INTERDISCIPLINARY CENTER
FOR ECONOMIC SCIENCE
GEORGE MASON UNIVERSITY

Do People Trust Humans More Than ChatGPT?

Joy Buchanan and William Hickman

November 2023

Discussion Paper

Interdisciplinary Center for Economic Science
4400 University Drive, MSN 1B2, Fairfax, VA 22030
Tel: +1-703-993-4850   Fax: +1-703-993-4851
ICES Website: http://ices.gmu.edu
ICES RePEc Archive Online at: http://edirc.repec.org/data/icgmuus.html

Do People Trust Humans More Than ChatGPT?

Joy Buchanan[*] William Hickman[†]

November 17, 2023

Abstract

We explore whether people trust the accuracy of statements produced by large language models (LLMs) versus those written by humans. While LLMs have showcased impressive capabilities in generating text, concerns have been raised regarding the potential for misinformation, bias, or false responses. In this experiment, participants rate the accuracy of statements under different information conditions. Participants who are not explicitly informed of authorship tend to trust statements they believe are human-written more than those attributed to ChatGPT. However, when informed about authorship, participants show equal skepticism towards both human and AI writers. There is an increase in the rate of costly fact-checking by participants who are explicitly informed. These outcomes suggest that trust in AI-generated content is context-dependent.

Keywords: Artificial intelligence, machine learning, trust, belief

JEL Codes: O33, C91, D8

[*] Brock School of Business, Samford University, Birmingham, AL, USA, jbuchan1@samford.edu, corresponding author

[†] The Interdisciplinary Center for Economic Science, George Mason University, Arlington, VA, USA

1 Introduction

The capabilities of Large Language Models (LLMs) have significantly improved in recent years, transforming natural language processing and generating human-like text with remarkable fluency. In particular, OpenAI's ChatGPT has demonstrated the ability to generate coherent and contextually relevant responses to various prompts. While LLMs have showcased impressive capabilities, concerns have been raised regarding the potential for misinformation or bias in their responses. Indeed, ChatGPT is known to produce inaccurate results sometimes. Understanding the level of trust humans place in LLMs is crucial, as these models are increasingly being deployed in various real-world applications, including content generation, automated customer support, and information retrieval.

LLM tools will be adopted where they are cost-effective because of demonstrated successes in workplace settings (Brynjolfsson, Li, and Raymond, 2023; Fumagalli, Rezaei, and Salomons, 2022). So, it is increasingly likely that human readers will encounter AI-written text.

We investigate whether humans trust the factual accuracy of LLMs. Our study is the first controlled experiment of this type that is incentivized and gives the human subjects the option of doing a costly fact-check, which mirrors the options available to workers and managers in the workplace. Subjects rate the accuracy of statements written by humans or ChatGPT. We vary whether each statement was written by a human or ChatGPT and whether we explicitly inform subjects about authorship.

Our findings paint a complex portrait of trust. If subjects are not informed about authorship but first have to guess whether a paragraph was written by AI or by a human, then the subjects who guess the paragraph was written by a human display higher trust in the statement's factual accuracy. However, in the treatments with explicit information about authorship, subjects are equally suspicious that the AI or human-written paragraph contains an error and needs to be fact-checked. We collect information about age and whether participants have had exposure to ChatGPT prior to the experiment. Neither of those characteristics has a significant impact on trust.

Our results indicate that trust in AI-written content is context-dependent, especially now that AI writing can closely mimic human style and content. We add to a growing literature finding that informed individuals are not inherently biased against the accuracy of AI outputs (Epstein, Arechar, and Rand, 2023).

This research will provide insights for researchers, developers, and policymakers, paving the way for a more informed integration of LLMs into our increasingly AI-driven world. We also present a novel design that can be extended and modified to study AI trust.

The remainder of this paper is structured as follows. Section 2 discusses the literature related to our research. Section 3 explains our experimental design and procedures, while Section 4 details our hypotheses. Section 5 presents our results, and Section 6 concludes.

2 Related Literature

We first emphasize the fact that LLMs can hallucinate falsehoods that misinform human readers. In a survey article, Ji et al. (2023) state that "deep learning based generation is prone to hallucinate unintended text, which degrades the system performance and fails to meet user expectations in many real-world scenarios." Zhang et al. (2023) find that LLMs cannot fully self-correct or recognize mistakes. Buchanan, Hill, and Shapoval (2023) demonstrated that ChatGPT produces fake citations for papers that do not exist, with both GPT-3.5 and the more advanced GPT-4. Open AI researchers concluded after running image processing tests that "Given the model's imperfect performance in this domain and the risks associated with inaccuracies, we do not consider the current version of GPT-4V to be fit for performing any medical function or substituting professional medical advice, diagnosis, or treatment, or judgment" (OpenAI, 2023). Regardless, it is very likely that people will try to use LLMs for many tasks, including medical advice. Measuring trust in LLMs is needed in order to gauge the impact they will have on real-world outcomes.

While Gillespie et al. (2023) and others have conducted surveys about trust in AI, we are among the first to do an incentivized study to test whether people trust AI generated writing. Scientists should not rely completely on self-reported data for this issue (Smith, 1994). Humans might think they are supposed to say humans are more trustworthy than AI in a survey. However, if fact checking is costly, then these same people might rely on AI-generated text in practice without screening it for inaccuracies.

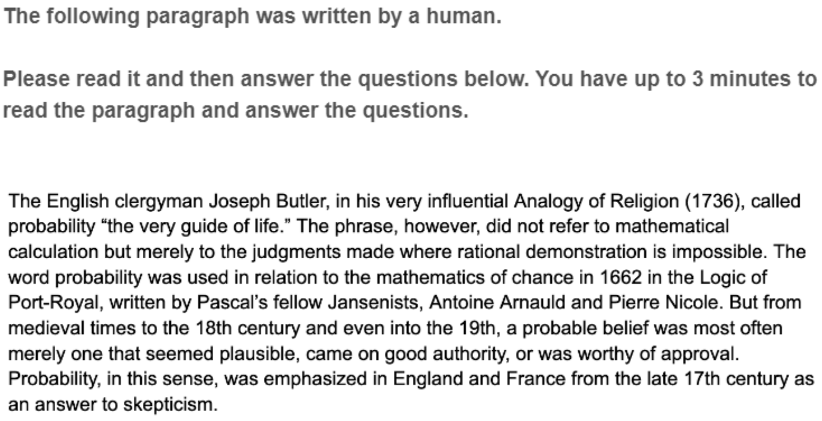
In an unincentivized study, Spitale, Biller-Andorno, and Germani (2023) recruit human subjects to evaluate tweets written by humans versus tweets written by AI. They conclude that, "In comparison with humans, [LLMs] can produce accurate information that is easier to understand, but it can also produce more compelling disinformation. We also show that humans cannot distinguish between tweets generated by GPT-3 and written by real Twitter users." Köbis and Mossink (2021) study whether respondents have a subjective preference for art created by humans versus poetry written by AI, with mixed results. Babin and Chauhan (2023) find that people prefer advice from humans in social dilemmas and have a higher willingness to pay for advice from humans.

To date, there has been more experimental work on lying than on unintentional AI mistakes. For example, Chen and Houser (2017) examine whether human subjects trust messages written by human counterparts in a trust game. Serra-Garcia and Gneezy (2021) ask people to detect messages with lies in an incentivized study. They find that subjects are overconfident in their own ability to detect lies (from audio visual sources) which has implications for what gets amplified on social media.

Pennycook et al. (2021) use a prompt that causes people to pause and think about accuracy before sharing a story on social media. They found that "subtly shifting attention to accuracy increases the quality of news that people subsequently share."[1] Our results might be interpreted as a replication of Pennycook et al. (2021) because when we draw attention to authorship our subjects become suspicious and more inclined to do a costly fact-check.

Chugunova and Sele (2022) and March (2021) provide overviews of studies on human interactions with autonomous agents. Humans treat AI counterparts differently from human strategic partners. It is evident that humans apply different cognitive processes when they know they are facing a computer. Chugunova and Sele (2022) state that "while humans seem willing to accept automated agents in areas considered more objective or analytical, they seem reluctant to do so in areas considered social or moral." We replicate that to some extent when we test human perception of an AI doing a fairly objective task which is to write a factual statement about an uncontroversial topic. Our subjects are about as suspicious of ChatGPT as human writers, when they are informed about authorship. However, we find that context affects the willingness to trust AI. This finding is consistent with their conclusion that humans are willing to incorporate the advice of computers.

[1] This result was replicated by Athey et al. (2023), who ran phone-based tutorials in Kenya about not impulsively sharing stories on social media that might be false.



The following paragraph was written by a human.

Please read it and then answer the questions below. You have up to 3 minutes to read the paragraph and answer the questions.

The English clergyman Joseph Butler, in his very influential Analogy of Religion (1736), called probability "the very guide of life." The phrase, however, did not refer to mathematical calculation but merely to the judgments made where rational demonstration is impossible. The word probability was used in relation to the mathematics of chance in 1662 in the Logic of Port-Royal, written by Pascal's fellow Jansenists, Antoine Arnauld and Pierre Nicole. But from medieval times to the 18th century and even into the 19th, a probable belief was most often merely one that seemed plausible, came on good authority, or was worthy of approval. Probability, in this sense, was emphasized in England and France from the late 17th century as an answer to skepticism.

Figure 1: Human-Written Paragraph A

Yet, this willingness varies based on the framing of the situation and the perception of autonomy. Sunstein and Reisch (2023) also find that framing affects the human perception of algorithms.

Because LLMs are new, we seek to provide novel evidence on how humans perceive natural language that they believe was entirely written by a computer.

3 Experimental Design and Procedures

Participants answered questions after reading a 5-sentence paragraph. Two types of paragraphs were used: human-written and AI-generated. The AI paragraph was generated by prompting ChatGPT to replicate the style, sentence count, and word count of the human-written paragraph without altering factual content.[2] Figure 1 shows Paragraph A written by a human, as it was presented to subjects. The Appendix contains an example of the AI-written Paragraph B along with screenshots of instructions for the experiment.

Subjects were randomly assigned to one of the treatments listed in Table 1.

Table 1: Treatments

Human-written (informed)	Human-written (uninformed)
AI-written (informed)	AI-written (uninformed)

Participants indicated whether they thought the paragraph was written by a human or AI, with

[2] For robustness, we used two sets of human and AI-written paragraphs (labeled A and B), instead of relying on one. Our analysis includes data from all paragraphs, allowing for differences in average trust based on the paragraph set reviewed.

an incentive for being correct. If the experimenter informed the participants about authorship, then this question served as an attention check, and otherwise it was an incentivized belief elicitation.

Participants could earn money for correctly guessing whether the paragraph contained any factual errors. They could pick from three responses:

1. There are not any factual errors in this paragraph
2. I would like to purchase a fact-check (cost: $0.20)
3. There are factual errors in this paragraph

Correct answers without a fact-check earned a $0.50 bonus. Opting for a fact-check guaranteed a $0.30 bonus ($0.50 - $0.20).

We use these answers to create our trust measure. A risk-neutral subject who guesses that there are no errors without requesting the fact-check indicates that they believe there is at least a 60% chance of the paragraph being entirely correct. If a risk-neutral subject decides that purchasing a fact-check maximizes their expected value, then they must estimate the probability of the paragraph being correct is between 40% and 60%. A risk-neutral subject who is confident that the paragraph is incorrect maximizes their payoff by answering that there are errors and not purchasing a fact-check. If subjects are risk-averse, that might lead to an overall increase in fact-checks, but this is unlikely to differ systematically between treatments.

We categorize responses as follows: no errors is “high trust”, errors is “low trust”, and fact-check requests are labeled “medium trust”.

The experiment concluded with a demographic survey and questions about familiarity with ChatGPT.

In July 2023, we recruited subjects via Prolific to take the survey. Performance bonuses were paid to subjects who completed the questions. As an attention check, subjects had to correctly count the number of sentences in the paragraph. Subjects in our sample are at least 18 years old, live in the United States, and have at least a 75% approval rating on Prolific.

Five hundred subjects participated in the experiment, which we pre-registered in the American Economic Association’s RCT registry. Table 2 shows the number of participants in each treatment.

Table 2: Number of Subjects

	Paragraph A	Paragraph B	Total
Human-written (informed)	78	86	164
Human-written (uninformed)	50	39	89
AI-written (informed)	79	69	148
AI-written (uninformed)	43	56	99

So as not to rely exclusively on a single paragraph, we used two similar human-written paragraphs and derived two AI-written paragraphs from them. Each of the paragraphs included details about the history of statistics, and we only used paragraphs that did not include controversial facts or polarizing topics. None of the paragraphs contain factual errors. In our analysis, we allow average trust to vary depending on which passage the subject viewed. Each subject only read one paragraph.

4 Hypotheses

Hypothesis 1: There will be higher trust in passages believed to have been written by a human author as opposed to generated by AI.

> In the Informed treatment, this will result in higher trust among subjects who are informed about human authorship. In the Uninformed treatment, this might result in higher trust for passages that are believed to be authored by a human. We expect this because LLMs are a relatively new technology and subjects should be familiar with published writing by humans being generally reliable.

Hypothesis 2: The passages written by humans versus AI will inspire equal levels of trust, apart from what is believed about authorship.

> We instructed ChatGPT to mimic the tone and language of the human-written paragraphs as closely as possible. We do not expect that uninformed subjects will be more or less trusting simply based on the style of writing (consistent with Casal and Kessler (2023)). We only expect to see a difference based on perceived authorship.

> If it were true that the paragraphs written by ChatGPT are of obviously poor quality, then we would not directly be testing trust in the author's ability to not make factual errors. We expect to find a null result, and our design allows us to make the direct comparison, with two slightly different topics (Paragraph A and Paragraph B).

Hypothesis 3: Among those who are informed that the text is written by AI, trust levels will decrease with the age of the participant.

> We collect information on age to test whether older participants are less trusting of new this technology.

Hypothesis 4: Among those who are informed that the text is written by AI, those who have used AI for writing will trust AI-written text less than those who have not used AI for writing.

> Subjects who have used ChatGPT for writing might have directly observed it making a mistake. Therefore, we test whether there is a significant effect on trust in this context.

5 Results

We surveyed American adults, half of whom have at least a 4-year college degree. Table 3 contains descriptive statistics of the sample. More than 80% of subjects had heard of ChatGPT but only one third had used it to write. We show the balance of our samples between treatments in Table A1.

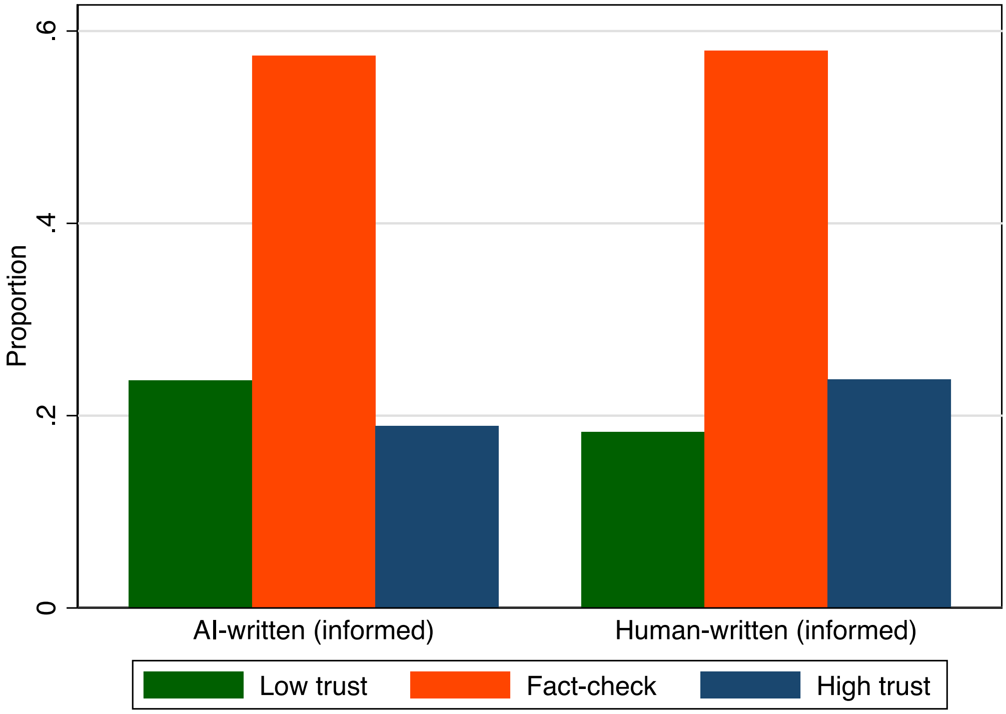
Table 3: Sample Description

	Mean	SD
Age	36.77	12.40
Female	0.48	0.50
4 year college degree or higher	0.56	0.50
Has heard news stories about AI	0.85	0.36
Has seen examples of writing by ChatGPT	0.86	0.35
Has used ChatGPT for writing	0.35	0.48
Observations	500	

Result 1a: Subjects are not more trusting of human-written text, when informed about authorship.

Figure 2 shows that the distribution of outcomes is similar when subjects are explicitly informed of authorship. The modal response is to purchase a fact-check, and subjects do not appear significantly more willing to trust a human author. In both treatments, only about 20% of participants have high trust in the accuracy of the paragraph. Being informed of authorship and being asked an artifactual question may have raised suspicion, equally so for AI and human authored paragraphs.

Figure 2: Comparison of trust among the informed



We use ordered logistic regressions to test for differences in trust between treatments.[3] The regression in column (1) of Table 4 confirms our visual analysis of behavior when participants are explicitly informed about authorship. The coefficient on human-written is positive but small and insignificant.

[3] In this regression model, the categories of the dependent variable are conceptualized as being underpinned by a continuous latent variable. In this case, the underlying trust (which can be between 0 (no trust) to 1 (complete trust) in the factual accuracy of text is the latent variable. While the coefficients themselves are not easily interpreted, we primarily focus on using this regression framework to test for significant differences in trust between treatments.

Table 4: Trust compared between treatments and by authorship belief

	(1)	(2)	(3)
Human-written	0.247	0.149	
	(0.171)	(0.276)	
Informed	0.206		
	(0.179)		
Believed human authorship			1.110***
			(0.300)
Paragraph B	-0.250	-0.155	-0.299
	(0.171)	(0.276)	(0.277)
Observations	500	188	188

Standard errors in parentheses

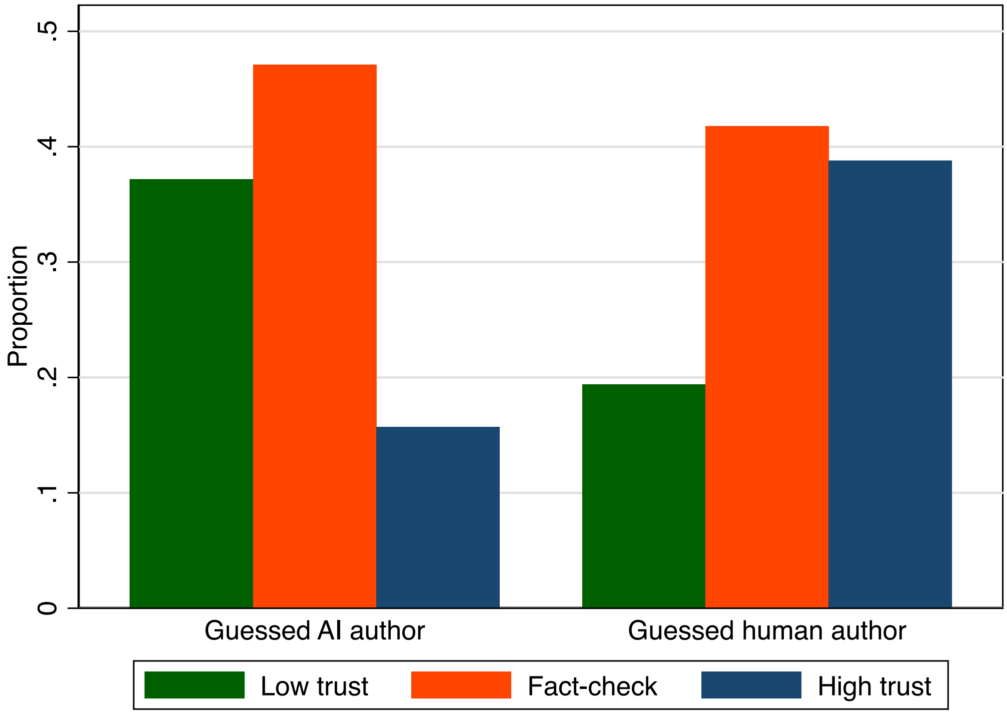
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: The dependent variable in each of these ordered logistic regressions is our measure of Trust. Column (1) includes all participants, while columns (2)-(3) only include participants in the Uninformed treatments. "Human-written" is a binary variable that =1 if a participant reviewed human-written text. "Informed" is a binary variable that =1 if a participant was informed about the authorship of the text they reviewed. "Believed human authorship" is a binary variable that =1 if a participant (in the Uninformed treatments) guessed the text they reviewed was human-written. "Paragraph B" is a binary variable that =1 if a participant saw the second set of paragraphs we included for robustness.

Result 1b: When subjects are not explicitly informed of authorship, there is higher trust in paragraphs believed to have been written by humans.

We find partial support for our first hypothesis when considering only the trust levels of people who were uninformed. Subjects who believe the paragraph they review is human-written are more confident in the accuracy of the paragraph. We show this visually in Figure 3, where "High trust" ("Low trust") is higher (lower) for those who believe the paragraph is human-written than for those who believe the paragraph is AI-written. We might conclude from this that subjects are more trusting of human authors or that they associate trustworthy writing with human authors. However, this relationship is only evident among the uninformed. We test for this relationship in column (3) of Table 4, and find that, among the uninformed, those who believe the text is human-written trust the text's factual accuracy significantly more than those who believe the text is AI-written.

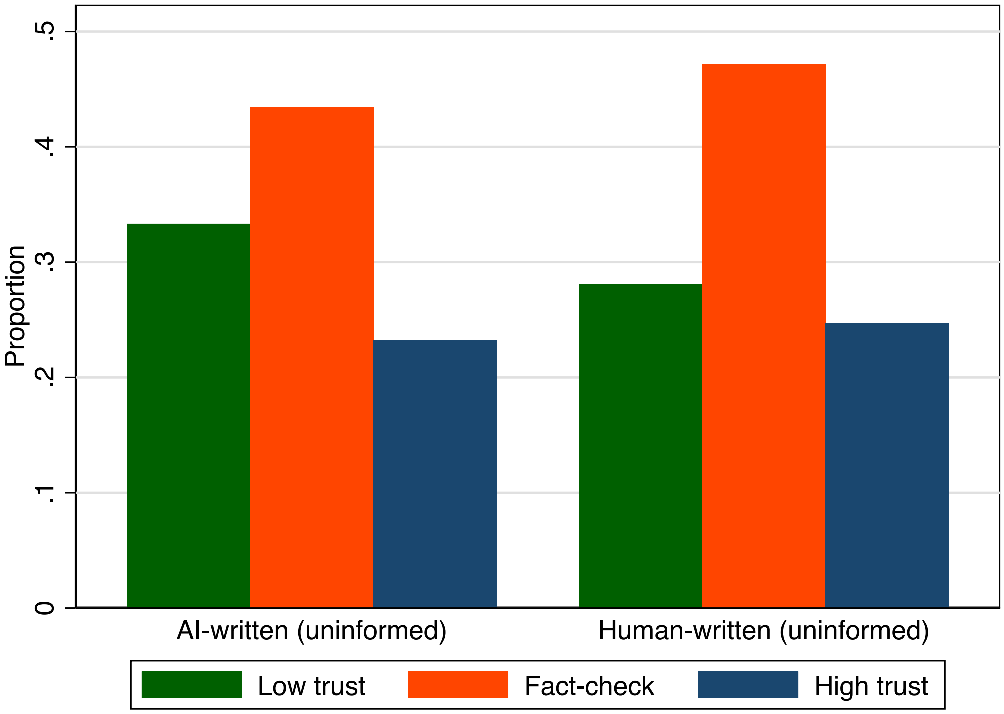
Figure 3: Comparison of trust among the uninformed, by authorship belief



Result 2: The text of the writing is not perceived differently between human and AI authors.

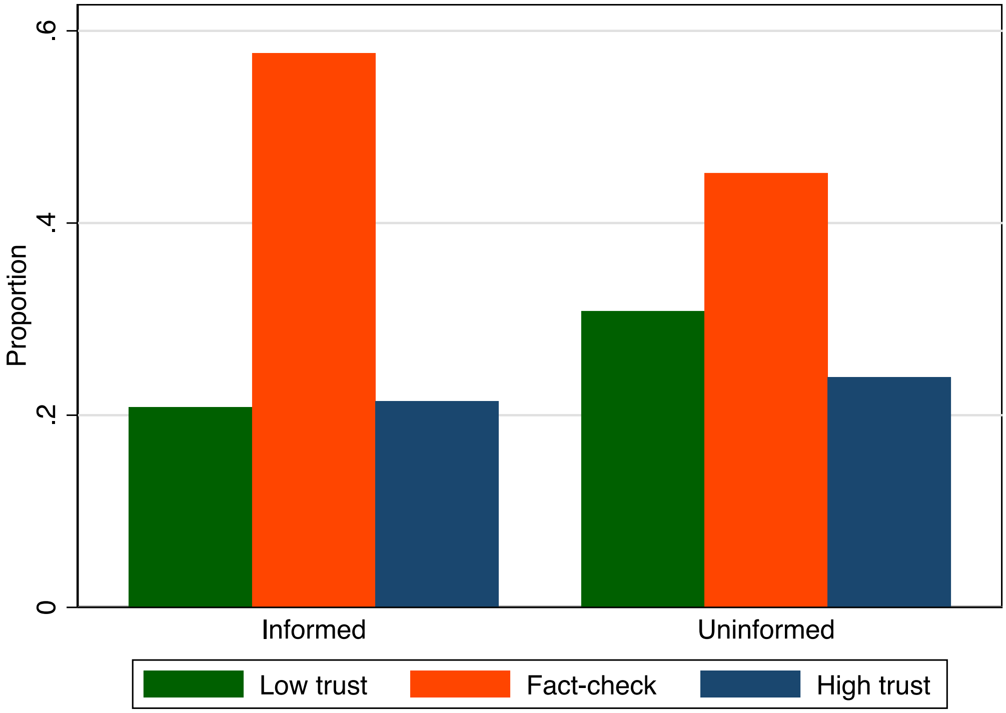
When we present the data visually in the same way for the Uninformed treatments in Figure 4, the pattern is similar to the Informed treatments although subjects purchase fewer fact-checks. We conclude that the paragraphs are not inherently perceived to be different, so we do not reject the null hypothesis for Hypothesis 2. We conclude that differences in trust are driven by perception of the nature of the author rather than the particular style of the paragraph. We test this is in column (2) of Table 4, where we restrict the sample to the uninformed and do not find a significant relationship between trust in the paragraphs and whether the paragraphs were human or AI-written.

Figure 4: Comparison of trust among the uninformed



As we show visually in Figure 5, the rate of fact-checking (although not overall trust) is significantly higher in the Informed treatments than in the Uninformed treatments, ($\chi^2(1) = 7.33$, $p = 0.007$). This implies that the level of confidence in the subjects' own ability to be a judge is lower when they are explicitly informed about authorship. This would be interesting to pursue toward a policy goal of making people more thoughtful about any information they see online.

Figure 5: Comparison of trust between informed and uninformed



We asked several demographic questions to test hypotheses about the interaction of trust and life experience with technology. The lack of significant results in Table A2 indicates that we cannot reject the null for hypothesis 3 or 4.

Result 3: Trust does not appear to decrease with age in our sample.

Result 4: Prior exposure to ChatGPT does not significantly impact trust.

6 Conclusion

Considering that generative AI can produce writing that sounds correct but contains factual errors, it is important to understand how human readers interact with tools like ChatGPT. Our paper contributes to the literature on disinformation and the behavioral aspects of artificial intelligence advances. We found some evidence that human subjects are more trusting of human authors, but we also show that trust is context-dependent.

AI is a powerful tool that will be widely adopted. Policymakers and researchers have expressed concern about "alignment" in this new era. Messages generated by AI might not always be true or convey the most helpful content to align with the goals of humans who query a tool like ChatGPT or who read an AI-generated website.

People tend to be overconfident in their own ability to assess text (Serra-Garcia and Gneezy, 2021). If certain framing prompts people to fact-check, then that could be a promising policy avenue. We find that people are more likely to fact-check when they are explicitly informed about authorship before being asked about errors in the text. Similarly, Pennycook et al. (2021) found that people are more hesitant to share stories on social media if they are prompted to consider accuracy.

Based on our results, there might be ways of displaying AI-generated writing that will encourage human readers to be alert and do fact-checking when necessary. That would allow humans to get the benefit of efficiency from AI while mitigating the risks of disinformation or bias. Although we did not find a significant correlation for the demographic information that we collected, future research can explore whether some groups of people are more trusting of AI and therefore more vulnerable to misinformation. We consider this to be an important new field of behavioral research, because ChatGPT is not infallible.

References

Athey, Susan et al. (2023). "Emotion-versus Reasoning-based Drivers of Misinformation Sharing: A field experiment using text message courses in Kenya". *Available at SSRN 4489759*.

Babin, J. Jobu and Haritima Chauhan (2023). "Chatbot or Humanaut? How the Source of Advice Impacts Behavior in One-shot Social Dilemmas". *Working Paper*.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond (2023). *Generative AI at work*. Tech. rep. National Bureau of Economic Research.

Buchanan, Joy, Stephen Hill, and Olga Shapoval (2023). "ChatGPT Hallucinates Nonexistent Citations: Evidence from Economics". *Working Paper*.

Casal, J. Elliott and Matt Kessler (2023). "Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing". *Research Methods in Applied Linguistics* 2.3, p. 100068. ISSN: 2772-7661. DOI: `https://doi.org/10.1016/j.rmal.2023.100068`.

Chen, Jingnan and Daniel Houser (2017). "Promises and lies: can observers detect deception in written messages". *Experimental Economics* 20, pp. 396–419.

Chugunova, Marina and Daniela Sele (2022). "We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines". *Journal of Behavioral and Experimental Economics* 99, p. 101897. ISSN: 2214-8043. DOI: `https://doi.org/10.1016/j.socec.2022.101897`.

Epstein, Ziv, Antonio Alonso Arechar, and David Rand (2023). "What label should be applied to content produced by generative AI?"

Fumagalli, Elena, Sarah Rezaei, and Anna Salomons (2022). "OK computer: Worker perceptions of algorithmic recruitment". *Research Policy* 51.2, p. 104420.

Gillespie, Nicole et al. (2023). "Trust in Artificial Intelligence: A global study".

Ji, Ziwei et al. (2023). "Survey of hallucination in natural language generation". *ACM Computing Surveys* 55.12, pp. 1–38.

Köbis, Nils and Luca D Mossink (2021). "Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry". *Computers in human behavior* 114, p. 106553.

March, Christoph (2021). “Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players”. *Journal of Economic Psychology* 87, p. 102426. ISSN: 0167-4870. DOI: https://doi.org/10.1016/j.joep.2021.102426. URL: https://www.sciencedirect.com/science/article/pii/S0167487021000593.

OpenAI (2023). “GPT-4V(ision) System Card”.

Pennycook, Gordon et al. (2021). “Shifting attention to accuracy can reduce misinformation online”. *Nature* 592.7855, pp. 590–595.

Serra-Garcia, Marta and Uri Gneezy (2021). “Mistakes, overconfidence, and the effect of sharing on detecting lies”. *American Economic Review* 111.10, pp. 3160–3183.

Smith, Vernon L (1994). “Economics in the Laboratory”. *Journal of economic perspectives* 8.1, pp. 113–131.

Spitale, Giovanni, Nikola Biller-Andorno, and Federico Germani (2023). “AI model GPT-3 (dis) informs us better than humans”. *arXiv preprint arXiv:2301.11924*.

Sunstein, Cass R and Lucia Reisch (2023). “Do People Like Algorithms? A Research Strategy”. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4544749.

Zhang, Muru et al. (2023). *How Language Model Hallucinations Can Snowball.* arXiv: 2305.13534 [cs.CL].

A Additional Results

There is one significant difference found in our balance tables. We have an ex-post explanation for it. The percentage of people who had engaged with ChatGPT increased from the first wave with Paragraph A to the second wave with Paragraph B. This survey was run at the time (June 2023) when news about ChatGPT was spreading quickly. That is likely the reason that there was a significant increase in people who had used ChatGPT from Paragraph A to Paragraph B. This should not affect our results or the interpretation of treatment effects.

Table A1: Treatment Balance

(a) Human-written text

Variable	Human-written (uninformed)	Human-written (informed)	Difference
Age	36.640	35.848	-0.793
	(10.976)	(12.445)	(1.573)
Female	0.517	0.500	-0.017
	(0.503)	(0.502)	(0.066)
4 year college degree or higher	0.539	0.555	0.016
	(0.501)	(0.499)	(0.066)
Has heard news stories about AI	0.820	0.848	0.027
	(0.386)	(0.361)	(0.049)
Has seen examples of writing by ChatGPT	0.798	0.884	0.086*
	(0.404)	(0.321)	(0.046)
Has used ChatGPT for writing	0.281	0.354	0.073
	(0.452)	(0.480)	(0.062)
Observations	89	164	253

Notes: Standard errors in parentheses. *$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

(b) AI-written text

Variable	AI-written (uninformed)	AI-written (informed)	Difference
Age	36.323	38.155	1.832
	(11.959)	(13.418)	(1.669)
Female	0.424	0.459	0.035
	(0.497)	(0.500)	(0.065)
4 year college degree or higher	0.545	0.595	0.049
	(0.500)	(0.493)	(0.064)
Has heard news stories about AI	0.818	0.899	0.080*
	(0.388)	(0.303)	(0.044)
Has seen examples of writing by ChatGPT	0.818	0.905	0.087**
	(0.388)	(0.294)	(0.043)
Has used ChatGPT for writing	0.394	0.351	-0.043
	(0.491)	(0.479)	(0.063)
Observations	99	148	247

Notes: Standard errors in parentheses. *$p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A2

	(1)	(2)	(3)
Human-written	0.246	0.167	
	(0.172)	(0.278)	
Informed	0.207		
	(0.179)		
Believed human authorship			1.115^{***}
			(0.300)
Age	0.000225	-0.00448	-0.00447
	(0.00706)	(0.0121)	(0.0122)
Has used ChatGPT for writing	-0.0480	0.115	0.122
	(0.180)	(0.287)	(0.288)
Paragraph B	-0.246	-0.152	-0.299
	(0.172)	(0.276)	(0.277)
Observations	500	188	188

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: The dependent variable in each of these ordered logistic regressions is our measure of Trust. Column (1) includes all participants, while columns (2)-(3) only include participants in the Uninformed treatments. "Human-written" is a binary variable that =1 if a participant reviewed human-written text. "Informed" is a binary variable that =1 if a participant was informed about the authorship of the text they reviewed. "Believed human authorship" is a binary variable that =1 if a participant (in the Uninformed treatments) guessed the text they reviewed was human-written. "Age" is measured in years. "Has used ChatGPT for writing" is a binary variable that =1 if a participant reports having used ChatGPT for writing. "Paragraph B" is a binary variable that =1 if a participant saw the second set of paragraphs we included for robustness.

B Experiment Instructions

The experiment began with an introduction and consent form. Figure B1 shows the first screen that subjects saw after consenting to participate. Next, subjects saw their assigned paragraph for the first time. Pictured in B2 is the second topic (Paragraph B) written by AI, in the informed treatment.

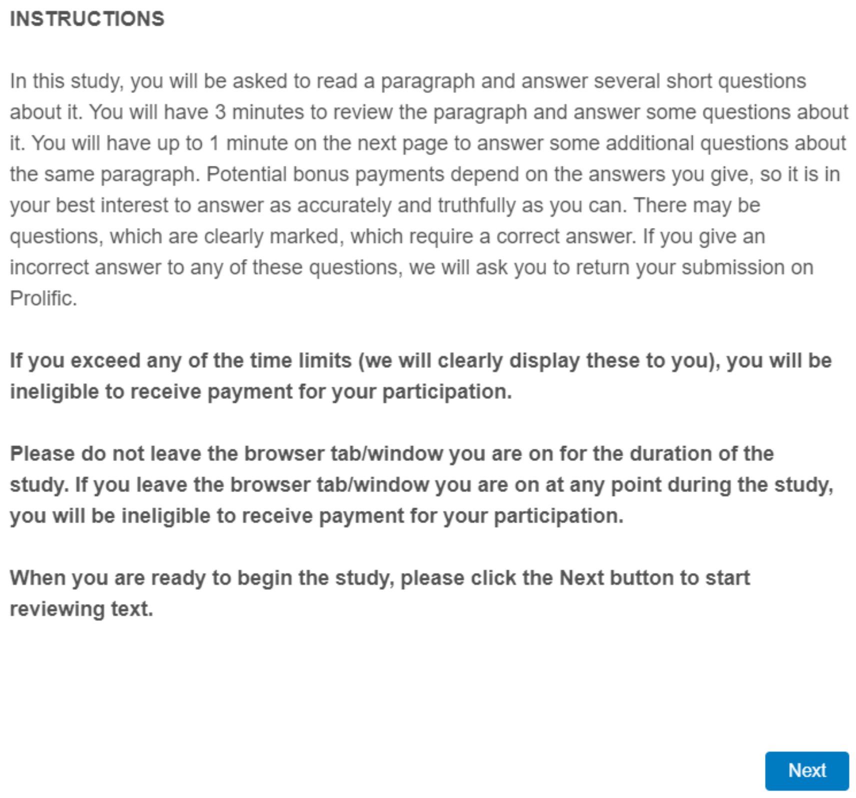
Figure B3 shows the questions that subjects answered about the paragraph before they had to consider accuracy. In the uninformed treatment, subjects were paid if their prediction about authorship was correct, but the answer was not given to them.

Subjects clicked "Next" to advance to the page shown in figure B4. This alerts them to the fact that they are about to answer questions about the paragraph under time pressure.

The next page explained how the financial incentive worked for the belief elicitation about accuracy, shown in figure B5. Subjects had to answer comprehension questions about the financial consequences of a fact-check.

The screen where subjects indicate how confident they are in the factual accuracy of the paragraph is shown in figure B6.

Figure B1: First Instructions After giving consent to participate in research



INSTRUCTIONS

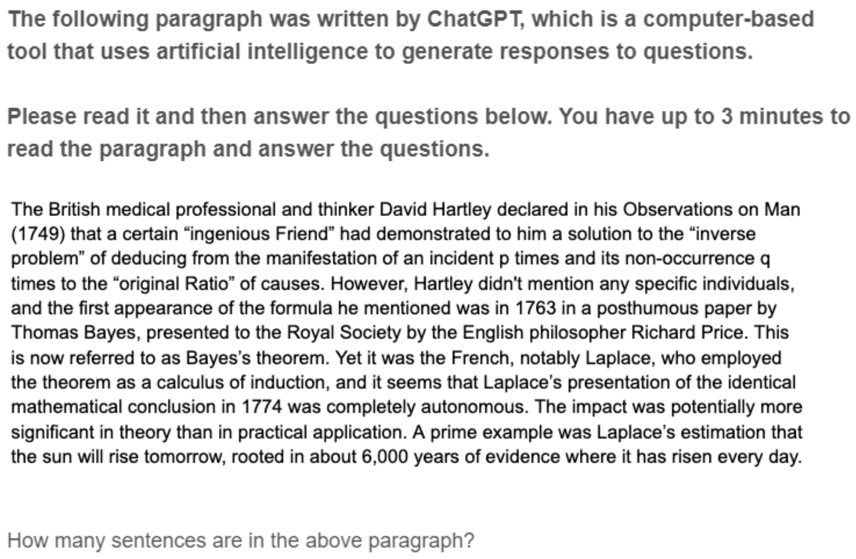
In this study, you will be asked to read a paragraph and answer several short questions about it. You will have 3 minutes to review the paragraph and answer some questions about it. You will have up to 1 minute on the next page to answer some additional questions about the same paragraph. Potential bonus payments depend on the answers you give, so it is in your best interest to answer as accurately and truthfully as you can. There may be questions, which are clearly marked, which require a correct answer. If you give an incorrect answer to any of these questions, we will ask you to return your submission on Prolific.

If you exceed any of the time limits (we will clearly display these to you), you will be ineligible to receive payment for your participation.

Please do not leave the browser tab/window you are on for the duration of the study. If you leave the browser tab/window you are on at any point during the study, you will be ineligible to receive payment for your participation.

When you are ready to begin the study, please click the Next button to start reviewing text.

Figure B2: AI-written Paragraph B



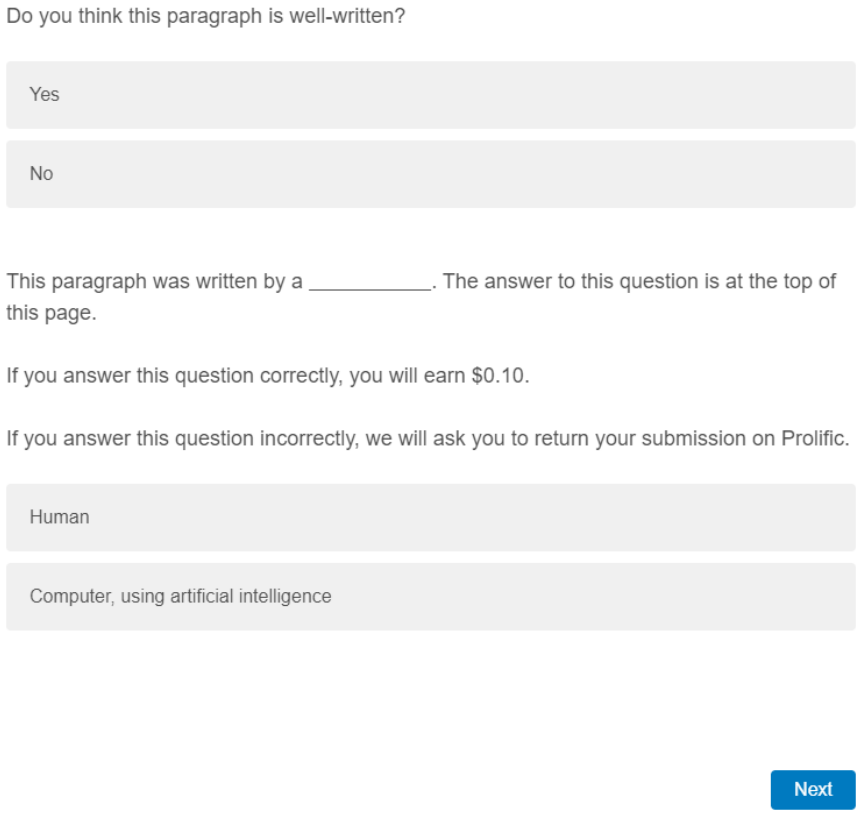
The following paragraph was written by ChatGPT, which is a computer-based tool that uses artificial intelligence to generate responses to questions.

Please read it and then answer the questions below. You have up to 3 minutes to read the paragraph and answer the questions.

The British medical professional and thinker David Hartley declared in his Observations on Man (1749) that a certain “ingenious Friend” had demonstrated to him a solution to the “inverse problem” of deducing from the manifestation of an incident p times and its non-occurrence q times to the “original Ratio” of causes. However, Hartley didn't mention any specific individuals, and the first appearance of the formula he mentioned was in 1763 in a posthumous paper by Thomas Bayes, presented to the Royal Society by the English philosopher Richard Price. This is now referred to as Bayes’s theorem. Yet it was the French, notably Laplace, who employed the theorem as a calculus of induction, and it seems that Laplace’s presentation of the identical mathematical conclusion in 1774 was completely autonomous. The impact was potentially more significant in theory than in practical application. A prime example was Laplace’s estimation that the sun will rise tomorrow, rooted in about 6,000 years of evidence where it has risen every day.

How many sentences are in the above paragraph?

Figure B3: First impression of the paragraph and beliefs about authorship



Do you think this paragraph is well-written?

Yes

No

This paragraph was written by a ___________. The answer to this question is at the top of this page.

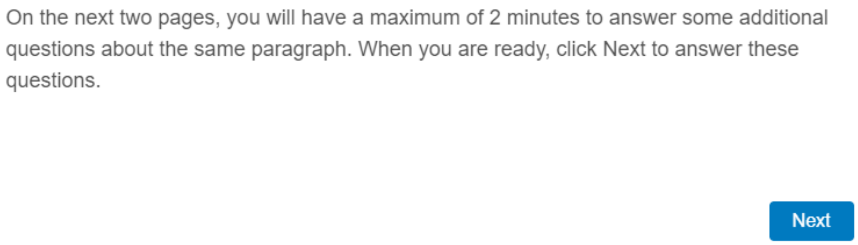
If you answer this question correctly, you will earn $0.10.

If you answer this question incorrectly, we will ask you to return your submission on Prolific.

Human

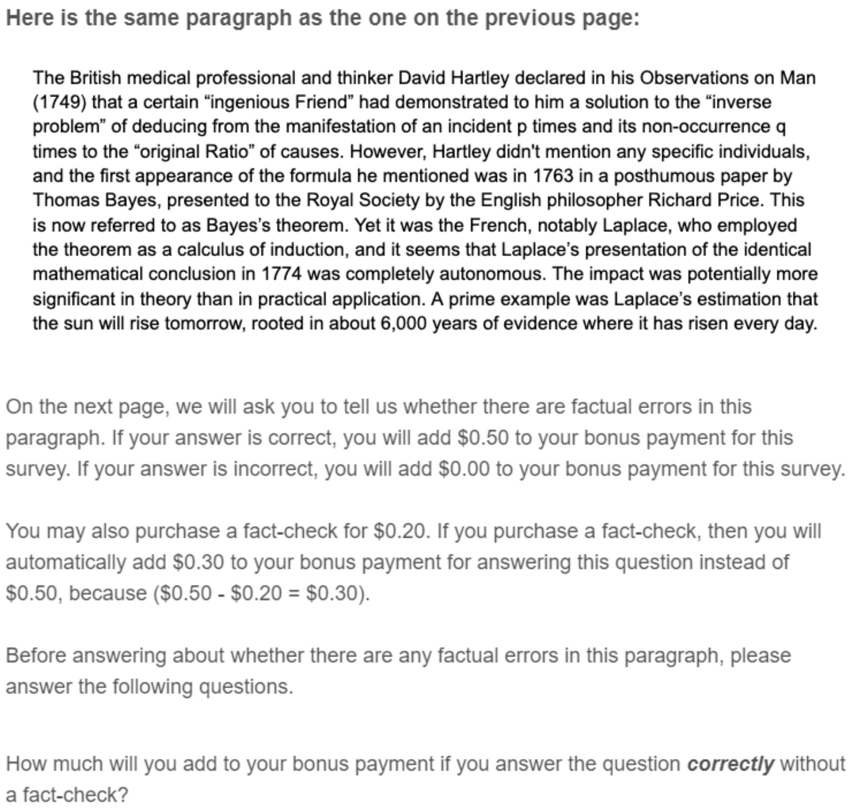
Computer, using artificial intelligence

Figure B4: Page between timed questions



On the next two pages, you will have a maximum of 2 minutes to answer some additional questions about the same paragraph. When you are ready, click Next to answer these questions.

Figure B5: Comprehension questions about the fact check



Here is the same paragraph as the one on the previous page:

The British medical professional and thinker David Hartley declared in his Observations on Man (1749) that a certain “ingenious Friend” had demonstrated to him a solution to the “inverse problem” of deducing from the manifestation of an incident p times and its non-occurrence q times to the “original Ratio” of causes. However, Hartley didn't mention any specific individuals, and the first appearance of the formula he mentioned was in 1763 in a posthumous paper by Thomas Bayes, presented to the Royal Society by the English philosopher Richard Price. This is now referred to as Bayes’s theorem. Yet it was the French, notably Laplace, who employed the theorem as a calculus of induction, and it seems that Laplace’s presentation of the identical mathematical conclusion in 1774 was completely autonomous. The impact was potentially more significant in theory than in practical application. A prime example was Laplace’s estimation that the sun will rise tomorrow, rooted in about 6,000 years of evidence where it has risen every day.

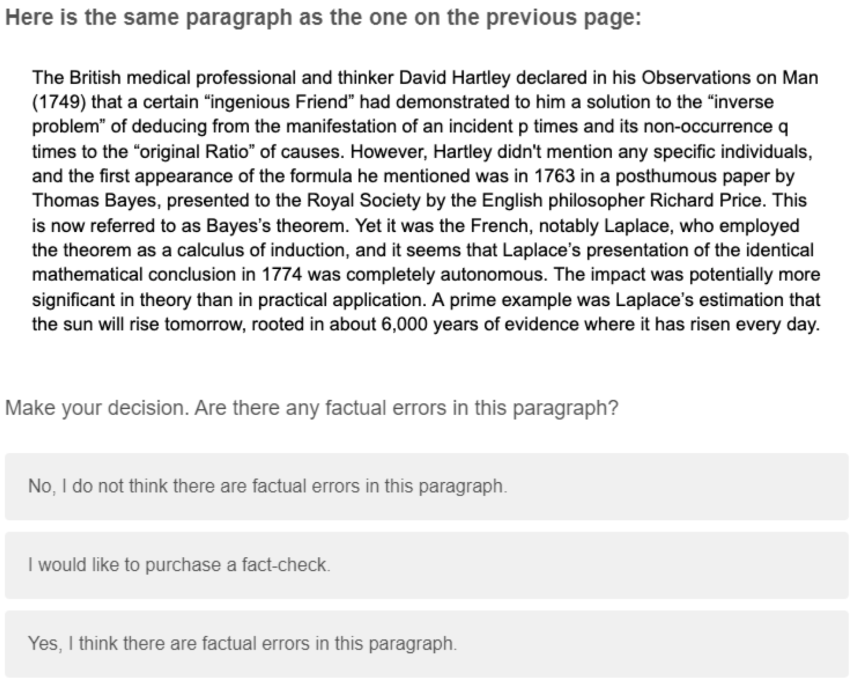
On the next page, we will ask you to tell us whether there are factual errors in this paragraph. If your answer is correct, you will add $0.50 to your bonus payment for this survey. If your answer is incorrect, you will add $0.00 to your bonus payment for this survey.

You may also purchase a fact-check for $0.20. If you purchase a fact-check, then you will automatically add $0.30 to your bonus payment for answering this question instead of $0.50, because ($0.50 - $0.20 = $0.30).

Before answering about whether there are any factual errors in this paragraph, please answer the following questions.

How much will you add to your bonus payment if you answer the question ***correctly*** without a fact-check?

Figure B6: Incentivized belief elicitation about accuracy



Here is the same paragraph as the one on the previous page:

The British medical professional and thinker David Hartley declared in his Observations on Man (1749) that a certain “ingenious Friend” had demonstrated to him a solution to the “inverse problem” of deducing from the manifestation of an incident p times and its non-occurrence q times to the “original Ratio” of causes. However, Hartley didn't mention any specific individuals, and the first appearance of the formula he mentioned was in 1763 in a posthumous paper by Thomas Bayes, presented to the Royal Society by the English philosopher Richard Price. This is now referred to as Bayes’s theorem. Yet it was the French, notably Laplace, who employed the theorem as a calculus of induction, and it seems that Laplace’s presentation of the identical mathematical conclusion in 1774 was completely autonomous. The impact was potentially more significant in theory than in practical application. A prime example was Laplace’s estimation that the sun will rise tomorrow, rooted in about 6,000 years of evidence where it has risen every day.

Make your decision. Are there any factual errors in this paragraph?

No, I do not think there are factual errors in this paragraph.

I would like to purchase a fact-check.

Yes, I think there are factual errors in this paragraph.