_i ln\, a_i - T_1\, ln\, d_1 + C_{11}\, ln\, f_{11} + O_{11}\, ln\, g_{11} + ln\, e_{ij}$$

Using the observed values for s_{ij} and the indicator variables for T_j, C_{ij}, and O_{ij} for each observed value, we can use linear regression to estimate the values of $ln\, \beta$, $ln\, d_1$, $ln\, f_{11}$, and $ln\, g_{11}$. We can then relabel these values as β_0, β_1, β_2, and β_3. Finally, we can take the exponentials of these logs to recover the estimated values of the original variables: β, a_i, d_1, f_{11}, and g_{11}. We estimate the subject random effects by taking $v_i = ln\, a_i$.

S3.7 Main Regressions

Tables S5 and S6 contain the Study 2 and programmers vs non-programmers regressions, respectively.

Table S5: Study 2 Regression. Comparison of programmers with and without GPT-3.

Effect	Estimate	SE	p	95% CI LL	95% CI UL	$e^{Estimate}$
Intercept	3.697	0.066	0.000	3.567	3.827	40.326
Condition	0.240	0.076	0.002	0.091	0.390	1.271
Task	-0.098	0.066	0.137	-0.227	0.031	0.376
Order	-0.329	0.059	0.000	-0.444	-0.214	0.720
Subject RE	0.046	0.063				

Table S6: Study 3 Regression. Comparison of programmers and non-programmers, both using GPT-3.

Effect	Estimate	SE	p	95% CI LL	95% CI UL	$e^{Estimate}$
Intercept	-3.792	0.099	0.000	-3.986	3.597	0.023
Condition	0.015	0.118	0.902	-0.217	0.247	1.015
Task	0.311	0.049	0.000	0.216	0.406	1.365
Order	-0.221	0.048	0.000	0.126	0.315	0.802
Subject RE	0.410	0.435				

S3.8 Estimating costs

We compute the total cost per task per subject for the different conditions by computing the human cost which depends on the time spent in a task and the hourly rate per subject; and the GPT-3 cost based on the number of calls and the average cost per call. The Davinci engine is \$0.06 per 1000 tokens, each call uses an average of 66 tokens plus an additional 578 tokens for the prompting. The average cost per GPT-3 call is \$0.039.

S3.9 Cost Regressions

Cost Regressions HC vs H and HC vs HC' (Programmers vs Non-programmers).

Table S7: Study 2 Cost Regression. Comparison of programmers with and without GPT-3.

Effect	Estimate	SE	p	95% CI LL	95% CI UL	$e^{Estimate}$
Intercept	2.507	0.068	0.000	2.373	2.640	12.268
Condition	-0.116	0.072	0.109	-0.258	0.026	0.890
Task	-0.147	0.063	0.019	-0.269	-0.024	0.863
Order	-0.311	0.056	0.000	-0.420	-0.201	0.733
Subject RE	0.196	0.159				

Table S8: Study 3 Cost Regression. Comparison of non-programmers and programmers with GPT-3.

Effect	Estimate	SE	p	95% CI LL	95% CI UL	$e^{Estimate}$
Intercept	2.135	0.108	0.000	1.922	2.347	8.457
Condition	0.337	0.130	0.010	0.082	0.591	1.401
Task	-0.351	0.046	0.000	-0.441	-0.261	0.704
Order	-0.199	0.046	0.000	-0.289	-0.110	0.819
Subject RE	0.512	0.553				

S3.10 Duration and Quality Distributions

Figure S8 and S9 show, respectively, the probability distribution functions and cumulative distribution functions for task duration in the three conditions. Figure S10 shows the probability distribution function for quality in the same three conditions.

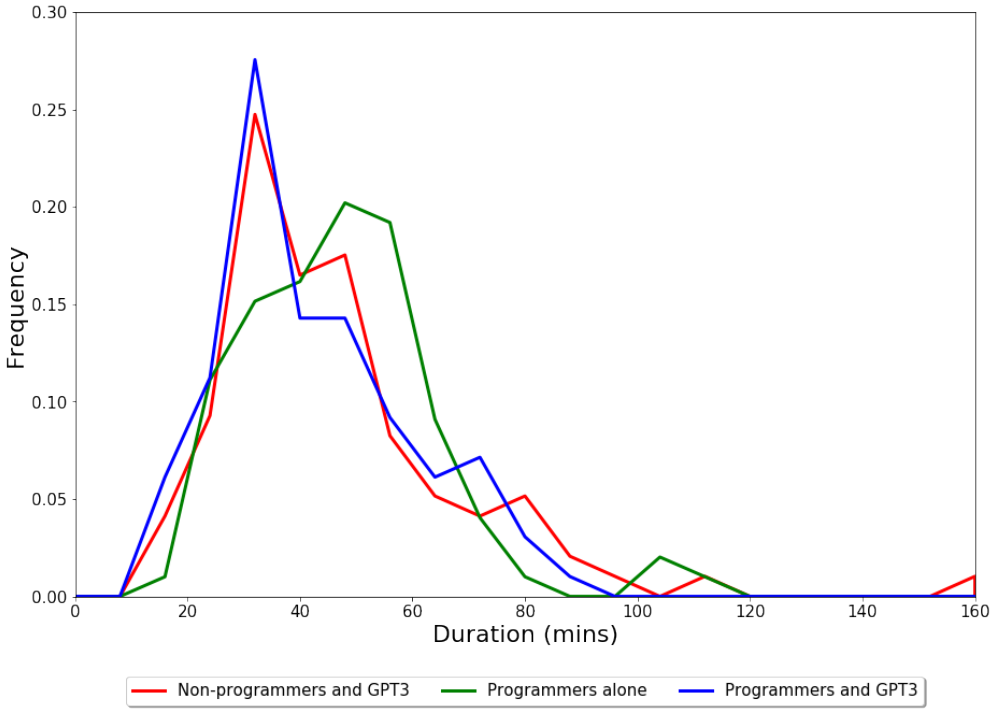


Figure S8: Probability distribution of subjects across task duration for the three conditions.

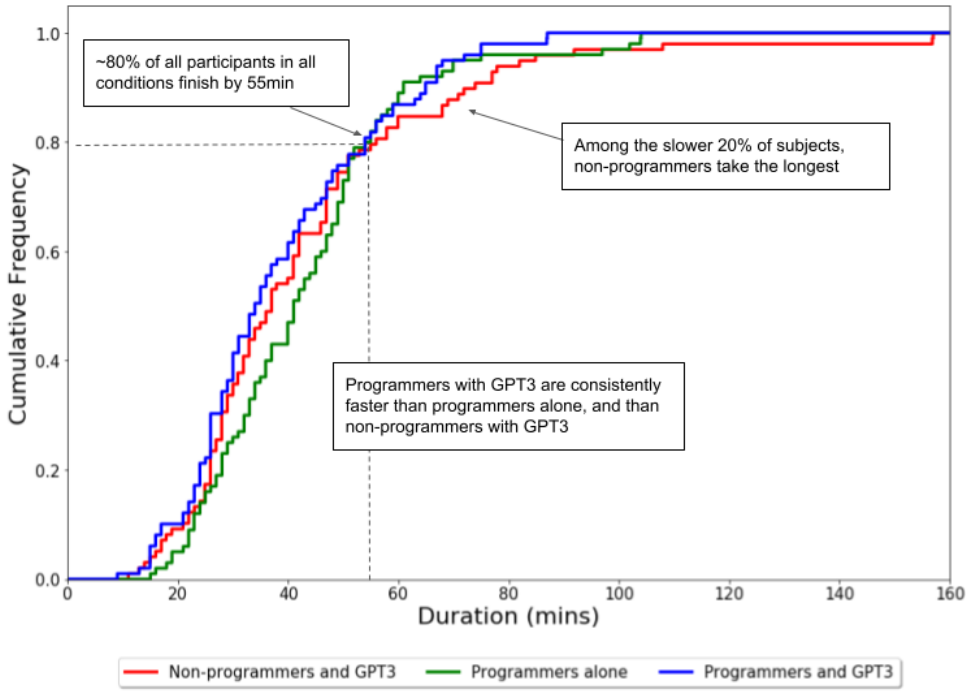


Figure S9: Cumulative distribution of subjects across task duration for the three conditions.

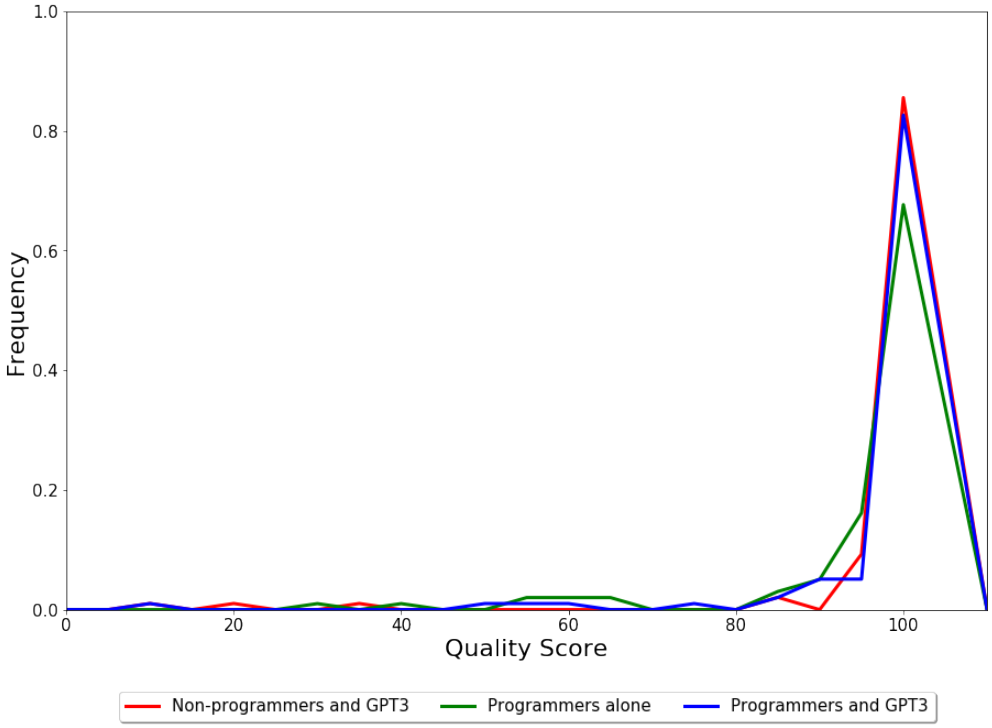


Figure S10: Probability distribution of subjects across quality for the three conditions.

References

[1] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, p. 677–680, 1946.

[2] R. J. Grissom and J. J. Kim, *Effect Sizes for Research: Univariate and Multivariate Applications*. Routledge, Taylor and Francis Group, 2nd ed., 2012. Chapter 2.

[3] F. Mosteller and J. Tukey, *Data Analysis and Regression: A Second Course in Statistics*. Addison Wesley, 1977.

[4] P. Velleman and L. Wilkinson, "Nominal, ordinal, interval and ratio typologies are misleading," *The American Statistician*, vol. 47, p. 65–72, 1993.

[5] S. S. Stevens, "Mathematics, measurement, and psychophysics," in *Handbook of Experimental Psychology* (S. S. Stevens, ed.), New York: John Wiley, 1951.

[6] K. S. Button, D. Kounali, L. Thomas, N. J. Wiles, N. J. Peters, A. E. Ades, and G. Lewis, "Minimal clinically important difference on the beck depression inventory—II according to patient's perspective," *Psychological Medicine*, vol. 45, p. 3269–3279, 2015. http:// doi.org/10.1017/S0033291715001270.

[7] T. J. B. Kline, *Psychological Testing: A Practical Approach to Design and Evaluation*. Sage, 2005.

[8] R. J. Grissom and J. J. Kim, *Effect Sizes for Research: Univariate and Multivariate Applications*. Routledge, Taylor and Francis Group, 2nd ed., 2012. Chapter 8.

[9] B. A. Kitchenham, *Procedures for Performing Systematic Reviews*. Keele University, 2004.

[10] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, "The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration," *British Medical Journal*, vol. 339, 2009.

[11] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. A. Ebel, "The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems," in *Hawaii International Conference on System Sciences*, 2019.

[12] A. Gerber, P. Derckx, D. A. Döppner, and D. Schoder, "Conceptualization of the human-machine symbiosis - A literature review," in *Hawaii International Conference on System Sciences*, 2020.

[13] A. V. Gonzalez, G. Bansal, A. Fan, Y. Mehdad, R. Jia, and S. Iyer, "Do explanations help users detect errors in open-domain QA? An evaluation of spoken vs. visual explanations," in *FINDINGS*, 2021.

[14] K. Holstein and V. Aleven, "Designing for human-AI complementarity in K-12 education," *ArXiv*, vol. abs/2104.01266, 2021.

[15] X. Fang, X. Wu, and Y. Qian, "Design and analyze one object recognition system with human in the loop for unmanned platforms with 5G communication," *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 570–575, 2021.

[16] M. Leimkühler, L. S. Gravemeier, T. Biester, and O. Thomas, "Deep learning object detection as an assistance system for complex image labeling tasks," in *Hawaii International Conference on System Sciences*, 2021.

[17] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, 2022.

[18] M. Desmond, Z. Ashktorab, M. Brachman, K. Brimijoin, E. Duesterwald, C. Dugan, C. Finegan-Dollak, M. J. Muller, N. N. Joshi, Q. Pan, and A. Sharma, "Increasing the speed and accuracy of data labeling through an AI assisted interface," *26th International Conference on Intelligent User Interfaces*, 2021.

[19] E. Bondi, R. Koster, H. R. Sheahan, M. J. Chadwick, Y. Bachrach, taylan. cemgil, U. Paquet, and K. Dvijotham, "Role of human-AI interaction in selective prediction," *ArXiv*, vol. abs/2112.06751, 2021.

[20] A. Fügener, J. Grahl, A. Gupta, and W. Ketter, "Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI," *Management Information Systems Quarterly*, vol. 45, 2021.

[21] W. Shin, J. Han, and W. Rhee, "AI-assistance for predictive maintenance of renewable energy systems," *Energy*, vol. 221, p. 119775, 2021.

[22] T. Baudel, M. Verbockhaven, V. Cousergue, G. Roy, and R. Laarach, "Objectivaize: Measuring performance and biases in augmented business decision systems," in *IFIP TC13 International Conference on Human-Computer Interaction*, (Berlin, Heidelberg), p. 300–320, Springer-Verlag, 2021.

[23] F. D. Pereira, F. Pires, S. C. Fonseca, E. H. T. Oliveira, L. S. G. Carvalho, D. B. F. Oliveira, and A. I. Cristea, "Towards a human-AI hybrid system for categorising programming problems," *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 2021.

[24] Z. Gao and J. Jiang, "Evaluating human-AI hybrid conversational systems with chatbot message suggestions," *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.

[25] H. Bastani, O. Bastani, and W. P. Sinchaisri, "Improving human decision-making with machine learning," *ArXiv*, vol. abs/2108.08454, 2021.

[26] H. Liu, V. Lai, and C. Tan, "Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 1 – 45, 2021.

[27] R. Zeleznik, J. Weiss, J. Taron, C. V. Guthier, D. S. Bitterman, C. I. Hancox, B. H. Kann, D. W. Kim, R. S. Punglia, J. S. Bredfeldt, B. Foldyna, P. Eslami, M. T. Lu, U. Hoffmann, R. H. Mak, and H. J. Aerts, "Deep-learning system to improve the quality and efficiency of volumetric heart segmentation for breast cancer," *NPJ Digital Medicine*, vol. 4, 2021.

[28] G. Pavoni, M. Corsini, F. Ponchio, A. Muntoni, C. B. Edwards, N. E. Pedersen, S. A. Sandin, and P. Cignoni, "Taglab: AI-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages," *Journal of Field Robotics*, vol. 39, pp. 246 – 262, 2022.

[29] B. Kleinberg and B. Verschuere, "How humans impair automated deception detection performance.," *Acta Psychologica*, vol. 213, p. 103250, 2021.

[30] A. Levy, M. Agrawal, A. Satyanarayan, and D. A. Sontag, "Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[31] S. M. Jesus, C. Bel'em, V. Balayan, J. Bento, P. Saleiro, P. Bizarro, and J. Gama, "How can I choose an explainer? An application-grounded evaluation of post-hoc explanations," *ArXiv*, vol. abs/2101.08758, 2021.

[32] J. Skirzynski, F. Becker, and F. Lieder, "Automatic discovery of interpretable planning strategies," *Machine Learning*, vol. 110, pp. 2641–2683, 2021.

[33] A. Jain, D. Way, V. Gupta, Y. Gao, G. de Oliveira Marinho, J. Hartford, R. Sayres, K. Kanada, C. Eng, K. Nagpal, K. B. Desalvo, G. S. Corrado, L. H. Peng, D. R. Webster, R. C. Dunn, D. Coz, S. J. Huang, Y. Liu, P. Bui, and Y. Liu, "Development and assessment of an artificial intelligence–based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices," *JAMA Network Open*, vol. 4, 2021.

[34] H. Kyriacou, A. Ramakrishnan, and J. Whitehill, "Learning to work in a materials recovery facility: Can humans and machines learn from each other?," *LAK21: 11th International Learning Analytics and Knowledge Conference*, 2021.

[35] M. L. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos, "How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection," *Translational Psychiatry*, vol. 11, 2021.

[36] Z. Buccinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 1 – 21, 2021.

[37] J. Higgins and J. Thomas, eds., *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane Training, 6.3 ed., 2022. https://training.cochrane.org/handbook/current.

[38] B. A. Kitchenham and S. L. Pfleeger, "Principles of survey research part 2: Designing a survey," *SIGSOFT Software Engineering Notes*, vol. 27, p. 18–20, Jan 2002.

[39] B. A. Kitchenham and S. L. Pfleeger, "Principles of survey research: Part 3: Constructing a survey instrument," *SIGSOFT Software Engineering Notes*, vol. 27, p. 20–24, Mar 2002.

[40] P. Diebold, C. Lampasona, S. Zverlov, and S. Voss, "Practitioners' and researchers' expectations on design space exploration for multicore systems in the automotive and avionics domains: A survey," in *Evaluation and Assessment in Software Engineering '14*, 2014.

[41] L. Jarvis and S. Macdonald, "What is cyberterrorism? Findings from a survey of researchers," *Terrorism and Political Violence*, vol. 27, pp. 657 – 678, 2015.

[42] H. Gehlbach and S. Barge, "Anchoring and adjusting in questionnaire responses," *Basic and Applied Social Psychology*, vol. 34, pp. 417 – 433, 2012.

[43] C. F. de Sousa Rodrigues, F. J. C. de Lima, and F. T. Barbosa, "Importance of using basic statistics adequately in clinical research," *Brazilian Journal of Anesthesiology (English Edition)*, vol. 67, no. 6, pp. 619–625, 2017.

[44] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[45] OpenAI, "GPT-4 Technical Report," *ArXiv*, 2023.

[46] D. G. Bonett and R. M. Price, "Confidence intervals for ratios of means and medians," *Journal of Educational and Behavioral Statistics*, vol. 45, no. 6, pp. 750–770, 2020. http:// doi.org/10.3102/1076998620934125.