Something Borrowed: Exploring the Influence of AI-Generated Explanation Text on the Composition of Human Explanations

Sharon Ferguson*
sharon.ferguson@mail.utoronto.ca
Mechanical and Industrial Engineering
University of Toronto
Toronto, Ontario, Canada

Paula Akemi Aoyagui†
paula.aoyagui@mail.utoronto.ca
Faculty of Information
University of Toronto
Toronto, Ontario, Canada

Anastasia Kuzminykh†
anastasia.kuzminykh@utoronto.ca
Faculty of Information
University of Toronto
Toronto, Ontario, Canada

ABSTRACT

Recent advances in Human-AI interaction have highlighted the possibility of employing AI in collaborative decision-making contexts, particularly in cases where the decision is subjective, without one ground truth. In these contexts, researchers argue that AI could be used not just to provide a final decision recommendation, but to surface new perspectives, rationales, and insights. In this late-breaking work, we describe the initial findings from an empirical study investigating how complementary AI input influences humans' rationale in ambiguous decision-making. We use subtle sexism as an example of this context, and GPT-3 to create explanation-like text. We find that participants change the language, level of detail, and even the argumentative stance of their explanations after seeing the AI explanation text. They often borrow language directly from this complementary text. We discuss the implications for collaborative decision-making and the next steps in this research agenda.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Explainable AI, GPT-3, Human Explanations, Human-AI Collaboration, User Study

ACM Reference Format:
Sharon Ferguson, Paula Akemi Aoyagui, and Anastasia Kuzminykh. 2023. Something Borrowed: Exploring the Influence of AI-Generated Explanation Text on the Composition of Human Explanations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585727

1 INTRODUCTION

Recently, users without any machine learning experience have been excited to explore AI's capability to produce text, notably with the scaling popularity of ChatGPT[1]. And while concerns around AI-assisted plagiarism are valid [18], one can also argue that tools such as ChatGPT can be leveraged as a starting point for inspiration to "improve, not do, the work" [10]. Similarly, there is a potential for AI systems to act as a soundboard companion in decision-making tasks where the human counterpart would be inspired by the AI explanation when building their own rationale. For instance, Lai et al. [20] argue that a Human-AI team would be particularly efficient in complex decision-making tasks such as recidivism prediction, by countering human biases with sets of predictions coupled with explanations of the rationale behind them, thus exposing patterns humans might have missed.

As the Human-AI interaction field grows, new possibilities for collaborative decision-making tasks emerge as well. While most research so far has explored *objective* decision tasks, such as predicting a person's credit score, some decisions are inherently *subjective* [20] and not based on ground truth. Thus suggesting that in ambiguous and open-to-interpretation scenarios [26], the accuracy in label classification is less important than the argumentation or rationale behind the decision. For example, in content moderation for social justice, what is considered offensive language by some, might not get the same reaction from others, resulting in low annotator agreement [9, 29]. In these ambiguous scenarios, AI can support a human decision-maker, not by offering the single most accurate prediction (that would not be possible to achieve since there is no ground truth), but by surfacing new perspectives, rationales and insights. In theory, this distinct goal or desiderata for interpretable systems has been discussed and named as 'informativeness' or AI for "discovery" [1, 21]. However, there is a notable paucity of empirical investigation into how humans co-construct a rationale using AI outputs. In summary, while it has been hypothesized that AI can assist humans in making decisions in ambiguous scenarios, there has been little empirical work to suggest this. Thus, as a first step in this investigation, we aim to identify the ways in which AI-generated text influences human rationale generation.

Some researchers argue that ultimately, the goal of Human-AI collaboration is to achieve complementary team performance (CTP) [5, 13, 17], where the combination of human and AI efforts results in a better outcome than either party could have reached on its own. For objective decision tasks, where there is ground truth, accuracy is often the metric of choice to measure the efficacy of the partnership [5]. An example is age prediction, a task with ground truth labels [11]. However, in open-to-interpretation scenarios, without ground truth labels, more subjective evaluation scales are required beyond accuracy [20], such as perceived trust [28], usefulness [22] and user satisfaction [12] to name a few - thus indicating Human-AI hybrid teams' success might also rely on more parameters that are

[1]https://openai.com/blog/chatgpt

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585727

still being uncovered. Furthermore, Zhang et al. [31] argue that a model should be built for "high performance, safety and fairness" while also guaranteeing the human counterpart can understand the machine's outputs. Explainable AI (XAI) aims to make AI outputs understandable to end users [1, 3, 14, 23, 27, 30] and to achieve this goal, different approaches have been explored. In terms of system design, a model can be directly interpretable (i.e., transparent, self-explainable and thus understandable by design) or require a post-hoc explanation (i.e., the model produces an output first and, afterwards, an explanation is generated to sustain why the output is valid [21]). In XAI-assisted decision-making, the explanation can take many forms, including natural language explanations that were found to be most effective for diverse audiences [8]. Hence, we hypothesize that the Large Language Model, GPT-3, could be used as a final step in XAI systems to produce explanation text.

In this paper, we present early findings on how human explanations are influenced by presenting complementary rationales in the form of AI-generated text. As an example of an ambiguous, open-to-interpretation scenario, we chose subtle sexism, where one situation can elicit divergent and subjective opinions, often based on personal values [24]. We present a qualitative, comparative analysis of how users explained whether a scenario was sexist before and after viewing complementary AI-generated text about the scenario. We describe how the participants took the AI-generated explanation text into consideration by changing the language used, level of detail and even argumentative stance between their first and final explanation. Specifically, participants adopted terms and phrases, but also adjusted the level of detail in their explanation, to match that in the AI-generated text.

On a short note, we acknowledge this work contains quotes with language that may be considered offensive. These are used to illustrate our findings and do not represent the authors' views.

2 METHODS

To develop an initial dataset, we collected naturally occurring scenarios of subtle sexism from internet discussions [1]. We excluded scenarios that contained explicit language. This complete scenario dataset contained 117 scenarios. To collect the AI explanation-format text for this dataset, we used GPT-3, [7], the state-of-the-art publicly available natural language model at the time. Furthermore, to conduct a user study, the model selected had to be able to provide English-text responses that were comprehensible to the end user without adding another layer of interpretation. Since GPT-3 is pre-trained, we did not train the model but only prompted it. We used GPT-3's question-answer feature and the following prompt, complete with the 10 chosen scenarios, to create the AI explanation-text dataset: *Is this sexist:* "[insert scenario here]"? *Why or why not?* We used the default temperature parameters and a max_token length of 240, and we prompted GPT-3 three times for each scenario, to ensure that at least one coherent explanation-like text was generated for each sceanrio. As a result, we had a dataset of AI-generated texts formatted as explanations. These outputs are not explanations to describe the inner workings of a model, as our work does not intend to explain how GPT-3 works, but instead explore how large language models may be used at the final stage of XAI systems to produce human-understandable explanations of AI outputs. The final AI dataset contained 351 data points. We chose 10 scenarios and corresponding explanations from this dataset such that the AI explanation text contained a semantically relevant argument, both scenarios and explanations were not longer than five sentences, and both argumentative directions (sexist vs. not sexist) were included.

To assess how AI-generated explanation text influences the composition of human explanations, we conducted a survey study with 21 participants (12 W, five M, one who identified outside of the gender binary, and three who did not specify). Regarding visible minority status, nine identified as not a visible minority, with nine identifying as a visible minority (South Asian, Chinese, Filipino, West Asian), and three leaving the category blank. We recruited four participants in the age group 18-24, nine from 25-34, six from 35-44, and two who did not provide an age. Lastly, the majority of our sample had Master's (10) or Bachelor's degrees (seven), with two not completing university and two who did not provide a level of education.

Study Procedure: The survey was conducted on SurveyMonkey [2] and was approved by the University's institutional review board. The survey contained a consent page, five randomized data collection pages, and demographic questions. Participants were randomly assigned to one of two versions of the survey with different scenarios and explanations.

Based on evidence [15, 17] that better decision-making performance can be achieved if the human counterpart is asked to make an initial decision before seeing an AI recommendation, we first asked participants to read a subtle sexism scenario and decide whether it was sexist, not sexist, or if they couldn't decide. They were then asked to explain their decision (at least three sentences), before being shown the scenario again, accompanied by the complementary AI-generated explanation text - a combination of the AI's assessment (should a behaviour be considered sexist: yes or no) with an explanation arguing why that assessment should be considered valid (why it is/is not sexist). Participants were told the text was AI-generated. Lastly, they were asked to provide a final assessment and explanation, taking into account all of the information provided, which could be as similar or as different to their original explanation as they felt necessary. As not every participant provided a full explanation pair for each scenario, Table 1 shows how many explanation pairs are used in the analysis, along with each scenario and explanation used.

To assess the influence of AI-generated text on the composition of humans' rationales, we qualitatively compared each initial explanation to the complementary AI-generated explanation text and the final explanation. We identified that an explanation had changed any time the first and final explanations were not identical. For explanations that had changed, we employed a qualitative analysis process [6], where we open coded the differences between the first and final explanation. We identified keywords and structural elements that differed between the explanations, and grouped these into higher-level categories, representing broad ways in which the explanation changed. For example, if an explanation pair was tagged with codes "same argument" and "added term ageism", this

[1] Reddit (www.reddit.com), The Everyday Sexism Project (www.everydaysexism.com) and Twitter (www.twitter.com)

[2] https://www.surveymonkey.com/

Table 1: Each Scenario and AI explanation text used in the user study. S = scenario, Ex = explanation pairs

S #	Scenario Text	AI Explanation Text	# Ex.
1	The massive amount of people who think it's a woman's duty to have children, even if she doesn't want them...	Not necessarily sexist. It's a joke about the difference in the expectation for men and women on the dating market. Also, the comedian is saying how when women hit 40, they aren't on the market anymore, but for men, it isn't the same. It's not sexist because it's about the idea that women are on the market for marriage instead of just being on the market for dating. It's still a joke, but it's not sexist.	10
2	My partner and I took his daughter shopping for new shoes, we took a number and waited for our turn. When the assistant came over I was standing and my partner and his daughter were sat on the seats waiting to be served, the assistant turned to me and asked how she could help, completely ignoring my partner.	It is not sexist. There is nothing wrong with being friendly to one person in a couple and ignoring the other. I have often been ignored by shop assistants when I have been with my husband, and it does not mean the person who is ignored is not worth anything.	9
3	I work in a large marketing department and recently whilst speaking about the voice we use on our adverts, we discussed the prospect of using a female voice over. My male senior, department deputy really, said to me that we don't use females because "all female voices sound sexual".	This is the kind of statement that both men and women make without thinking about it. While on some level it does not seem to be overtly sexist it does assume that all female voices are sexual voices. Regardless of the fact that the voice may sound great or be exactly what the company needs to get the right message across it is more about the idea that it makes women sound like sex objects. This is the kind of thinking that leads to common stereotypes about women and it should be something that we should be working to eliminate.	10
4	I return to work from maternity leave (working 30 hours a week), to be told by my boss, "hey you can't do special projects because you only work part time", " you have to be there 7 til 7" and "no-one is going to take you seriously now because you work part time" and "don't apply for a promotion cause you have more important things to worry about at home"	I think this is incredibly sexist. It is unfair and doesn't give you any room to prove yourself. If you can do the job, you have to be given the opportunity to prove it. The fact that she is only working 30hrs a week is questionable, maybe she has a family to take care of and is working full time! Your employer should respect you and your decision on how you want to spend your time and family comes before work. I think you should state that you are a hard worker, that you are capable and will take the opportunity given to you. If you don't work there anymore, that's okay, maybe they weren't the right fit for you!	11
5	men insisting that men have better friendships and women's friendships are always tumultuous or strained because they're innately two-faced and petty.	No, it's not sexist. It's true. And the reason why it's true is because the overall social contract between men and women is different. Men are typically more focussed on stuff outside of the friendship, like status or political stuff, while women are much more focussed on how someone is going to make them feel. And that creates a dynamic where the friendship is full of petty back-and-forth and mind games.	10
6	The massive amount of people who think it's a woman's duty to have children, even if she doesn't want them...	I don't think it's sexist. It could have been said about men who wanted children. I think it's a comment on how some people have children they don't want. I don't think the language in the post is sexist.	9
7	When I'm the only woman with a group of men and one of them swears and apologises only to me.	Yes. The issue here is not the swear word, but rather the subtle (or maybe not so subtle) implication that the man is swearing in front of a woman because he feels he must. He is apologising to the woman for swearing, not because he feels badly about the swear word, but rather because he feels bad that a woman heard it. This is a sexist assumption that women are delicate flowers who can't handle some language. And it is an assumption that this man may very well not even be aware he is making.	9
8	The statement "women should dress modestly because men can't help but think dirty thoughts"	It is sexist. It is implying that men are animals that are unable to control their sexual urges and women must dress to not provoke men into thinking dirty thoughts. It is also saying that the responsibility lies entirely on women to maintain men's sexual purity.	8
9	Using "guys" to address or describe a group that includes men and women	Yes, this is sexist. The word "guys" is often used by men to include women and thus belittle women. It is also used by men to include women and thereby stereotype women as the weaker sex, as if women don't know how to speak for themselves. Women who use the word "guys," however, are sometimes using it in the correct sense. But these are highly educated, well-known, and can be counted on one hand, so it's not something to aspire to. We have a society that creates and perpetuates the use of sexist language. The more we are exposed to this language, the more we use it. We don't even think about it. We are unconsciously trained to associate women with sex and men with power. I have always been a strong advocate of gender equality. I am a mother with three daughters and I have always encouraged them to be strong and independent, encouraging them to aim high and achieve their goals, whatever they may be. I am a feminist in the correct sense of the word. I believe in gender equality.	9
10	Last week, I dyed my hair from my natural brunette to a subtle purple. I posted it on one of my social medias and I quickly got a comment from another girl, saying what I'd done was stupid and silly. I asked why. She responded that I had only dyed my hair for a guy. I said "why the hell would I do that?" and she commented back "every girl does."	I don't think it's sexist, but I think it's wrong, because there's so many different reasons why people alter their appearance. Women are increasingly faced with the pressure to look a certain way, and it's hard to avoid, especially if you are on social media. I think that it's important to remember that sometimes people do things because they're fun, because they like to experiment and try new things.	9

was categorized as a language change. Then, in order to understand the influence of AI explanation text, we looked for these keywords and structural elements in the AI-generated explanation text, suggesting that this text included elements that were borrowed in the final argumentation. Additionally, although the actual decision (sexist or not sexist) was not a major consideration in this work, we discuss instances where the decision was changed from the initial explanation to the final explanation, as this represents an extreme case of influence, where the participant not only changed their rationale, but their assessment as well.

3 RESULTS

Early analysis of the human explanation pairs identified two distinct ways in which participants' explanations changed after viewing the complementary AI-generated explanation text: language, and level of detail. We identified these changes across participants and scenarios. Language changes were identified when the argument, or reasons why they made the decision, were the same, but the specific words and phrases used to describe the argument changed. In some cases, words and phrases were adopted from the AI-generated text. Changes in the level of detail were identified when the final explanation had more or less detail than the initial explanation.

3.1 Language Changes

Language changes occurred in 31 (33%) of explanation pairs. Participants sometimes borrowed terms directly from the AI explanation text:

> **Initial Explanation:** "It is sexist because the deputy considers the female voice as sexual by default. Why doesn't the male voice sound sexual? It is unfair." [H; S3]
> **Final Explanation:** "It is sexist because the deputy considers the female voice as sexual by default *or a gender stereotype.* Why doesn't the male voice sound sexual? It is unfair." [H; S3]

Although the argument does not change, as "considering the female voice as sexual by default" is a stereotype, the participant adds the phrase "gender stereotype", where the word stereotype comes directly from the AI explanation-text: *"...This is the kind of thinking that leads to common stereotypes about women..."*.

The changes in phrasing also may be inspired by, but not directly borrowed from the AI explanation text. Consider this example where similarities suggest language changes might be influenced by the AI text:

> **Initial Exp:** "it is sexist because it shouldn't be the responsibility or burden of the woman to be responsible for the thoughts a man might have..." [H; S8]
> **Final Exp:** "yes, assigning too much responsibility on the women for the thoughts of men..." [H; S8]

Perhaps this portion of the AI-generated text *"...It is also saying that the responsibility lies entirely on women to maintain men's sexual purity"* [AI; S8], particularly the use of the word *"entirely"*, led the participant to change their phrasing from just "responsibility" to "too much responsibility."

On the other hand, in some cases the language changes do not seem to be influenced by the AI-generated text at all:

> **Initial Exp:** "It is sexist because again a certain sex-based stereotype is being promoted. This stereotype being that women's choices around their appearances has much or everything to do with garnering male approval and attention." [H; S10]
> **Final Exp:** "I think it is sexist, particularly due to the comment that "every girl does". It is reinforcing a known stereotype that women are actively seeking male attention and approval, particularly where looks are concerned." [H; S10].

These terms and arguments were not used in the AI explanation text; the AI text focused more on the pressure women feel to look a certain way based on social media. Additionally, in this example, the human and AI explanations contained contradicting argumentative directions. Thus, we see that adapting the language used in the explanation to be closer to the complementary AI text is an argumentative strategy used by humans in collaborative decision-making settings primarily when they agree with the other collaborative party.

Evidence of language from AI text adopted in human explanations suggests that the participants are reading and understanding the complementary AI explanation text, and considering this information when forming their final explanation. It also may mean that participants are creating a common ground with the system, or attempting to describe their reasoning in similar terms.

3.2 Level of Detail

Thus, in addition to changes in the language used in explanations, we also see changes in the level of detail provided in explanations. This occurred in 45 (48%) of explanation pairs. Often, the level of detail in the final explanation is closer to the level of detail contained in the AI text than the initial explanation. We see participants add definitions and examples to their explanations to match the AI-generated text, or roll their explanation up to a higher level. This happens in cases where the participant's stance is similar to the one in the AI text, and when they disagree. In most cases, the final explanations contained more detail, often in the form of an example:

> **Initial Exp:** "Depends on two factors: - The gender of the narrator (not specified here) - The assistant considers standing customers as waiting to be served or sitting customers." [H; S2]
> **Final Exp:** "Depends on two factors: - The gender of the narrator (not specified here) - The assistant considers standing customers as waiting to be served or sitting customers It's hard to tell without further context, the assistant might just not [be] seeing the sitting partner and his daughter."[H; S2]

The participant may be adding an example in response to the example in the AI explanation, *"I have often been ignored by shop assistants when I have been with my husband..."* [AI; S2].

The participants also added detail by defining a term used in their initial explanation:

> **Initial Exp:** "Jokes about generalized differences between men and women are sexist by definition." [H; S1]
> **Final Exp:** "I still think it is a sexist joke. Hard to definitively say based on one sentence. However, it seems to me the comedian is generalizing the difference between men and women above 40, their attractiveness "on the market", and what they are looking for. It is stereotyping on the basis of sex, and therefore sexist." [H; S1].

Here, the participant uses similar language and level of detail to the AI-generated text, even though they are arguing different points of view (H: sexist, AI: not sexist).

Lastly, we saw participants providing a final explanation that was more high-level than their initial explanation, particularly when the complementary AI explanation text shown to the participant is also high-level (Table 1):

> **Initial Exp:** "The statement doesn't account for a woman's autonomy. There is no consideration around whether a woman wants children or not, and it makes an assumption about someone's role based solely on gender. The 'even if she doesn't want them' part is especially concerning because, again, it raises a question of whether consent has been considered." [H; S6]
> **Final Exp:** "This scenario is still sexist. There is no evidence of consideration around bodily autonomy. There is an assumption being made based on traditional, outdated notions of what it means to be a woman, and it comes off as non-consensual." [H; S6]

This suggests that participants are considering the complementary AI text when formulating their final explanation, and may attempt to "respond" to the collaborative partner with the same amount of information.

3.3 Argumentative Direction Change

In addition to the initial analysis of the changes in the explanations, we analyzed the effect of AI explanation text on the argumentative direction. There were numerous examples of this assessment change in our dataset, across scenarios (six out of the ten scenarios showed a participant who flipped their decision after viewing the AI explanation) and participants (12 participants flipped in 1/5 scenarios, two participants in 2/5). Since participants were asked to rate the scenario as sexist, not sexist, or 'it depends,' we identified a switch when this label changed from the first explanation to the final explanation. This happened in 16, or 17% of explanation pairs. The majority of decisions were changed from the initial uncertain 'it depends':

> **Initial Exp:** "I'm a little bit torn with this scenario. On one hand, I could see it being sexist by the notion it reinforces (i.e. that a woman's role in society is to to bear and rear children). But on the other hand, I could [see] this 'role' being imposed on younger generations, men, etc etc" [H; S6].
> **Final Exp:** "It is not sexist because I agree with the AI explanation. The same could be said about men as well" [H; S6].

On the other hand, there were instances when the participant initially provided a more definitive label (i.e. sexist or not sexist) and then became more unsure in their final explanation. Most of the time this happened when the initial label was 'not sexist':

> **Initial Exp:** "It is not sexist because this person cannot make it to work full-time, given her current status. According to the phrases of the boss, doing special projects and promotions may need more time commitment..."[H; S4]
> **Final Exp:** "It depends. If the boss's judgment is based on gender, i.e., women cannot perform outstandingly due to maternity and family issues, it is sexist. However, it is not sexist if these phrases are based on time commitment..."[H; S4].

While in this case, the participant becomes less certain in their final explanation, viewing the complementary AI explanation text provided new perspectives that influenced their rationale.

4 DISCUSSION

In this work, we aimed to understand how people adapt their explanations after being exposed to collaborative input from AI for the decisions made in ambiguous scenarios. Our early results suggest that humans might be adapting the language and adjusting the level of detail in their explanations, and even changing their decision, when presented with complementary AI-generated argumentation. For instance, we saw participants include additional terminology used in AI-generated explanation text, signalling the efforts in developing a "shared language" between collaborative partners. These results provide early empirical evidence of the influence of AI collaborative input on human rationales in ambiguous scenarios. In contrast to objective decisions, where complementary team performance [4, 13] equals accuracy in prediction, subjective and personal decisions benefit from new information and perspectives thrown into the mix. This is a distinct purpose for XAI where unlike explaining predictions, AI can expose patterns and insights humans would have ignored [1, 20, 21]. We find that humans might adapt their explanations in response to the AI explanation text both when they agree and disagree with the argumentative direction present in the AI text. This suggests that our participants may not be blindly following the decision presented in the AI text, a concern when it comes to trust calibration in XAI-assisted decision-making [28]. The influence of AI-generated text on human decisions can be compared to the study of social learning in humans, where recent work provides evidence for different learning styles: some individuals make decisions based only on personal experience, while others base it only on external advice [25]. Future work can investigate individual differences in the adoption of AI-generated ideas in rationale co-creation.

This study describes human-in-command decision-making [2, 16, 31], where AI input may be provided, but humans make the final decision. This is similar to past studies of collaborative, complementary performance [4, 5] and is recommended for risky situations where an algorithm should not operate autonomously [16, 31]. While these studies are usually conducted in the context of complex, high-stakes decisions, like medical readmission and recidivism

cases [17], we show that AI-generated collaborative input influences explanations even in equally ambiguous, though less critical, settings. Additionally, participants were told that the collaborative input was AI-generated. Researchers have frequently found that humans are averse to AI decisions [12] and AI-generated text [19]; in line with this, we found instances of humans disagreeing with the AI explanation text. However, we also found participants adopting language and level of detail from AI-generated text, suggesting that human perception of AI input in ambiguous and collaborative scenarios may differ from findings in past work.

As a next step, we plan to investigate how the thematic content of the explanations changes – specifically, what new information is being provided in the AI-generated text, and if this information is used in the participant's decision-making process. Future work can expand by using complete text-based XAI systems for ambiguous scenarios as they are developed. This study was scoped to focus on only 10 examples of subtle sexism as an open-to-interpretation case with implications for hate speech detection, and thus cannot be generalized to all scenarios and contexts. Future work can test these findings using other ambiguous contexts, such as judging creative work (e.g. short stories or visual art), another subjective task. Further, our recruited participants skewed towards younger individuals with a university education, thus caution must be exercised when generalizing these findings to broader populations. Lastly, while the qualitative analysis presented here allows us to explore the influence of AI-generated text in detail, future work could use natural language processing algorithms to quantify these changes.

5 CONCLUSION

In this work, we described the initial results from a user study investigating how complementary AI-generated text influences humans' rationale in ambiguous decision-making. Using subtle sexism as an example of this context, we found that participants changed the language used and level of detail in their explanations after viewing the AI-generated explanation text. We provide evidence that AI-generated text can be useful in offering new perspectives, rationales and insights for ambiguous scenarios that aid in human decision-making. Further analysis will investigate how AI-generated text influences the themes contained in human explanations, and replicate this analysis using complete XAI systems.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] HLEG AI. 2019. *High-level expert group on artificial intelligence.* Technical Report. 6 pages.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[8] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Information Processing & Management* 60, 1 (2023), 103111.

[9] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[10] Brian X Chen. 2022. How to Use ChatGPT and Still Be a Good Person. *The New York Times* (12 2022). https://www.nytimes.com/2022/12/21/technology/personaltech/how-to-use-chatgpt-ethically.htmln

[11] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).

[12] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[13] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1639–1656. https://doi.org/10.1145/3531146.3533221

[14] D Doran, SC Schulz, and TR Besold. 2018. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. In *CEUR Workshop Proceedings*, Vol. 2071.

[15] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[16] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. 2022. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine* 17, 1 (2022), 72–85.

[17] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* (2021), 78.

[18] Alex Hern. 2022. AI-assisted plagiarism? ChatGPT bot says it has an answer for that. *The Guardian* (2022). https://www.theguardian.com/technology/2022/dec/31/ai-assisted-plagiarism-chatgpt-bot-says-it-has-an-answer-for-that

[19] Nils Köbis 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 (2021), 106553.

[20] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[21] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[22] Sean McGrath, Parth Mehta, Alexandra Zytek, Isaac Lage, and Himabindu Lakkaraju. 2020. When does uncertainty matter?: Understanding the impact of predictive uncertainty in ML assisted decision making. *arXiv preprint arXiv:2011.06167* (2020).

[23] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[24] Chelsea Mitamura, Lynnsey Erickson, and Patricia G Devine. 2017. Value-based standards guide sexism inferences for self and others. *Journal of Experimental Social Psychology* 72 (2017), 101–117.

[25] Maayan Pereg, Uri Hertz, Ido Ben-Artzi, and Nitzan Shahar. 2022. Disentangling the contribution of individual and social learning processes in human advice-taking behavior. https://doi.org/10.31234/osf.io/28js3

[26] Schaekermann, Mike. 2020. *Human-AI Interaction in the Presence of Ambiguity: From Deliberation-based Labeling to Ambiguity-aware AI.* PhD Thesis. UWSpace. http://hdl.handle.net/10012/16284

[27] Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093* (2020).

[28] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International*

Conference on Intelligent User Interfaces. 318–328.

[29] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[30] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*. Springer, 563–574.

[31] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

Received 19 January 2023