*Judgment and Decision Making* (2023), **Vol. 18**:e41 1–16
doi:10.1017/jdm.2023.37

EMPIRICAL ARTICLE

# Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation

Yunhao Zhang[1] and Renée Gosline[2]

[1]MIT Sloan, Berkeley Haas, Berkeley, CA, USA and [2]MIT Sloan, Cambridge, MA, USA

**Corresponding author: Yunhao Zhang; Email:** zyhjerry@mit.edu

**Received:** 13 July 2023; **Revised:** 10 September 2023; **Accepted:** 22 September 2023

**Keywords:** Generative AI; Algorithm Aversion; Human Favortism; Consumer Perception

**Abstract**

With the wide availability of large language models and generative AI, there are four primary paradigms for human–AI collaboration: human-only, AI-only (ChatGPT-4), augmented human (where a human makes the final decision with AI output as a reference), or augmented AI (where the AI makes the final decision with human output as a reference). In partnership with one of the world's leading consulting firms, we enlisted professional content creators and ChatGPT-4 to create advertising content for products and persuasive content for campaigns following the aforementioned paradigms. First, we find that, contrary to the expectations of some of the existing algorithm aversion literature on conventional predictive AI, the content generated by generative AI and augmented AI is perceived as of higher quality than that produced by human experts and augmented human experts. Second, revealing the source of content production reduces—but does not reverse—the perceived quality gap between human- and AI-generated content. This bias in evaluation is predominantly driven by human favoritism rather than AI aversion: Knowing that the same content is created by a human expert increases its (reported) perceived quality, but knowing that AI is involved in the creation process does not affect its perceived quality. Further analysis suggests this bias is not due to a 'quality prime' as knowing the content they are about to evaluate comes from competent creators (e.g., industry professionals and state-of-the-art AI) without knowing exactly that the creator of each piece of content does not increase participants' perceived quality.

## 1. Introduction

As large language models such as OpenAI's ChatGPT become increasingly accessible, generative artificial intelligence (GAI) is bound to revolutionize the way human beings work and live. While the academia and the public have expressed both concerns and excitement about this new technology (Botha & Pieterse, 2020; Clayton, 2023; Haupt & Marks, 2023; Khan, 2023; Li et al., 2023), the world is also curious about how GAI such as ChatGPT-4 would affect businesses and industries (Berg et al., 2023).

Recent research has shown that GAI could enhance labor productivity, for instance, in customer communication (Brynjolfsson et al. 2023) or essay writing (Noy & Zhang, 2023). While the cited studies focus on examining the impact of LLMs (i.e., generative pre-trained transformer, or GPT) on the *workers*, our research, to our knowledge, is the first to thoroughly study the perception of LLMs from the *people's* perspective (or the *consumer's* perspective). We hereby describe the novelty in our set-up compared with previous relevant research. First, existing research on people's perception of the content generated by LLMs has focused only on content generated solely by humans or AI, and they have not examined content produced by human–AI interactions. For example, Ayers et al. (2023) examined how healthcare professionals would evaluate responses to medical questions generated by physicians vs ChatGPT in an anonymized setting and found that responses generated by ChatGPT received higher-quality ratings. Nevertheless, in real life, the responses could be generated by human–AI interactions or human in the loop. For example, content generators may first obtain a response from ChatGPT as a reference before making their final decision (e.g., 'augmented human'), or they may enter their response as an input to ChatGPT along with the question prompt, letting ChatGPT edit their response and make the final decision ('augmented AI'). Our research compares the quality of persuasive content (i.e., advertising content[1] for products and persuasive content for campaigns, details described in Section 2) generated under all four paradigms: (1) human expert-only (i.e., professional content creators from one of the world's leading consulting firms create the content individually), (2) AI (ChatGPT-4)-only, (3) augmented human (i.e., a human expert makes the final decision on the output but is given the content first generated by ChatGPT-4 for the task, which they may edit or use as an inspiration), and (4) augmented AI (i.e., ChatGPT-4 makes the final decision on the output but is given the content first generated by a human expert, which it may edit or use as an inspiration). Furthermore, our set-up also allows us to shed light on the debate about whether humans or AI should make the final decision in our content generation context (McKendrick & Thurai, 2022).

After obtaining the generated content, we recruit and randomly assign online participants to rate the content quality in different conditions. In the baseline condition, participants are completely unaware of the content generation paradigms at all, thus basing their evaluations solely on the textual output (e.g., they are presented with the content without any mention of humans or AI throughout the study). We found that the content that ChatGPT-4 solely or ultimately determines the output is perceived as higher quality. This result is robust in a 'partially informed' condition when we add the contextual influence such that participants are informed of the content generation paradigms at the beginning of the survey, but they are unaware of the specific process for each piece. These results are somewhat surprising given predictions made by the literature mostly focused on traditional discriminative AI.[2] Frey and Osborne (2017) predict that creative tasks requiring creative and social intelligence, which are crucial knowledge in advertising and persuasion, will be the last to be automated. Castelo et al. (2019) suggest that people would prefer humans to AI in subjective task domains (e.g., composing a song and writing news articles). Our results suggest that people's perceptions of GAI might be different than previously postulated.

Additionally, our research also examines people's bias toward the content generation paradigms (i.e., given the same piece of content, whether knowing its creator affects people's evaluation). Liu et al. (2022) examined a similar question and found that—when writing emails to console others—the recipients display aversion toward the senders who use AI to write the message. Nevertheless, their study did not involve any emails actually generated by AI, but deceptively revealed and varied the human-generated messages to be either human-generated or AI-mediated.

In other words, since the content generated by humans and LLMs may have inherently different features and quality, their research is not about how people perceive content generated by LLMs, but how people perceive human-generated content being framed as AI-mediated. Our research examines potential bias toward human experts, AI, or human–AI collaboration without deception. About one-third of the recruited participants are randomly assigned to evaluate content quality in the 'informed' condition, in which they are not only informed upfront about the four paradigms as in the 'partially informed' condition, but they are also explicitly told under which paradigm a piece of content is generated when they evaluate the content. By comparing the baseline and the 'informed' condition,

[1] Advertising content also falls into the category of persuasive content because the goal of advertisement is to persuade people to be interested in the product.

[2] Discriminative AI is best suited for classification and prediction tasks, while generative AI, such as ChatGPT, can be used to produce content, rather than analyze it (Berg et al., 2023).

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

we find that people perceive the same piece of content generated solely by a human expert as of higher quality (e.g., state a higher level of satisfaction and higher willingness to pay) if they are aware the content is generated solely by a human expert. However, we do not find this change in perception among the other three paradigms. Furthermore, this phenomenon is robust when we compare the 'partially informed' condition with the 'informed' condition. Therefore, contrary to the arguments made by the 'algorithm aversion' literature (e.g., Castelo et al., 2019; Dietvorst et al., 2015) that people display aversion toward AI, we do not find aversion toward AI's involvement in generating advertising and persuasive content. Rather, we have evidence for human favoritism. In addition, further analysis suggests that this favoritism is unlikely to be driven by a quality prime (e.g., knowing that the content creators are competent). As there is no increase in the perceived quality in the 'partially informed' condition relative to the baseline, this means that knowing the creators are competent (e.g., top industry professionals and state-of-the-art AI) does not increase perceived quality.

However, even with this bias, the content generated by ChatGPT-4—when it makes the sole or final decision on the output—is still perceived as on par or better than human-generated content. Nevertheless, our results by no means suggest that LLMs should completely replace human agents—especially human oversight, which we discuss in Section 5 of our paper.

We provide a brief overview of our research design and paper structure. To examine how people perceive content generated under the four paradigms previously outlined, we first need to create the content to be used as stimuli for the experiment. Section 2 offers a detailed explanation of how we engage professional content creators from one of the world's leading consulting firms, as well as ChatGPT-4, to produce advertising content for products and persuasive content for campaigns. Section 3 outlines the methods and procedures for our 3x4 between-subjects experiments. In these experiments, participants evaluate content produced under one of the four paradigms (human-only, AI-only, augmented human, or augmented AI) and within one of the three conditions (baseline, partially informed, or fully informed). These conditions vary based on whether participants are informed about the content creation paradigms. Section 4 presents the experimental results, while Section 5 discusses the implications of our findings and concludes the paper.

## 2. Content generation process

In this section, we describe in detail how the content used for studies is generated. First, we pre-select five retail products (e.g., air fryer, projector, electric bike, emergency kit, and tumbler) from a retail website and five campaign goals that are uncontroversially benign (e.g., stop racism, start recycling, do more physical exercises, wash hands more often, and eat less junk food). Readers may find screenshots of the instructions given to the professional human content generators in Supplementary Information (SI) Section 1. All content generated under the four paradigms is included in SI Section 2.

### *2.1 Content generated by human expert individually and human expert with access to content first generated by ChatGPT-4*

We enlist ten professional content creators from one of the world's leading consulting firms (over $175 billion USD market capitalization) to generate advertising content for the products and persuasive content for the campaigns. This research context offers unique advantages that facilitate the exploration of our research questions. First, the ten content creators are industry professionals who have experience with writing advertising content for corporate products and campaign messages for NGOs.[3] One of our research questions is to compare the quality of content generated by human experts and AI. Engaging top-tier professionals for this purpose grants our study a significant degree of external validity, more so

[3] When deciding on the task prompts, we worked with our liaisons to ensure the participating content creators had past experience and expertise in this type of task.

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

than if we were to recruit subjects from platforms such as MTurk or Upwork.[4] Second, according to our liaisons from the firm, these content creators took the tasks very seriously. The department from which we sourced these creators was aware of their higher-level managers' interest in this study, instilling an intrinsic motivation to excel in the task. Moreover, participation was strictly voluntary, ensuring that those involved were genuinely interested in contributing to our research. Our liaisons also facilitated a one-hour meeting with the ten content creators during regular working hours to distribute the tasks via Qualtrics links.[5] These links remained accessible for one day, allowing the creators ample time to engage with the tasks. Overall, we are able to compare the performance of high-quality real industry professionals who write advertising and persuasive content as a part of their daily jobs with that of ChatGPT-4.

Each content creator completed two content generation tasks—one advertising content for one of the five products and one persuasive content for one of the five campaigns. For the campaign tasks, the experts were provided with the following instructions: '*Your task is to write persuasive content for a campaign in fewer than 100 words. Your goal is to persuade people to change their behavior after seeing your content'*. After each expert had been assigned a campaign, they were asked '*please create persuasive content for a client (e.g., an NGO doing a campaign) to convince people to [perform the action advocated by the campaign]*[6] *in fewer than 100 words'*.

For the advertising content task, the experts were provided with these instructions: '*The task is to write advertising content for a product in fewer than 100 words without relying on LLMs. Your goal is to get people interested in this product after seeing your content'*. After each expert had been assigned a product, they were given a picture of the product along with a description of the product's features sourced directly from the product description section of a retail website. The experts were then prompted '*please create advertising content for the above product in fewer than 100 words'*. The 100-word limit was chosen as it approximates the amount of text that can be effectively communicated within a 30-second span.

Very importantly, the first task the experts completed, be it about the product or the campaign, required the content creators to perform the task without using LLMs.[7] This resembled the typical human content generation paradigm before LLMs became popular. For the second task, however, the content generators were shown the content generated (solely) by ChatGPT-4. Before beginning their second task, they were informed '*we will provide you with the textual content generated by ChatGPT-4, which you may use either as inspiration for your own content or as a first draft that you can edit upon.*

*In order to obtain the persuasive / advertising content generated by ChatGPT-4, we provided ChatGPT-4 with a prompt on the campaign topic, and asked it to create persuasive content in fewer than 100 words / we provided ChatGPT-4 with a prompt for the product to be advertised (including the product's name and feature descriptions), and asked it to create an advertisement in fewer than 100 words. Both the input prompt and the output by ChatGPT-4 will be available to you on the next page'*.

Then, in addition to the aforementioned prompt for the first task, the experts are also presented with content generated solely by ChatGPT-4. The content produced during the second task will be referred to as the 'augmented human' content generation paradigm in our paper, distinguishing it from the 'human-only' paradigm. This is because, although the final output was determined by the humans, they were assisted by AI. In summary, we obtained twenty pieces of content generated by ten professional content creators—ten tasks (i.e., five products + five campaigns) multiplied by two content per person (i.e., one with the assistance of ChatGPT-4 and one without.)

[4] Anecdotally, we asked our liaison about the cost of recruiting these experts to do the tasks without the partnership. The manager responded 'probably not affordable'.

[5] This was done also for the purpose of maintaining their anonymity.

[6] In the bracket is one of the five campaigns mentioned previously.

[7] This was made explicitly clear by the instructions in our survey and our liaison when they held an internal meeting during which they were randomly assigned the tasks.

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

### 2.2 *Content generated by ChatGPT-4 solely and ChatGPT-4 with access to content first generated by a human expert*

We paid a $20 monthly subscription fee to obtain access to ChatGPT-4. While each human expert generated two pieces of content, ChatGPT-4 generated twenty pieces of content (ten without access to the content first generated by a human expert and ten with). We presented ChatGPT-4 with prompts nearly identical to those given to human experts. The prompt for the campaign was as follows: '*please create persuasive content for a client (e.g., an NGO doing a campaign) to convince people to [perform the action advocated by the campaign] in fewer than 100 words'*. The prompt for the products was '*please create advertising content to get people interested in the following product in fewer than 100 words. The product is [the product's name]. For your information, the product has the following features: [the product's features taken from the retail website]'*. The outputs generated following these prompts were the AI-generated content we gave to the human experts in their second task as mentioned previously.

In addition, we took the first response generated under each prompt as the output used in our studies. We understand that ChatGPT is known to give stochastic outputs. Therefore, after we have obtained all the responses, we tried asking ChatGPT the same prompt several times in a row. We notice that the outputs—though using different words—are quite similar in structure and overall arguments. Therefore, we highly doubt that using the output from, say, the fifth try instead of the first try would significantly affect our results. Furthermore, using the later responses affects internal validity as we are interested in knowing how ChatGPT performs against human experts. Using responses from repeated prompting can be viewed as the researchers influencing or helping ChatGPT, since ChatGPT could take repeated prompting as a sign that the researchers are not satisfied with its current output.

When we also gave ChatGPT-4 the content generated solely by a human expert for the same campaign or product, the prompt had the following additional paragraph: '*Below is the advertising/persuasive content generated by a professional content creator from one of the world's best consulting firms, which you may use as inspiration or a first draft you edit upon when generating your own content'*. Then, the content generated by a human expert without access to content generated by ChatGPT-4 is appended below the paragraph as part of the prompt so that the human experts may refer to the AI's output while creating their own. We will refer to the content generated in response to this type of prompt as the 'augmented AI' content generation paradigm in our paper. This term distinguishes it from the 'AI-only' paradigm because, even though the AI made the final output decision, it had human expert's assistance.

## 3. Methods

### 3.1 *Study design*

The study protocol and all main analyses are pre-registered at https://aspredicted.org/2qg6y.pdf. The data and codes are also available in the supplement. Readers may find screenshots of the instructions given to the online participants of this study in SI Section 3. Participants were randomly assigned to one of the three different conditions: the baseline condition, the 'partially informed' condition, and the 'informed' condition. In the baseline condition, participants evaluate content quality completely ignorant of the context, meaning they do have any knowledge of the content creation paradigms (i.e., there is no mention of humans, AI, or human–AI collaboration throughout the study. Their judgment of content quality should be solely based on the textual output). In this condition, we are able to measure the quality of the content created under each paradigm without any contextual influence.

In the 'partially informed' condition, participants rate the content quality with partial knowledge of the content creation paradigms. Participants in this condition are briefed about the four content creation paradigms at the beginning of the survey. In particular, participants are informed that '*the textual content for the products and campaigns you are about to evaluate were generated under one of the four paradigms*' and then they are given the definition of the content generation paradigms (e.g., AI is the state-of-the-art language model ChatGPT-4 developed by OpenAI, and human experts are

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

recruited from one of the world's best consulting firms) and the information available to the content creators (see SI Section 3 'Page 2 of the instructions' for a screenshot). However, very importantly, participants will not know *exactly* how each piece of content they evaluate is created as they rate each content. This condition resembles real-life situations such as when people are reading a news article. This condition mirrors real-life situations such as reading a news article where readers might be aware that some articles are AI-generated, but they cannot be sure whether the article they are reading is human, AI, or human–AI-produced. This condition allowed us to measure how people assess content when AI might be involved in the content generation process.

In the 'informed' condition, participants rate each content's quality with full knowledge of the content creation paradigms. Not only are participants briefed about the content generation paradigm participants as those in the 'partially informed' condition, but they also know exactly how each piece of content they evaluate is created. This condition enabled us to determine whether any potential bias toward a specific content generation paradigm exists. By comparing perceived content quality between the baseline and 'informed' conditions, we could examine whether awareness (or lack thereof) of the content generation paradigm for a given piece of content affects people's quality evaluation. For instance, suppose that given the same piece of content produced solely by a human expert, participants perceive the content to be of higher quality when they are aware that it was solely human-generated compared with when they are not aware. This would clearly indicate a bias in favor of content created under the 'human-only' paradigm. In addition to the comparison with the baseline for all four paradigms, we also compared the 'informed' with the 'partially informed' condition.

### 3.2 Participants

All participants were recruited from the research panel platform CloudResearch Connect. A total of 1212 participants entered our survey, and nine participants failed an attention check and were not allowed to finish the survey. All remaining 1203 participants (50% female, M_age = 38) who finished the survey are included in analyses. The median time to complete the survey was 10.3 minutes. The survey completion fee was $1.5.

### 3.3 Procedures

We had a between-subjects 3-by-4 design: three conditions (baseline, partially informed, and informed) and four content generation paradigms (human-only, AI-only, augmented human, or augmented AI). After signing the consent form, all participants initially received the same survey overview. It stated, '*we have generated some advertising content for five different products and some persuasive content for five different campaigns. We want you to evaluate the quality of the text you will read'*. Participants in the 'partially informed' and 'informed' conditions were further informed about the four content generation paradigms as described previously. Those randomized into the baseline condition skipped this step, thus remaining entirely unaware of the content generation paradigms. After passing the first attention check, all participants were informed that they would first evaluate the quality of five advertising content for five different products and given a description of the outcome measures for quality evaluation:

(Satisfaction) '*suppose you are the seller of the product, to what extent are you satisfied or unsatisfied with the generated advertising content on a scale from 1 to 7*' with 1 being very unsatisfied and 7 being very satisfied.

(Willingness to pay) '*suppose you are the seller of the product and the content has a copyright, what is the maximum amount of money you are willing to pay to use the content as your advertisement? (Assuming you have a budget of $1000, please state your max willingness-to-pay between $1 and $1000)'*.

(Interest) '*to what extent you are interested in learning more about the product (e.g., its price, user reviews, complete product descriptions, etc.) on a scale from 1 to 7*' with 1 being not interested at all and 7 being very interested.

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

Participants then proceeded to rate the content quality for the five products in a random order. They were given a picture of each product and the piece of advertising content generated by one of the four paradigms for the product when they provided responses to the three key outcome measures. Very importantly, within each of the three conditions, all participants provided quality evaluations for content produced under only one of the four paradigms (e.g., if a participant were randomly assigned to rate the quality of content generated solely by human experts, they would only be given content generated by human experts throughout the study). Participants in the 'informed' condition were additionally informed of under which paradigm the content they were viewing was generated[8], but participants in the baseline or the 'partially informed' condition were not. After evaluating the advertising content, participants needed to pass another attention check. Then, they were told that they would next evaluate the quality of five persuasive content for five different campaigns in a random order. They were again given a description of the key outcome measures. While the first two outcome measures were the same, the third outcome measure was as follows:

(Persuasion) '*to what extent you are convinced by the above content to [perform the action advocated by the campaign] on a scale from 1 to 7'* with 1 being not convinced at all and 7 being very convinced.

After rating the quality of the advertising content and reading the outcome measures used for measuring the quality of persuasive content, participants proceeded to rate the quality of the persuasive content generated by one of the four paradigms for the five campaigns in a random order. Same as before, participants in the 'informed' condition were additionally made aware of under which paradigm the content they were viewing was generated, but participants in the baseline or the 'partially informed' condition were not. Participants then answered some demographic questions before finishing the study.

In summary, with approximately 1200 total participants, we obtained around 100 evaluations for content generated under each of the four paradigms in each of the three conditions.

## 4. Results

### *4.1 Purely evaluating content quality*

We first examine participants' stated level of satisfaction with the content generated under different paradigms when their focus is solely on the content quality. As depicted by the baseline condition in Figure 1, participants expressed varying degrees of satisfaction with content produced under different paradigms (one-way ANOVA, $p$ = 0.000). Furthermore, on average, content generated solely by ChatGPT-4 resulted in the highest satisfaction level, and it is on par with content generated by augmented AI (see Row 1 of Table 1 for statistics). However, the content generated by a human expert resulted in a similar level of satisfaction as content generated by an augmented human expert (see Row 2 of Table 1 for statistics). Interestingly, content generated when AI makes the sole or final decision on the output resulted in a higher satisfaction level compared with content generated when a human expert makes the sole or final decision on the output (see Row 3 and Row 4 of Table 1 for statistics).

The correlation between the level of satisfaction and log willingness to pay is 0.51. Similar patterns emerge when we use participants' willingness to pay for the content as a measure of content quality. As depicted by the baseline condition in Figure 2, participants had a varying willingness to pay for content generated under different paradigms (one-way ANOVA, $p$ = 0.001). Furthermore, on average participants' willingness to pay was almost the same for content generated by AI or augmented AI (see Row 5 of Table 1 for statistics). However, the content generated by human experts and augmented human experts had a similar willingness to pay (see Row 6 of Table 1 for statistics). Consistent with the previous results, content generated when AI makes the sole or final decision on the output resulted in a higher willingness to pay compared with content generated when a human expert makes the sole

[8] The additional information subjects in the informed condition receive when evaluating the content quality for advertising and campaign messages is the following sentence '*Below is the advertising / persuasive content created by [one of the four paradigms]'*.

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

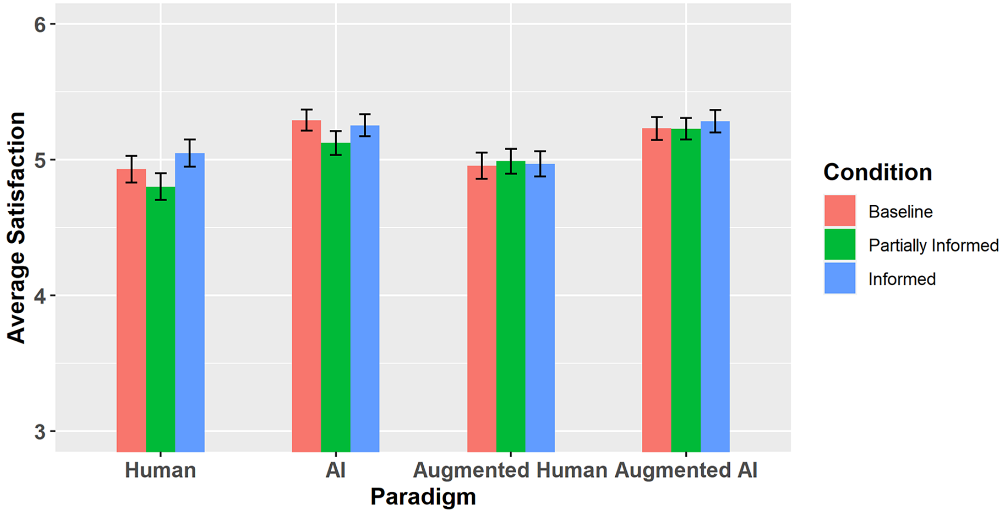

***Figure 1.*** *X-axis is the content generation paradigm: human expert-only, AI-only, a human expert who finalizes the content first generated by AI, and an AI that finalizes the content first generated by a human expert. The y-axis is the subjects' average level of satisfaction, pooling all ten contents together for each paradigm. It starts from 3 instead of 0 for better visualization. The colors represent the different conditions. The bars indicate 95% confidence intervals.*

***Table 1.*** *Comparisons of subjects' average level of satisfaction and average log willingness to pay (WTP) among the four paradigms using a two-sided two-sample t-test in the baseline condition. The test statistics are all based on the first-mentioned paradigm minus the second-mentioned paradigm (e.g., for 'human expert' vs. 'augmented human expert', the 't = −0.36' is the first minus the second).*

| | Paradigm comparison | DV | Averages | *t*-value | *p*-value |
|---|---|---|---|---|---|
| 1 | AI vs augmented AI | Satisfaction | 5.29 vs 5.23 | 1.04 | 0.30 |
| 2 | Human expert vs augmented human expert | Satisfaction | 4.93 vs 4.95 | −0.36 | 0.72 |
| 3 | AI vs human expert | Satisfaction | 5.29 vs 4.93 | 5.65 | 0.000 |
| 4 | Augmented AI vs augmented human expert | Satisfaction | 5.23 vs 4.95 | 4.25 | 0.000 |
| 5 | AI vs augmented AI | log (WTP) | 4.83 vs 4.85 | −0.21 | 0.83 |
| 6 | Human expert vs augmented human expert | log (WTP) | 4.61 vs 4.59 | 0.27 | 0.79 |
| 7 | AI vs human expert | log (WTP) | 4.83 vs 4.61 | 2.52 | 0.01 |
| 8 | Augmented AI vs augmented human expert | log (WTP) | 4.85 vs 4.59 | 3.03 | 0.003 |

or final decision on the output (see Row 7 and Row 8 of Table 1 for statistics). Our results suggest that, although the size of the difference is not large, AI has the capability to produce creative content with higher quality.

Since the third outcome measures are different for products (level of interest) and campaigns (degree of persuasion), we defer all our analysis regarding the third outcome measures to Section 4.5 in which we separately examine the two task categories (i.e., product advertisement versus campaign messages).

### *4.2 Evaluating content quality with partial knowledge of AI's potential involvement*

In the real world, while human-to-human interaction remains the norm, people might sometimes wonder whether they are interacting with a human or an AI agent (e.g., could this message or news article have been written by AI?). Therefore, we examine participants' stated level of satisfaction when

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

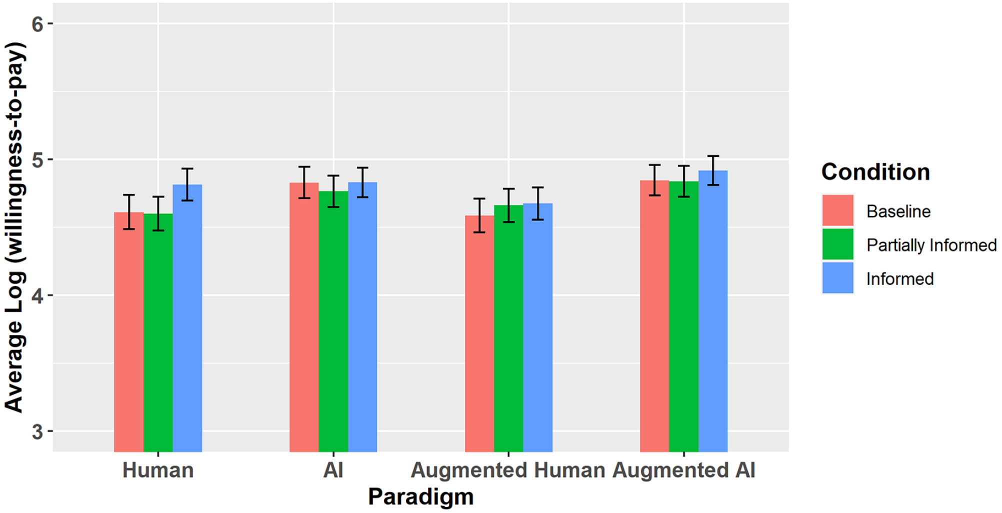

***Figure 2.*** *X-axis is the content generation paradigm: human expert-only, AI-only, a human expert who finalizes the content first generated by AI, and an AI that finalizes the content first generated by a human expert. The y-axis is the average of the logarithm of subjects' stated willingness to pay for the content (pooling all ten contents together for each paradigm). It starts from 3 instead of 0 for better visualization. The colors represent the different conditions. The bars indicate 95% confidence intervals.*

they are aware that the AI could *potentially* be involved in the content generation process. The crucial difference between the baseline condition and this 'partially informed' condition is that participants in the baseline are completely ignorant of AI's potential involvement in content generation; hence, the identity of the content creators is unlikely to be a factor affecting their judgment. However, since participants are not informed of how exactly each content they see is generated, we hereby examine the contextual effect of potential AI involvement on their evaluation.

As depicted by the 'partially informed' condition in Figure 1, participants had different levels of satisfaction with content generated under different paradigms (one-way ANOVA, $p = 0.000$). On average, content generated by augmented AI resulted in the highest satisfaction level, and it is on par with content generated solely by AI (see Row 1 of Table 2 for statistics). However, the content generated by an augmented human expert was better than the content generated solely by a human expert (see Row 2 of Table 2 for statistics).[9] Similar to the baseline condition, content generated when AI makes the sole or final decision on the output resulted in higher satisfaction level compared with content generated when a human expert makes the sole or final decision on the output (see Row 3 and Row 4 of Table 2 for statistics).

Furthermore, as depicted by the 'partially informed' condition in Figure 2, participants had different willingness to pay for content generated under different paradigms (one-way ANOVA, $p = 0.02$). The willingness to pay for content is nearly identical when a human expert makes the sole or final decision on the output (see Row 6 of Table 2 for statistics). Similarly, the willingness to pay is almost the same for content generated when AI made the sole or final decision (see Row 5 of Table 2 for statistics). Nevertheless, the willingness to pay for content generated when AI made the sole or final decision on the output is still slightly higher than that for content generated by human experts or augmented human experts (see Row 7 and Row 8 of Table 2 for statistics).

[9] Although the difference in the level of satisfaction between content generated by human experts and augmented human experts is statistically significant, we need to be cautious in interpreting the result because (1) the effect is not present when we examine willingness to pay (see Row 6 of Table 2), level of interest, or degree of persuasion (see Figure 5); (2) this effect does not exist in the baseline condition.

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

***Table 2.*** *Comparisons of subjects' average level of satisfaction and average log willingness to pay (WTP) among the four paradigms using a two-sided two-sample t-test in the 'partially informed' condition. The test statistics are all based on the first-mentioned paradigm minus the second-mentioned paradigm.*

| | Paradigm comparison | DV | Averages | *t*-value | *p*-value |
|---|---|---|---|---|---|
| 1 | AI vs augmented AI | Satisfaction | 5.23 vs 5.12 | 1.72 | 0.09 |
| 2 | Human expert vs augmented human expert | Satisfaction | 4.80 vs 4.99 | −2.73 | 0.006 |
| 3 | AI vs human expert | Satisfaction | 5.12 vs 4.80 | 4.76 | 0.000 |
| 4 | Augmented AI vs augmented human expert | Satisfaction | 5.22 vs 4.99 | 3.87 | 0.0001 |
| 5 | AI vs augmented AI | log (WTP) | 4.76 vs 4.84 | −0.88 | 0.36 |
| 6 | Human expert vs augmented human expert | log (WTP) | 4.60 vs 4.66 | −0.7 | 0.47 |
| 7 | AI vs human expert | log (WTP) | 4.76 vs 4.60 | 1.90 | 0.054 |
| 8 | Augmented AI vs augmented human expert | log (WTP) | 4.84 vs 4.66 | 2.08 | 0.03 |

### *4.3 Is there any bias toward any content creation paradigm?*

**Baseline vs informed**

Next, we explore whether individuals show any bias toward any of the content generation paradigms by comparing how participants' subjective content quality evaluations differ when they are fully aware (the 'informed' condition) versus completely ignorant (i.e., the baseline condition) of the content generation process. We will have evidence for bias if, for example, given the same piece of content, people express greater or less satisfaction or willingness to pay if they are informed the content is generated by a human expert. The results are visually illustrated by a comparison between the 'pinkish' bar (baseline, on the left) and the 'bluish' bar (informed, on the right) within each paradigm in Figures 1 and 2. To analyze this more systematically, and following our pre-registration, within each of the four paradigms, we (separately) regress the dependent variable (i.e., level of satisfaction or log willingness to pay) on the dummy variable indicating the condition (1 = informed, 0 = baseline) with task fixed effects.[10] A positive (negative) coefficient indicates favoritism (aversion) toward a particular content generation paradigm (throughout the text, the reported regression coefficient '*b*' refers to the original regression coefficient).

First, we find that given the same content generated solely by a human expert, participants felt more satisfied with the content ($b = 0.09$, $t = 2.96$, $p = 0.003$) and were willing to pay more ($b = 0.18$, $t = 4.14$, $p = 0.0000$) if they were informed that the content was created solely by a human expert. However, we do not find any significant effect on the satisfaction level or willingness to pay for the other content generation paradigms (see the footnote for the regression coefficients for the other three paradigms).[11] In addition, the results are robust when we examine product advertisement and campaign messages separately (see SI Section 6 for detailed results). Therefore, we do not have any evidence for aversion toward AI or the involvement of AI. Instead, we have evidence for human favoritism in our context.[12]

[10] 'Task' refers to the ten pieces of content (in this particular regression, it refers to the ten pieces of human-generated content) participants provide their quality evaluations for. We pre-registered the fixed-effects model because (1) it increases the precision of our main estimator by controlling for unobserved heterogeneity across the ten pieces of content, which is desirable given our pre-registered sample size (based on our pilot study, we expect the effect size to be small); (2) it minimizes the potential risk of false negatives when we examine the other three paradigms—the null results when comparing the baseline and the informed condition within each of the other three paradigms are robust, even at higher levels of precision.

[11] AI: $b\_satisfaction$ = -0.05, $p$ = 0.35, $b\_wtp$ = -0.01, $p$ = 0.87; augmented human: $b\_satisfaction$ = -0.004, $p$ = 0.91, $b\_wtp$ = 0.1, $p$ = 0.23; augmented AI: $b\_satisfaction$ = 0.08, $p$ = 0.12, $b\_wtp$ = 0.12, $p$ = 0.10

[12] See SI Section 4 for a comparison of the partially informed and the informed condition.

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

### 4.4 Priming of quality or human favoritism?

Bar-Hillel et al. (2012) show that revealing the name of a highly regarded poet serves as a quality prime, which enhances the reading experience that subsequently results in a higher perceived quality of the poem. One might wonder whether a higher perceived quality of human-generated content in the 'informed' condition is due to a quality prime (i.e., knowing that the human creators are top industry professionals affects the 'evaluation experience') or due to 'biased favoritism' toward human experts. Our results support the latter. Although the participants are 'context-blind' in the baseline condition, they possess the knowledge—in the 'partially informed' condition—that the content they are about to evaluate is generated by human experts recruited from one of the world's best consulting firms, the state-of-the-art language model ChatGPT-4 developed by OpenAI, and their collaboration. If the effect is driven by a quality prime, we should expect the perceived quality to be higher in the 'partially informed' condition compared with the 'baseline' condition.[13] Nevertheless, this is clearly not the case, as one may observe from the comparisons between the pinkish bars and green bars in Figures 1 and 2. In addition, we use a similar regression as in the previous section to examine the results more analytically. Within the paradigm of human-generated content, we (separately) regress the dependent variable (i.e., level of satisfaction or log willingness to pay) on the dummy variable indicating the condition (1 = partially informed, 0 = baseline) with task fixed effects. The results contradict the alternative explanation that the observed human favoritism is due to a quality prime. For log willingness to pay, the regression coefficient on the dummy variable is −0.015 ($t = -0.2$, $p = 0.85$), which suggests that informing subjects about the content could be driven by human experts, or state-of-the-art AI, or their collaboration, which has almost no effect on willingness to pay. For the level of satisfaction, the regression coefficient on the dummy variable is −0.013 ($t = -3.24$, $p = 0.001$), which suggests that the quality prime, if anything, could lower—as opposed to increase—the level of satisfaction with the human-generated content relative to the baseline condition.

Furthermore, the perceived quality does not increase for the other three paradigms either. In addition, the results in Sections 4.3 and 4.4 are robust when we examine participants' level of interest in the products after seeing the generated advertisement and the degree to which they are persuaded after seeing the campaign messages. We also analytically examine the comparisons between the baseline and both the 'informed' and 'partially informed' conditions (more details provided in Sections 4.5). The regression models are similar to the ones used in Sections 4.3 and 4.4. Within human-generated content, we separately regress the level of interest in products and the degree of persuasion by campaign messages on the dummy variable (1 = informed, 0 = baseline) with task fixed effects. The positive coefficients suggest that there is an increase in the level of interest in the product ($b = 0.26$, $t = 3.95$, $p = 0.0000$) and the degree of persuasion by campaign messages ($b = 0.104$, $t = 2.3$, $p = 0.022$) from the baseline to the 'informed' condition for the content generated solely by a human expert. Nevertheless, when running the same regressions between the baseline and the 'partially informed' condition (i.e., the dummy variable is now coded as '1 = partially informed' and '0 = baseline'), we do not see a significant increase in the level of interest in the product ($b = -0.10$, $t = -2.41$, $p = 0.016$) and the degree of persuasion ($b = 0.06$, $t = 1.37$, $p = 0.17$). These results confirm that there is not a significant positive effect of a 'quality prime'. Therefore, we argue that the observed increase in the perceived quality—evident only when participants are explicitly informed that the content was generated by a human expert—is attributable to human favoritism.

### 4.5 Persuasive content for campaigns vs advertising content for products

One might wonder whether the primary results differ when we separately analyze persuasive content for campaigns and advertising content for products. While persuasive content for campaigns might necessitate a deeper understanding of human psychology and more creativity, advertising content in

[13] The assumption, which is mostly likely to be true, is that participants do not automatically infer that the content is generated by top industry professionals.

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

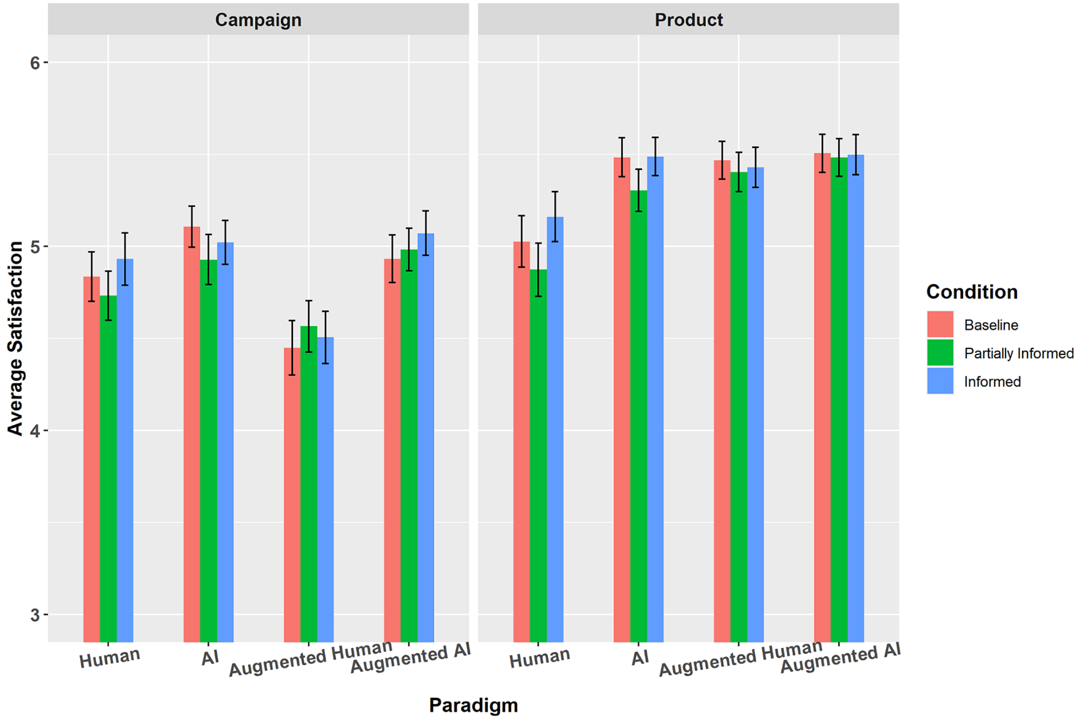

***Figure 3.*** *X-axis is the content generation paradigm. The y-axis is the subjects' level of satisfaction, pooling the five contents together for each paradigm given a task category. The left panel depicts persuasive contents generated for five campaigns, and the right panel depicts advertising contents generated for five products. The colors represent the different conditions. The bars indicate 95% confidence intervals. The y-axis starts from 3 instead of 0 for better visualization.*

our case tends to be more standardized, given that a significant portion of the text consists of product feature descriptions.

In the baseline and the 'partially informed' condition for both task categories, participants either felt at least as satisfied, or even more satisfied, with the content generated with AI's involvement than without. They were either willing to pay an equivalent amount, or more, for the content generated, and they became either equally interested or more interested in the product or persuaded to support the campaign when AI made the sole or final decision on the output. (The results are obtained by comparing the 'pinkish' bars across the paradigms in Figures 3–5; see SI Section 5 for detailed statistics). In addition, the correlation between ad satisfaction and product interest is 0.62. Between campaign message satisfaction and persuasion level, the correlation is 0.86. The correlation between log willingness to pay for the ad and product interest is 0.41. Between log willingness to pay for the campaign messages and persuasion level, the correlation is 0.50.

Interestingly, in the baseline condition, although the performance gap between human experts and AI is similar between the two categories, the gap between augmented human experts and augmented AI is smaller in content generated for products than for campaigns. For example, we separately regress the dependent variable (level of satisfaction or log willingness to pay) on the content generation paradigm (0 = augmented human, 1 = augmented AI), category (0 = campaign, 1 = product), and their interaction. The negative coefficient of the interaction term suggests a reduction in the gap in the level of satisfaction and willingness to pay between augmented human experts and augmented AI in the product category relative to the campaign category ($b\_satisfaction$ = −0.44, $t$ = −3.58, $p$ = 0.0003; $b\_wtp$ = −0.54,

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

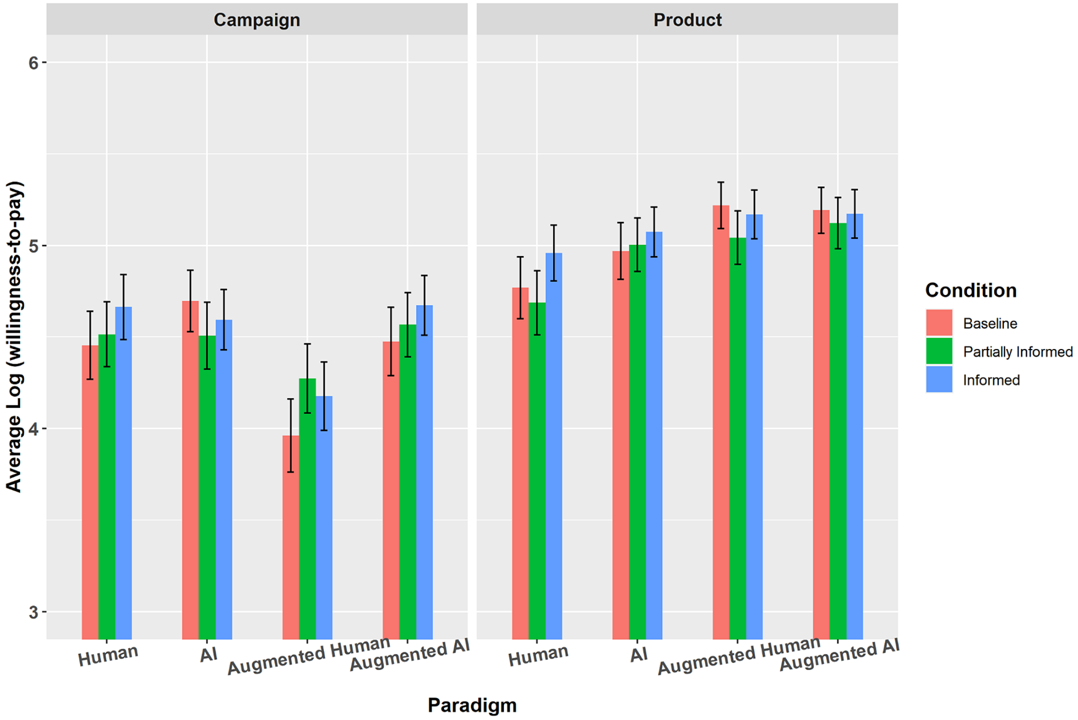

***Figure 4.*** *X-axis is the content generation paradigm. The y-axis is the average of the logarithm of subjects' willingness to pay, pooling the five contents together for each paradigm given a task category. The left panel depicts persuasive contents generated for five campaigns, and the right panel depicts advertising contents generated for five products. The colors represent the different conditions. The bars indicate 95% confidence intervals. The y-axis starts from 3 instead of 0 for better visualization.*

$t = -3.24$, $p = 0.001$). Furthermore, this effect is driven by an increase in the perceived quality of content generated by the 'augmented human' paradigm in the product category, rather than a decrease in the 'augmented AI' paradigm.

Furthermore, this result remains true when we run the above regression in the 'uninformed' condition. When the dependent variable is the level of satisfaction, the coefficient on the interaction is significantly negative ($b\_satisfaction = -0.34$, $t = -2.82$, $p = 0.005$), which suggests a reduction in the performance gap between augmented humans and augmented AI in the product category. When the dependent variable is log willingness to pay, the coefficient on the interaction is directionally negative ($b\_wtp = -0.21$, $t = -1.28$, $p = 0.20$).

Last but not least, we observe favoritism toward content generated solely by human experts for both task categories, but do not find consistent evidence of aversion[14] toward content generated with AI's involvement. In addition, the level of human favoritism is the same for the two task categories[15] (detailed statistics are provided in SI Section 6). To summarize the results in this section, although the primary outcomes are qualitatively similar when we examine the two task categories separately, we do

[14] The only exception where the data suggest bias (aversion) toward AI is that there is a decrease in the degree of persuasion from the baseline to the 'informed' condition for AI-solely-generated campaign messages (two-sample t-test, 5.07 vs 4.88, $p = 0.03$). However, this comparison is not significant if we compare the 'informed' condition with the 'partially informed' condition for AI-solely-generated campaign messages (two-sample t-test, 4.88 vs 4.81, $p = 0.49$).

[15] Within a content generation paradigm, we regress the DVs on the condition (1 = informed, 0 = baseline), task category (1 = product, 0 = campaign), and their interaction. Within the paradigm of human-generated content, the coefficient of the interaction term for satisfaction and log willingness to pay is 0.04 ($p = 0.79$) and -0.02 ($p = 0.91$).

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

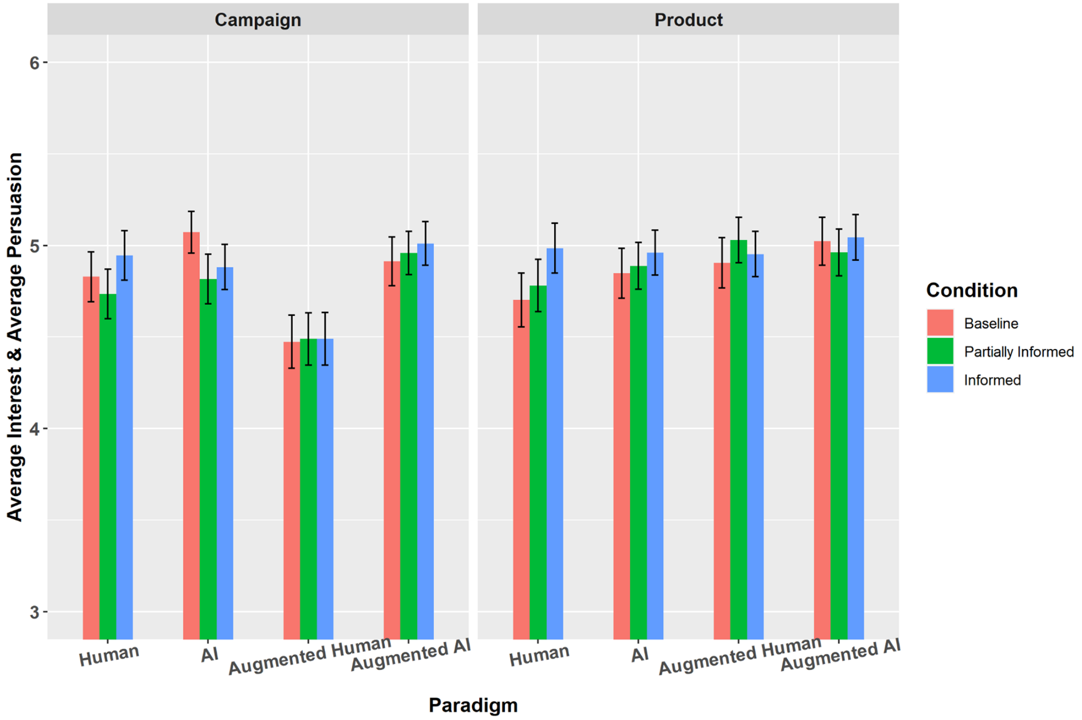

***Figure 5.*** *X-axis is the content generation paradigm. The y-axis for the left panel (pooling persuasive contents generated for five campaigns) is the (average) extent to which participants are persuaded by the persuasive content. The y-axis for the right panel (pooling advertising contents generated for five products) is the (average) extent to which participants are interested in learning more about the product after seeing the advertising content. The colors represent the different conditions. The bars indicate 95% confidence intervals. The y-axis starts from 3 instead of 0 for better visualization.*

observe that the performance gap between the 'augmented human' and 'augmented AI' paradigms is smaller when they are creating content for products. This is because human experts are more likely to adopt the texts written by AI in this case because a large fraction of the text is just standard product description.

## 5. Discussion and conclusion

Although our results suggest that ChatGPT-4 outperforms human experts in generating advertising content for products and persuasive content for campaigns, and it has the potential to reduce human labor in content generation, we by no means suggest that GAI should completely displace human workers, especially human oversight. For example, Bai et al. (2023) suggest that AI can persuade humans on political issues. In our contexts, we carefully choose the products and campaigns to be harmless. Nevertheless, human oversight is still needed to ensure that the content produced by GAI is appropriate for more sensitive topics, and inappropriate content is never distributed. Furthermore, our study does not examine non-textual content (e.g., graphical and audio), which are also popular means of communication. The performance between human experts and GAI in these domains remains to be explored by future research.

Nevertheless, our results indeed serve as evidence that GAI can benefit capital owners and consumers by raising productivity (e.g., it takes ChatGPT-4 a matter of seconds to produce the content

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

of on par or higher quality than the human experts in our context) and lowering prices (e.g., the monthly subscription fee for ChatGPT-4 is $20) (Acemoglu & Restrepo, 2018; 2020; Agrawal et al., 2019).

Our result also contributes to the discussion on algorithm aversion and appreciation (Dietvorst et al., 2015; Logg et al., 2019) vs human favoritism (Morewedge, 2022) in the domain of GAI. Instead of aversion, we demonstrate human favoritism as a form of bias—simply knowing that a piece of content being generated by human experts increases the reported perceived quality of the content. However, we do not find strong evidence of algorithm aversion in our context (i.e., knowing that a piece of content being generated with AI's involvement does not lower the level of satisfaction and willingness to pay for the content). This result is somewhat surprising given that Castelo et al. (2019) clearly show that people display aversion toward AI in subjective task contexts (e.g., evaluating joke funniness).[16] This suggests that conclusions made by existing research on conventional predictive AI may not necessarily apply in the context of generative AI. To our knowledge, our research is the first to document people's perception of persuasive content generated by industry professionals, LLMs, and their collaboration, as well as people's bias (favoritism) toward content generated solely by human experts. Future research could further investigate people's perception of the performance of LLMs (e.g., the cognitive mechanisms underlying the observed favoritism toward humans) and refine the human-in-the-loop protocol.

**Acknowledgments.** The authors appreciate Jon Baron, the two anonymous reviewers, and seminar participants from MIT B-Lab for their constructive feedback. The authors are grateful to Accenture Applied Intelligence and the MIT Initiative on the Digital Economy for financial support. We thank Patrick Connolly, Shraavya Koppole, and Philippe Roussiere for facilitating our communication and collaboration with Accenture and helping us recruit and collect responses from the professional content creators from the firm that were used as part of our stimulus.

**Funding statement.** The authors do not have financial support to report.

**Competing interest.** The authors do not have competing interests to disclose.

**Supplementary materials.** The supplementary material for this article can be found at http://doi.org/10.1017/jdm.2023.37.

## Reference

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, *108*(6),1488–1542.

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, *128*(6), 2188–2244.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, *33*(2), 31–50.

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*, 589–596.

Bai, H., Voelkel, J., Eichstaedt, J., & Willer, R. (2023). Artificial intelligence can persuade humans on political issues.

Bar-Hillel, M., Maharshak, A., Moshinsky, A., & Nofech, R. (2012). A rose by any other name: A social-cognitive perspective on poets and poetry. *Judgment and Decision Making*, *7*(2), 149–164.

Berg, J., Raj, M., & Seamans, R. (2023). *Capturing value from artificial intelligence*. Academy of Management Discoveries, In press.

Botha, J., & Pieterse, H. (2020, March). Fake news and deepfakes: A dangerous threat for 21st century information security. In *ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and Publishing Limited* (p. 57). https://researchspace.csir.co.za/dspace/handle/10204/11946.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work (No. w31161). *National Bureau of Economic Research*.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Clayton, J. (2023). Sam Altman: CEO of OpenAI calls for US to regulate artificial intelligence. BBC News. 16 May 2023. https://www.bbc.com/news/world-us-canada-65616866.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

[16] This finding is also replicated by Zhang and Gosline (2023).

https://doi.org/10.1017/jdm.2023.37 Published online by Cambridge University Press

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280.

Haupt, C. E., & Marks, M. (2023). AI-generated medical advice-GPT and beyond. *JAMA*, *329*(16), 1349–1350. https://doi.org/10.1001/jama.2023.5321

Khan, G. (2023). “Will AI-generated images create a new crisis for fact-checkers? Experts are not so sure.” Reuters Institute, 11 April 2023. https://reutersinstitute.politics.ox.ac.uk/news/will-ai-generated-images-create-new-crisis-fact-checkers-experts-are-not-so-sure.

Li, R., Kumar, A., & Chen, J. H. (2023). How Chatbots and large language model artificial intelligence systems will reshape modern medicine: Fountain of creativity or Pandora’s box? *JAMA Internal Medicine*, *183*, 596–597.

Liu, Y., Mittal, A., Yang, D., & Bruckman, A. (2022, April). Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1–13). New York, NY: Association for Computing Machinery.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

McKendrick, J., & Thurai, A. (2022). AI isn’t ready to make unsupervised decisions. *Harvard Business Review*, *15*. https://hbr.org/2022/03/overcoming-the-c-suites-distrust-of-ai.

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, *26*, 824–826.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*, 187–192. https://doi.org/10.1126/science.adh2586

Zhang, Y., & Gosline, R. (2023). Understanding Algorithm Aversion: When Do People Abandon AI After Seeing It Err? Available at SSRN 4299576.

**Cite this article:** Zhang, Y. and Gosline, R. (2023). Human favoritism, not AI aversion: People’s perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making*, e41. https://doi.org/10.1017/jdm.2023.37