The Explanation That Hits Home: The Characteristics of Verbal Explanations That Affect Human Perception in Subjective Decision-Making

SHARON FERGUSON, Mechanical and Industrial Engineering, University of Toronto, Canada
PAULA AKEMI AOYAGUI, Faculty of Information, University of Toronto, Canada
RIMSHA RIZVI, Faculty of Information, University of Toronto, Canada
YOUNG-HO KIM, NAVER AI Lab, Republic of Korea
ANASTASIA KUZMINYKH, Faculty of Information, University of Toronto, Canada

Human-AI collaborative decision-making can achieve better outcomes than either party individually. The success of this collaboration can depend on whether the human decision-maker perceives the AI contribution as beneficial to the decision-making process. Beneficial AI explanations are often described as relevant, convincing, and trustworthy. Yet, we know little about the characteristics of explanations that result in these perceptions. Focusing on collaborative subjective decision-making, using the context of subtle sexism, where explanations can surface new interpretations, we conducted a user study (N=20) to explore the structural and content characteristics that affect perceptions of human and AI-generated verbal (text and audio) explanations. We find four groups of characteristics (*Tone, Grammatical Elements, Argumentative Sophistication* and *Relation to User*), and that the effect of these characteristics on the perception of explanations for subtle sexism depends on the perceived author. Thus, we also identify which explanation characteristics participants use to identify the author of an explanation. Demonstrating the relationship between these characteristics and explanation perceptions, we present a categorized set of characteristics that system builders can leverage to produce the appropriate perception of an explanation for various sensitive contexts. We also highlight human perception biases and associated issues resulting from these perceptions.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence.*

Additional Key Words and Phrases: Collaborative Decision-Making; Explainable AI; Verbal Explanation; Explanation Characteristics; Perceptions; Subjectivity

ACM Reference Format:
Sharon Ferguson, Paula Akemi Aoyagui, Rimsha Rizvi, Young-Ho Kim, and Anastasia Kuzminykh. 2024. The Explanation That Hits Home: The Characteristics of Verbal Explanations That Affect Human Perception in Subjective Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 517 (November 2024), 38 pages. https://doi.org/10.1145/3687056

Authors' addresses: Sharon Ferguson, sharon.ferguson@mail.utoronto.ca, Mechanical and Industrial Engineering, and University of Toronto, Toronto, Ontario, Canada; Paula Akemi Aoyagui, paula.aoyagui@mail.utoronto.ca, Faculty of Information, and University of Toronto, Toronto, Ontario, Canada; Rimsha Rizvi, rimsha.rizvi@mail.utoronto.ca, Faculty of Information, and University of Toronto, Toronto, Ontario, Canada; Young-Ho Kim, yghokim@younghokim.net, NAVER AI Lab, Seoul, Republic of Korea; Anastasia Kuzminykh, anastasia.kuzminykh@utoronto.ca, Faculty of Information, and University of Toronto, Toronto, Ontario, Canada.

2573-0142/2024/11-ART517 $15.00
https://doi.org/10.1145/3687056

1 INTRODUCTION

Human-Artificial Intelligence (AI) collaborative decision-making aims for complementary performance [9, 10, 40, 67, 121], where human and AI partners together achieve a better outcome than they would individually. The nature of this collaboration might vary based on the decision-making scenario [93], particularly whether the decision to be made is objective (i.e., based on a ground truth) or subjective (i.e., open to interpretation). In any collaboration scenario, the human user needs to understand the AI output, which is commonly addressed through AI explainability features [1, 5, 105, 108, 141], and researchers have been actively exploring the effects of these explanations on the collaboration process [99, 149, 154]. Specifically, in subjective decision-making, AI explanations can additionally play an argumentative role [50, 93, 98], supporting the human user by providing multiple rationales and helping them to consider a broader set of arguments.

To consider the suggested perspectives—to effectively incorporate them into the decision-making—the user should see these AI explanations as beneficial to the decision-making process. Prior work suggests that for an explanation to be considered, it should be perceived as *relevant* [11, 62, 89, 129, 132], appropriately *convincing* [20, 56, 70, 118, 123], and appropriately *trustworthy* [47, 63, 64, 78, 90, 91, 161]. However, although the importance of these perceptions is extensively discussed, little is known about the characteristics of an explanation contributing to these perceptions [16, 74, 126, 149].

The effective incorporation of collaborative input into decision-making is further influenced by the perception of the collaborative party, both in human-human [41, 55] and, perhaps even to a greater extent, in human-AI partnership [94, 160]. For example, recent studies show that humans tend to misinterpret the characteristics of AI-generated images due to misjudgment of AI abilities [104]. Additionally, humans are reported to be uncomfortable when artificial systems produce uniquely human-like characteristics [110], e.g., intelligent agents displaying human-like emotional responses can trigger discomfort [142] or frustration [100, 103]. We suggest that the influence of these perceptions on the effective incorporation of collaborative input could be expected to be even more critical in subjective decision-making, where culture, beliefs, and values [93, 107], as well as uniquely human opinions and experiences, play a large role in the interpretations of data [61]. Correspondingly, we hypothesize that, in the context of subjective decision-making, the perceptions of an explanation's relevance, convincingness, and trustworthiness may differ based on the perceived source of the explanation (human vs. AI).

In this work, we investigate which characteristics of a verbal AI explanation affect its perceptions in subjective decision-making. Verbal explanations are made up of words, phrases, and natural language [109], and are sometimes referred to as "natural language explanations" [22]. While XAI research has explored various modalities [60, 69, 79, 112], we choose to focus on verbal explanations, which were shown to be more effective than graphic explanations [127], are often regarded as more equitable [146], require less expertise for comprehension [143] and become particularly promising with the growing popularity of large language models (LLMs) [13, 24, 35, 97, 163]. Given that modality can affect the perception of AI explanations [127], we consider both text and audio formats of verbal explanations, especially since audio modality has received little attention in XAI research so far [3, 79, 127].

Given the complexity and nuanced nature of human-AI collaboration processes, research exploring these processes is commonly performed on an example of a specific domain, e.g., work by Lee et al. [96] on an understanding of complementary strengths in human-AI collaboration, performed through an example of a healthcare setting. Human-AI collaboration in subjective decision-making has been previously studied in a number of domains, including medical [132], judiciary [27], and hiring decisions [113], yet has not been explored in the domain of hate speech detection, where decisions on what is or is not harmful also heavily rely on personal beliefs and values [107]. This use case represents an existing pressing issue in the domain of automated hate speech detection [23, 36, 61, 102, 128, 152] where human-AI collaboration could thrive. Although machine learning models can successfully identify overt hate speech from the internet by relying on swear or stereotypical words [36, 152], subtle sexism cases do not follow this pattern [117] and are significantly harder to identify. In this context, moderation relies on subjective decisions and requires the human user's final assessment. Given that one's interpretation of sexism can be very culturally dependent and influenced by one's beliefs and experiences [107], additional perspectives become particularly useful for such decisions. Previous work has shown that these additional perspectives can indeed be surfaced through AI explanations of its rationale for identifying context as sexist [50]. However, to be effectively incorporated into the decision-making, they should be perceived by a human collaborator as relevant, appropriately convincing, and appropriately trustworthy.

Correspondingly, to inform the design of AI verbal explanations as a collaborative input for human-AI subjective decision-making, we first explored (RQ1) which *characteristics of a verbal explanation* affect users' perception of it as relevant, convincing, and trustworthy, and whether these characteristics differ based on the perceived author (human vs. AI). In an interview-based study with 20 participants, we provided eight scenario-explanation pairs from a database of subtle sexism scenarios where an explanation was either produced by a human or generated by a GPT model [15]. Participants were asked to predict and explain their choice of the explanation author (human or AI), evaluate the explanations' relevance, convincingness, and trustworthiness, and provide a rationale for these evaluations. Our findings revealed that the perceived relevance, convincingness, and trustworthiness of verbal explanations for subtle sexism detection are affected by a set of explanation characteristics: *Tone, Argumentative Sophistication* and *Relation to User* (Figure 1, Table 7). Notably, the *Relation to User*, i.e., the alignment with the user's personal experiences and opinions, played a strongly positive role in all three evaluation categories, suggesting a concerning confirmation bias that requires particular attention from explanation designers. Further, there were notable differences in how characteristics were interpreted based on the perceived author of the explanation, such as a neutral tone being convincing and trustworthy if coming from humans, but distinctly unconvincing and untrustworthy if believed to come from AI.

Since we hypothesize that the perceived author of an explanation would impact how explanation characteristics contribute to the overall perception of an explanation, we additionally examined (RQ2) which *characteristics of a verbal explanation* for subtle sexism do people consider when predicting its author (human vs. AI). Although participants did only slighter better than chance when predicting the author of an explanation, they systematically relied on explanation characteristics that they strongly associated with either human or AI authorship (see Figure 2). These characteristics fell into the above three categories, with the addition of *Grammatical Elements*. Broadly, people used binary strategies in identifying the author, such as an emotional tone representing human authorship, while the opposite, a logical tone, represented AI output. Only a few characteristics (short length, repetition, completeness of argument, and directness) were commonly ascribed to both human and AI authorship. Additionally, while participants had equally low overall accuracy in predicting authorship across modalities, they were more accurate at identifying human-authored explanations in text and AI-authored in audio.

The results of this work contribute a categorized set of characteristics that influence the users' perceptions of authorship, relevance, convincingness, and trustworthiness of verbal explanations in subjective decision-making in sensitive contexts. We describe how each category affects the corresponding perceptions and discuss these findings in the context of existing literature. Finally, informed by our findings, we propose design considerations for AI explanations in collaborative subjective decision-making.

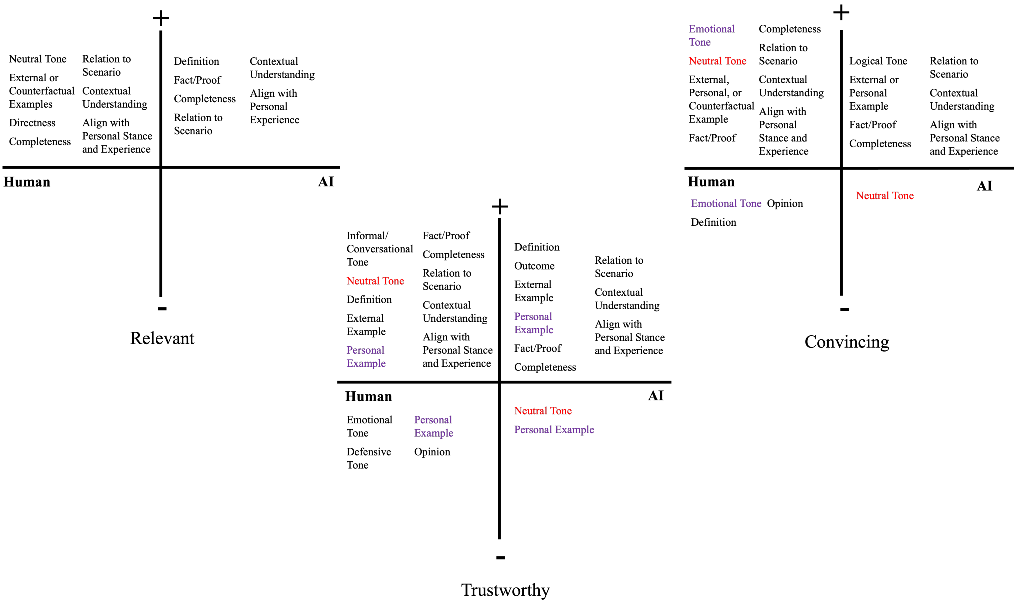


Fig. 1. Visualization of the explanation characteristics that impact different perceptions. Red text corresponds to characteristics that are perceived distinctly differently based on the perceived author, and purple text corresponds to characteristics that are perceived both positively and negatively.

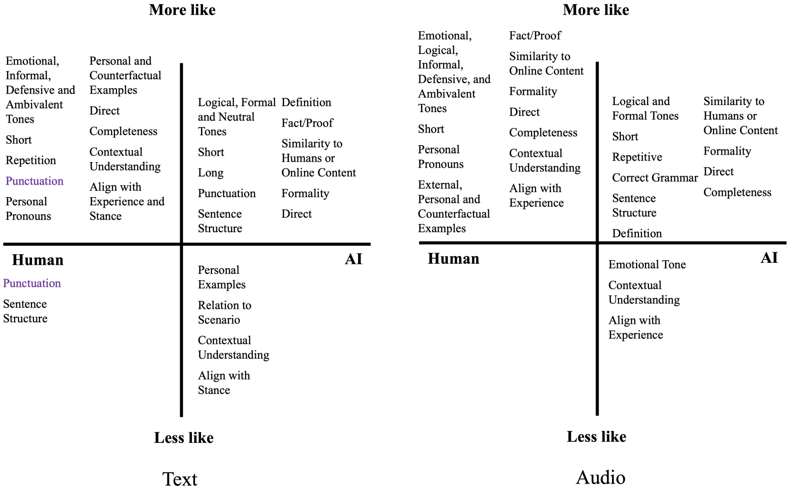


Fig. 2. Visualization of the explanation characteristics that impact perception of authorship. Purple text corresponds to characteristics that are perceived both to make an explanation more like (human or AI) and less like (humans or AI).

2 BACKGROUND WORK

In this section, we review relevant work on human-AI collaboration, AI explanation generation, and human perceptions.

2.1 Human-AI Collaboration

The ideal Human-AI partnership aims to achieve complementary performance, where both parties together are more accurate or higher performing than either could be on their own [10, 40, 67]. For example, in identifying lesions in the body, Reverberi et al. [126] found that medical professionals rely on the AI more frequently when it is correct, resulting in higher performance than if the AI or human worked on their own. In evaluating poetry, Hitsuwari et al. [72] found that AI-generated poems with human intervention were rated as the most beautiful. However, in a study in the education domain, Ren et al. [124] found that human-AI collaboration produces faster, but less accurate, tagging. Researchers have also focused on describing the mechanisms of successful collaboration. Hemmer et al. [68] found that humans perform better on a task when an AI model breaks down that task for them, and Lai et al. [92] found that having humans identify guidelines for when to rely on AI, and when not to, was successful. Ma et al. [101] have used AI not just to provide decisions but also to promote human reflection and discussion during AI-assisted decision-making to prevent over-reliance.

There are a number of examples of human-AI collaboration towards decision-making [126], gameplay [6], qualitative analysis [52], User Experience research [46], content moderation [92], and peer support [139]. Further, Human-AI collaboration has been explored in a number of different fields, such as medicine [20, 126, 132], software engineering [19], security, education, and accessibility [147].

One important area for human-AI collaboration is hate speech detection; both because the volume of hate speech on the internet is beyond that which a human could handle [2], and because fully automated hate speech detection methods can miss out on important nuances in language and context [59]. For instance, Oliva et al. [117], Kim et al. [84] and Dias Oliva et al. [39] have demonstrated how automated content moderation systems can be biased against marginalized communities by flagging their content as toxic more often. Additionally, hate speech judgements can be inherently subjective [59], evidenced by the high disagreement rates when humans annotate datasets for toxic content [4, 43, 151, 153] to be used in classification systems. Therefore, in these decision-making scenarios where there is no ground-truth, the AI system could suggest multiple arguments from different perspectives for the human counterpart to make a final decision [93]. In fact, LLMs are being increasingly used to as part of content moderation systems [75]. Hence, the opportunity here is to combine human and AI efforts to achieve complementary performance [10, 40, 67, 121].

2.2 Explainable AI

To support this successful Human-AI collaboration, the field of Explainable AI (XAI) has been gaining attention and relevance both in academia and in industry settings. The core problem XAI tackles is how to make AI systems and outputs understandable to end-users who need to interact with them, supporting the development of an accurate mental model of AI capabilities [1, 33, 37, 105, 148].

In fact, it might come as a surprise, but Eiband et al. [44] found not only that AI recommendations accompanied by explanations generate more trust in human decision-makers, but that even placebo explanations have the same effect. Similarly, in medical settings, Panigutti et al. [119] also found humans preferred recommendations from an automated system to be accompanied by an explanation to justify it, preferably in natural language.

As explanations are often preferred to be in natural language, researchers have been exploring using LLMs to generate these explanations. For example, Balog et al. [7] demonstrated LLM-generated explanations for recommender systems have an influence on human decision-makers. Feng et al. [48] explored how LLM-based agents can effectively collaborate with humans on complex tasks, generating both their next action and a rationale for it. Further, Danry et al. [34] experimented with GPT-3 and suggests that AI-generated explanations and AI-framed questions could instigate human critical thinking. Interestingly, Wiegreffe et al. [154] compared crowd-sourced and GPT-3 authored explanations to justify classification decisions and found that humans preferred the LLM-generated ones. Past work by Ferguson et al. [50] has also shown how an LLM's output can influence how a human will write their final decision in text format.

However, when discussing LLMs, it is not possible not ignore evidence that these systems can mirror and reproduce existing social biases and stereotypes found in society [18, 25, 38], including gender biases [87]. Thus, this area of research requires more exploration and attention to understand how to apply LLMs for human-AI collaboration.

2.2.1 *Modality.*

Another choice required when designing AI decisions is the modality of the explanation, with text (natural language) and graphic formats being the most commonly explored in XAI research currently, but little is known about the use of audio [79, 127]. For example, research by Van Berkel et al. [146] indicates that text-based explanations are often regarded as more equitable than visual explanations, as they lead to better user performance. Conversely, Szymanski et al. [143] suggest that while lay users favor concise visual explanations due to their potential to counter cognitive bias, these explanations require a certain level of expertise for complete comprehension. It is also possible to borrow from research related to Conversational Agents, considering it might be expected that audio will create a personal connection with the user and can be more engaging when compared to text [86]. On the other hand, in some instances, audio modality might require a higher cognitive effort in the interaction [130]. Ultimately, Schuller et al. [134]'s work on *sonified XAI* highlights the potential of AI-generated audio explanations as a "novel problem" to be explored in the field with many questions still to be answered.

2.3 Explanation Perception and Evaluation

With the wide variety of different forms of explanations, researchers have started exploring ways to evaluate AI explanations [16, 74, 126], from automated measures based on required properties [69, 129], to human-centric evaluations of how people perform on decision-making tasks using these explanations [16, 80, 127, 132, 149].

Additionally, explanations are evaluated through subjective measures, often using scales [16, 20, 70, 74, 89, 111, 132], such as humanness [31], and credibility [127], often comparing these scale measures for different types of explanations. For example, Wiegreffe et al. [154] found that subjective acceptability of AI explanations is correlated with explanation characteristics such as grammar and factuality. However, perhaps the most common metrics are relevance [69, 89, 129, 132, 149], convincingness [20, 70, 118], and trustworthiness [51, 63, 76, 90, 91, 127, 161, 162].

2.3.1 *Relevance.*

Hendricks et al. [69] found that explanations rated as relevant are more useful to humans in objective visual recognition tasks, and Schaekermann et al. [132] showed that relevant explanations improve the accuracy of human-AI collaborative decision-making, particularly in subjective cases, such as ours.

In past work, relevance was measured by whether random arguments were added [132], whether the explanation reflects the content of an image [69], or whether an explanation was helpful to the end user [89]. Even when relevance is not explicitly measured, it can be implied as an important characteristic, as many XAI systems are designed to filter out irrelevant content from the explanation [132].

Despite this importance, little is known about the characteristics of explanations that make them relevant. Russell et al. [129] found that counterfactual explanations are described as more relevant and preferred to non-counterfactual ones, and some studies conclude that the relevance can depend on the audience [62] and the context of use [9].

2.3.2 Convincingness. Research states that explanations need to be convincing in order to enable groups to come to a decision in collaborative settings, mimicking human deliberation [131]. Explanations that are not convincing would not be considered by the decision-maker at all.

Correspondingly, whether an explanation is convincing is often measured based on whether a user is convinced to take an action, e.g., whether the explanation prompts a user to review a medical case [132], to accept the AI classification [69], or to buy a recommended product [70].

While convincingness itself is not often a part of subjective evaluations, in the past, participants were asked, for example, to rate whether an explanation helped them [16] or was useful [74]. Importantly, overly convincing explanations pose a risk, as AI-generated content is not always correct and should not be blindly followed [149, 156, 161].

Past work provides little understanding of which characteristics make an explanation convincing, showing only that convincing explanations are the ones that explain why a classification is A and not B [20] and that convincing explanations are explainable themselves [123].

2.3.3 Trustworthiness. Trust greatly influences system adoption in collaborative decision-making (trust-dependant behavior [95]), and the role of perceived trustworthiness in AI explanations has been actively explored [8, 51, 63, 73, 162]. Bansal et. al [10] demonstrated that showing explanations next to AI recommendations increased humans' trust in the system even when the AI output was wrong, thus, inherently breaking the goal of complementary performance with over-reliance. Correspondingly, Zhang et al [161] suggest that it is necessary to "calibrate trust in AI" [156] to help end-users decide when they should trust the model, which differs from "enhancing trust in AI". Additionally, Wang and Yin [150] has found that changing AI explanations influence subjective trust in the explanation.

While multiple factors are known to influence human trust in automation [73], little is known about which explanation characteristics impact perceived trustworthiness.

2.3.4 Authorship. Additionally, there is evidence that perceived authorship of explanations might impact perceptions of explanations. For example, Kunkel et al. [91] described how personal movie recommendations written by humans tend to be perceived as more complex and trustworthy than outputs written by a system. Additionally, even subtle visual cues to convey the author's expertise on the domain were shown to enhance users' trust [90]. Jakesch et al. [76] showed that when faced with a mix of human and AI-generated recommendations, humans will mistrust the AI-authored ones more, whether they know or simply suspect that a system wrote it. Further, Ashktorab et al. [6] found that in playing a game, when participants thought they were playing with a human rather than an AI, they rated their companions as more likeable, intelligent, creative and had more rapport, although there was no difference in performance.

An important caveat in these studies of perception is that humans are notoriously bad at identifying content written by AI. Studies have applied customized variations of the Turing Test to evaluate human capabilities to detect AI-generated content in various formats such as poetry [85] and self-identification texts [77], consistently showing that humans are bad at detecting the difference — the accuracy rate being comparable to a chance [31, 76]. This could be caused by LLM's ability to generate human-like characteristics. For example, Sharma et al. [140] show that LLMs can be trained to produce human-like attributes such as agency, consisting of intention, motivation, self-efficacy, and self-regulation. However flawed human's perceptions of AI are, these perceptions can be expected to affect how human decision-makers receive, and later act on, an AI-generated recommendation.

2.3.5 Comparison. Lastly, research has also shown how the similarity between AI explanations and humans' beliefs, values, or cognitive processes can influence how the explanation is perceived [26, 49, 106]. For example, Chen et al. [26] described how complementary human-AI performance depends on how well the AI explanation fits into the human's intuition, and Miller [106] argues that current explanation design does not take into account the cognitive processes humans follow when making decisions. Miller [106] instead recommends an AI which provides evidence, instead of recommendations. Further, Fok and Weld [54] describe how many human-AI collaborative systems fail to reach complementary performance because the explanations do not provide information to the human in a way which allows them to verify the AI recommendation.

To summarize, whether human-AI collaboration results in complementary performance often depends on the human decision-maker's perception and evaluation of the AI explanation. While previous work recognized the importance of subjective perceptions of AI explanations and explored different ways to measure these perceptions as outlined above, there is still little understanding of which specific characteristics of an explanation make it acceptable to humans in a collaborative decision-making setting. Our work aims to contribute to the design of AI explanations by investigating the interplay of explanation characteristics, explanation modality, and perceptions of relevancy, convincingness and trustworthiness.

3 METHOD

To investigate which characteristics of verbal explanations affect their perceived authorship, relevance, convincingness, and trustworthiness, we designed an interview-based study where participants were presented with scenario-explanation pairs. We chose interviews as a research method to probe participants on their rationale, resulting in more in-depth responses relating to the specific features of the explanation that lead to a specific perception. For example, we regularly had to ask participants to further explain what they meant by "structure", or to specify which aspects of the sentences influenced their decisions. This is supported by past work, which shows that surveys or unmoderated studies rarely contain participants' rationale for making a decision [71]. The study design was approved by the institution's research ethics board (Ethics ID number 41256). This section describes the dataset creation, study procedure, and participants, followed by the data analysis process.

3.1 Explanation Dataset

We first created a dataset of subtle sexism scenarios, each accompanied by a human- and AI-generated explanation text. To collect naturally occurring subtle sexism discussions, three researchers searched online platforms using search terms *"subtle sexism", "everyday sexism", "Why is this sexist?", "Is this sexist?"*, etc. We then followed a tree search through related posts on the identified pages. As we required both the scenario and the human explanation, we searched for scenarios that were coupled with an interpretation or explanation of why the scenario was or was not sexist. We excluded scenarios that contained explicit language. The scenario descriptions were decoupled from the explanation or interpretation to create two separate datasets. The scenario dataset contained 117 examples: 83 from Reddit[1], 40 from The Everyday Sexism Project[2], and five from Twitter (X)[3]. The final human explanation dataset contained 128 explanations (some scenarios having more than one interpretation).

We then used GPT-3 (the state-of-the-art available at the time of data collection) to generate explanation-format text for each scenario, similar to Wiegreffe et al. [154], simulating the last stage in a complete XAI system. We used the question-answer feature with no fine-tuning (due to GPT-3's capability in zero-shot learning [15]). We used a basic question-and-answer prompting technique to produce "baseline" LLM-generated explanations, without prompting the model to produce any specific explanation characteristics. We used Python scripts with OpenAI's Completion API to obtain the explanations, keeping all default parameters, with the exception of the max_token length, which we increased to 240 words to prevent cut-off explanations. Due to the nature of the probabilistic generation of LLMs, GPT-3 did not always generate the same quality of responses. Hence, we generated explanations from each scenario three times. We used the following input:

Input: `Q: Is this sexist: “{{scenario}}” Why or why not?`
Output (example): `No, it’s not sexist. It’s true. Because the overall social contract between men and women is different.`

The final AI explanation dataset contained 351 data points (117 scenarios × 3 generations). We then chose eight scenario-explanation pairs (four with human-generated explanations and four with AI-generated explanation text) based on the following criteria: 1) the scenario had a corresponding coherent explanation text. 2) We attempted to balance the argumentative stance (i.e. sexist vs. not sexist)—however, due to the nature of our data collection, we still see an overrepresentation of "it is sexist" argumentation, as humans are unlikely to take the time to post a scenario on a discussion site if they found it non-sexist. 3) The scenario-explanation pair is no longer than five sentences – for participants' convenience. We chose to use eight scenarios in order to balance the length of the interview while also covering enough different scenarios to pull on various participant beliefs and values. Full scenarios and explanations are excluded from this paper to protect the original poster's privacy, though examples are shown in Table 1[4]. To explore the effects of modality, we created two versions of the interview protocol: in each version, the participant was provided half of the explanations in text, and the other half in audio. Those explanations provided in text in version one were provided in audio in version two. To create the audio version of the explanations, we used the Google Speech-to-Text program, with the 'en-US-Wavenet-F' voice (a relatively neutral tone that was not identical to popular conversational assistants). We inputted the entire human or AI-generated explanation into the Speech-to-Text program, without any edits. Thus, the text and audio versions of the explanation were identical in wording.

[1] https://www.reddit.com/
[2] https://everydaysexism.com/
[3] https://twitter.com/home
[4] The online hate speech detection community recommends against publishing entire posts from the internet as they could be re-identified. However, we still presented the scenario and explanation as phrased online to participants as we were interested in the exact characteristics of the explanation that led to different perceptions, and did not want this to be influenced by our paraphrasing.

Table 1. Paraphrased scenarios and explanations used in the interview study. All scenarios and human explanations are paraphrased to protect the privacy of the human poster. AI explanations are verbatim. H=Human-authored, AI = AI-authored.

Scenario	Explanation	Author
A scenario describing a professional setting where a man refers to a fully grown woman as a girl.	An explanation suggesting that it drives the individual insane, because someone would not do the same thing to a man.	H
A scenario describing guys blaming everything on women's "biology".	It is sexist. It's a way of undermining women by reducing their capabilities as human beings to nothing more than their physical features. It's a way of saying that women have no control over their behavior. It's insulting. And it's not helping.	AI
A scenario describing using "guys" to address or describe a mixed-gender group	No, it isn't sexist. The word 'guys' is not a gendered word. It is not sexist.	AI

3.2 Participants

We recruited 20 participants (P1–20; 10 women, 9 men, 1 non-binary) by snowball sampling through the researchers' network. During this sampling process, we aimed to recruit participants balancing different genders, ages, and levels of experience with/understanding of AI technologies. In each interview, we asked participants if they knew anyone else who had different levels of experience with AI who would be interested in completing an interview. For example, some participants recruited colleagues from other departments, or classmates from other university programs. In this recruitment process, we shared that the study aimed to collect perceptions of human and AI explanations of subtle sexism scenarios—participants were told of the nature of the study before committing. Table 2 displays the demographic information for each participant. Participants were aged between 20 and 56 ($M = 30$), and their occupations spanned a variety of roles from students to company executives, both within AI (e.g., data scientists) and outside of it (e.g., visual effects artist). In terms of experience with AI-generated audio, four participants said they never used conversational agents, while 11 said that they always use them. Chatbots, or text-based AI tools, were used less often, with 13 participants saying they were only used sometimes. Most participants (11) had some knowledge of AI from the news or pop culture, while seven had academic or technical knowledge of AI, and two had little to no experience or exposure.

3.3 Procedure

We invited each participant to an hour-long interview session on Zoom. Interviews were conducted by the first two authors. Initially, a pilot interview was conducted with both authors to identify any challenges with the setup of the decision-making context and flow between portions of the interview. For the first third of the interviews conducted, both researchers were present, with one leading the interview and the other asking follow-up questions where appropriate. The remainder of the interviews were conducted by one of the two researchers. Both of these authors have formal training and multiple years of experience conducting interviews as a research method. Having a consistent set of interview questions and conducting the initial interviews together reduced discrepancies between the data collected by each interviewer.

We began with questions on demographics and the participants' familiarity with AI technology, which they described by mentioning various recent advances, e.g., DALL-E[1], Miquela[2], Netflix

[1] https://openai.com/dall-e-2

[2] https://www.instagram.com/lilmiquela/

Table 2. Profile of participants in the study

Participant Number	Age	Gender	Role	Use of Voice Assistants	Use of Chatbots	Knowledge of AI
P1	35	Man	Production Coordinator	Rarely	Sometimes	Some knowledge (news and pop culture)
P2	34	Man	Engineer	Never	Sometimes	Some knowledge (news and pop culture)
P3	26	Woman	Product Designer	Always	Sometimes	Some knowledge (news and pop culture)
P4	25	Woman	Product Marketing Manager	Never	Sometimes	Academic or technical knowledge
P5	25	Woman	Graduate Student	Always	Sometimes	Some knowledge (news and pop culture)
P6	27	Man	Network Security Developer	Aways	Sometimes	Academic or technical knowledge
P7	25	Man	Data Scientist	Sometimes	Sometimes	Academic or technical knowledge
P8	31	Non-binary	Product Manager	Always	Often	Some knowledge (news and pop culture)
P9	24	Woman	Student	Always	Sometimes	Some knowledge (news and pop culture)
P10	36	Woman	Graphic Designer	Never	Never	Some knowledge (news and pop culture)
P11	28	Man	Visual Effects Artist	Always	Sometimes	Some knowledge (news and pop culture)
P12	38	Woman	Student	Always	Sometimes	Some knowledge (news and pop culture)
P13	29	Woman	Digital Strategist	Rarely	Never	Some knowledge (news and pop culture)
P14	36	Man	Head of Design	Always	Sometimes	Academic or technical knowledge
P15	37	Man	Strategy Lead	Always	Rarely	Some knowledge (news and pop culture)
P16	21	Man	Student	Always	Sometimes	Academic or technical knowledge
P17	21	Woman	Student	Never	Never	Little to no knowledge
P18	20	Man	Student	Sometimes	Sometimes	Academic or technical knowledge
P19	25	Woman	Product Manager	Sometimes	Sometimes	Academic or technical knowledge
P20	56	Woman	Retired	Always	Never	Little to no knowledge

documentaries[3], biases in AI models, the debate regarding Google's LaMDA's sentience[4]. We then provided instructions, explaining the format of the rest of the study, and describing the context of subtle sexism through an example. This example explained that the participants were responsible for deciding whether a scenario was sexist or not, and they were provided with a perspective on this scenario from either a human counterpart or AI. In making this decision, they are asked to

[3] https://www.netflix.com/ca/title/81254224

[4] https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

consider whether they perceived that provided perspective as relevant, convincing, and trustworthy. Each evaluation characteristic—relevance, convincingness, and trustworthiness—was defined to participants at the beginning of the study using Webster's dictionary definitions, and we would repeat the definition at any time the participant requested.

Next, participants were presented with eight scenario-explanation pairs (in three cases, participants chose to skip a scenario), four human-authored and four AI-authored explanations. All scenarios were provided in text; each explanation was provided in text to half of the participants and in audio to the other half (i.e., a between-subjects experiment). For each scenario-explanation pair, we asked (1) whether a participant thought the explanation was authored by a human or an AI model, and why; (2) how relevant the explanation was, and why; (3) how convincing the explanation was, and why; (4) how trustworthy the explanation was, and why. Participants were encouraged to identify any words, phrases, or other elements of the explanation that contributed to their perception. Participants were able to re-read or re-listen to scenarios/explanations at any time up until the end of the interview. Interviews lasted an average of one hour, with few taking slightly longer, and few participants completing all scenarios in under one hour.

Each session concluded with debrief questions on the participants' overall experience and how the perceived author and explanation modality influenced their evaluation. We continued interviews until we had begun to reach saturation, and participants were no longer bringing up new explanation characteristics that we had not heard yet.

3.4 Data Analysis

Table 3. Table of high-level codes used in the study with definitions. Full codebook included in Appendix A

Characteristic		**Definition**
Tone	Emotionality	Level of emotion present in an explanation
	Formality	Level of formality (or conversationality) present in an explanation
	Neutrality	How much an explanation argues for a distinct opinion or remains neutral
Grammatical Elements	Length	The length of the explanation
	Grammar	Specific grammatical choices or adherence to grammatic principles
	Pronoun Type	The type of pronoun (personal or impersonal) used in the explanation
Argumentative Sophistication	Source of Argument	Where the main argument is sourced from (definitions, outcome, examples, etc.)
	Argument Structure	How the explanation is structured in terms of an explicit, direct, or complete structure
	Relation to Scenario	How much the explanation relates to the scenario
	Context Consideration	Whether the explanation shows some contextual understanding, generalizing the scenario to a broader societal context
Relation	Relation to User	Whether the explanation matches the user's personal opinions and experiences

All 20 interviews were anonymized and transcribed using the Rev[4] platform. Following the qualitative thematic analysis process [14], two researchers went through two rounds of initial open coding. First, three interviews were randomly chosen from the sample and independently open-coded by both researchers. Then, codes were consolidated and reorganized, resulting in the

[4] https://www.rev.com/

first iteration of the coding scheme, with separate schemes for the perception of authorship and the assessment of relevance, convincingness, and trustworthiness, respectively. Three more randomly chosen interviews were again independently coded by two researchers using this coding scheme, new codes were added and modifications were made until the scheme contained all identified open codes. Iterations of this process led to the realization that each of the four separate schemes (author, relevant, convincing, and trustworthy) had begun to merge—participants brought up the same explanation characteristics when explaining their perceptions. The coding schemes were merged into the final coding scheme (high-level codes shown in Table 3, entire codebook in Appendix A Table 7). Four randomly chosen interviews were coded by both researchers and once again arbitrated for a final consistency check; the remaining 16 interviews were distributed equally among the two researchers for full analysis.

We then conducted Wilcoxon signed-rank tests [158] to compare the achieved accuracy of author assessment (Human vs. AI) by modality and explanation characteristic (when a characteristic was perceived in the explanation vs. when it was not). We report on the statistically significant values in the result section.

4 RESULTS

Our qualitative analysis identified four major categories of explanation characteristics mentioned by participants when assessing the relevance, convincingness, trustworthiness, and authorship of explanations – *Tone, Grammatical Elements, Argumentative Sophistication* and *Relation to User* (Table 7). In this section, we describe how these affect the perception of its relevance, convincingness, and trustworthiness (RQ1) and the explanation authorship (RQ2), and discuss how each of these findings relates to previous work.

[P] denotes a participant; [H] – perceived human authorship; [AI] – perceived AI authorship. If [H] or [AI] is not listed, the quote was gathered at the time of debriefing and not directly related to the authorship assessment.

4.1 RQ1: Which explanation characteristics affect its assessment as relevant, convincing, and trustworthy? Does this differ by perceived author?

We discuss the effects of the explanation characteristics on its evaluation and compare the effects across perceived authorship (Table 4) and explanation modality (Table 5). We found that characteristics under *Grammatical Elements* did not affect the evaluation of the explanation in terms of relevance, convincingness and trustworthiness (and thus were omitted in Tables 4 and 5). In contrast, we saw that the *Relation to User* characteristics played a strongly positive role across conditions, while the effects of *Tone* and *Argumentative Sophistication* varied between perceived authorship, explanation modality, and evaluation category. We begin with the characteristics that affect relevance, convincingness, and trustworthiness differently for differing authors and modalities, and conclude with the characteristics believed to be beneficial across authors, modalities, and evaluations.

4.1.1 Relevant. Participants rated how relevant the explanation was to the scenario described and the question at hand (whether the scenario was sexist). In general, participants spoke to *features* present in the explanation when discussing relevance, more often than the tone of the explanation. Participants focused on the content of explanations, comparing this to the content of the scenario. This perception varied based on the perceived author: human explanations were relevant when neutral, direct, and containing examples, while AI explanations were relevant when they contained definitions and facts.

Table 4. Summary of characteristics and direction used to evaluate explanations, by perceived authorship. H = explanation perceived to be written by a human, AI = explanation perceived to be written by an AI. Characteristics that positively impact the evaluation are marked with a "+" in red, negative impacts are marked with a "-" in blue, and characteristics with both positive and negative impacts are marked with "+/-" in purple. Numbers represents the number of instances when a characteristic was discussed in the noted direction. Note: *Grammatical Elements* were not commonly used in the evaluation of explanations and thus are not in this table.

				Relevant				Convincing				Trustworthy			
Characteristic				H		AI		H		AI		H		AI	
Tone	**Emotionality**	**Emotional**						+/-	3/4			-	6		
		Logical								+	2				
	Formality	**Formal/Academic**													
		Informal/ Conversational										+	2		
	Neutrality	**Neutral**		+	2			+	3	-	2	+	2	-	2
		Defensive										-	3		
		Ambivalent/ Contradictory													
Argumentative Sophistication	**Source of Argument**	**Definition**				+	2	-	3			+	2	+	2
		Outcome												+	3
		Example	**External**	+	2			+	5	+	2	+	6	+	3
			Personal Experience					+	6	+	2	+/-	4/3	+/-	2/2
			Counterfactua	+	8			+	2						
		Opinion						-	2			-	5		
		Authority													
		Fact/Proof				+	2	+	8	+	4	+	7	+	5
		Similarity	**Online Content**												
			Humans												
	Argument Structure	**Formality**													
		Directness	**Direct**	+	2										
			Verbose												
		Completeness	**Completeness of Argument Consideration**	+	6	+	3	+	5	+	5	+	7	+	4
	Relation to Scenario	**Relation to Scenario**		+	16	+	15	+	7	+	3	+	4	+	6
	Context Consideration	**Contextual Understanding**		+	2	+	3	+	2	+	2	+	7	+	3
Relation	**Relation to User**	**Alignment with Personal Experience**		+	4	+	2	+	7	+	4	+	4	+	4
		Alignment with Personal Stance		+	9			+	10	+	5	+	10	+	6

Participants described that perceived human-authored, neutral-toned explanations (i.e. the explanation does not argue strongly for one side) were relevant:

> "[relevance] I personally like the nuance, **it is able to hit the two possible explanations for the behaviour at once**.." [P2; H]

Humans using counterfactual examples that described how an outcome would be different if the scenario were different in some way were identified as highly relevant to the scenario:

> "It's directly responding to the scenario and kind of flipping it and reversing...It wants you to consider **the foil to that scenario**...it's very relevant, and they're very connected" [P9; H]

Participants seemed to see the relevance of human-authored explanations when they were direct, as perhaps verbose explanations would stray from the scenario.

> "**It is getting to the point** of why I would think the scenario is sexist...So that's why I think relevance is there." [P12; H]

Table 5. Summary of characteristics and direction used to evaluate explanations, by modality. Characteristics described to positively impact the evaluation are marked with a "+" in red, negative impacts are marked with a "-" in blue, and characteristics with both positive and negative impacts are marked with a "+/-" in purple. Numbers represent the number of instances when a characteristic was discussed in the noted direction. Note that the theme *Grammatical Elements* was not commonly used in the evaluation of explanations and is thus not in this table.

				Relevant				Convincing				Trustworthy			
Characteristic				Text		Audio		Text		Audio		Text		Audio	
Tone	**Emotionality**	**Emotional**						+/-	2/2	+/-	2/2	+/-	2/6	-	4
		Logical						+	2			+	2		
	Formality	**Formal/ Academic**													
		Informal/ Conversational												+	2
	Neutrality	**Neutral**						-	2	+	2			+/-	2/2
		Defensive										-	2		
		Ambivalent/ Contradictory													
Argumentative Sophistication	**Source of Argument**	**Definition**						-	2	+	2	+	2	+	2
		Outcome								+	2	+	2		
		Example	**External**	+	3			+	4	+	4	+	4	+	3
			Personal Experience			+	2	+/-	3/2	+	6	+/-	3/2	+/-	5/3
			Counterfactua	+	5	+	2	+	2	+	2			+	2
		Opinion								-	2	-	5	-	2
		Authority												+	2
		Fact/Proof						+	7	+	7	+	9	+	8
		Similarity	**Online Content**												
			Humans												
	Argument Structure	**Formality**													
		Directness	**Direct**					+	2						
			Verbose												
		Completeness	**Completeness of Argument Consideration**	+	7	+	3	+	5	+	8	+	8	+	7
	Relation to Scenario	**Relation to Scenario**		+	17	+	15	+	3	+	6	+	7	+	3
	Context Consideration	**Contextual Understanding**		+	3	+	3	+	2	+	4	+	5	+	5
Relation	**Relation to User**	**Alignment with Personal Experience**				+	5	+	8	+	5	+	3	+	6
		Alignment with Personal Stance		+	7	+	7	+	13	+	7	+	10	+	13

Facts and definitions, while believed to be AI-generated, were also identified as relevant as they corresponded directly to the terms used in the scenario:

> "[relevance] They **define** what the term is and [why] they think it's important..."[P4; AI]

4.1.2 *Convincing.*

Participants provided a variety of characteristics which influenced how convinced they were by an explanation. Both the tone of the explanation and the argumentative sophistication played a role in its ability to convince. The presence of external or personal examples and facts/proof were convincing regardless of the author they came from. Participants also described characteristics as convincing only when written by a human, such as a neutral tone or counterfactual examples. Similarly, a logical tone was only convincing when AI-generated. Convincingness varied across modalities, with neutral text explanations seeming less convincing and neutral audio explanations more convincing.

Across perceived human and AI-generated explanations, facts and external examples helped to convince participants:

> "In order to really convince completely, then I'd say **some data and peer review studies and whatnot** could add to that." [P2; H]

Personal experiences were also convincing across authors, even though participants often attributed these experiences to humans:

> "If I was in a conversation and there was a human woman saying this in rebuttal, and I would say [it is convincing] because it's someone who's speaking **from lived experience**. So if I'm leaning on the train of this is human, then...it's very convincing" [P15; H]

An emotional tone was only attributed to humans, and it had mixed effects on how convincing the explanation was perceived to be:

> "I think when you use the **feelings** that [it] evoke[s]...is more appealing to people. So I think it's convincing" [P10; H] **VS.** "[convincingness] I had a moment of doubt as to whether there was some **overreaction** in the explanation. " [P9; H]

Though the opposite of an emotional tone — a logical one — was determined to be convincing to participants when AI-authored:

> "[convincingness] it's a very **rational explanation**...It kind of feels like there is some...emotional distance when they're explaining it because they're wanting you to...take on a more rational stance" [P9; AI]

A neutral (impartial) tone was seen as convincing when thought to come from humans, as this might display nuanced thinking, but less convincing if perceived as AI-authored:

> "It would be convincing...[because] it starts out with, 'it might be used in a different context now'. So it's **trying to be very unbiased**" [P7; H] **VS** "...It's more like trying to be more understanding, which is also why it's throwing me off. It's kind of **trying to be neutral instead of picking one either side**." [P5; AI].

Modality also played a role in the perception of a neutral tone — it was less convincing in text form, and more in audio form.

It was also uniquely human for definitions and opinions to be perceived as unconvincing; participants looked for facts or personal experiences as convincing arguments:

> "[convingcingness]...in my belief, the origin of the word doesn't really have anything to do with certainly my perception of hysterical." [P20; H]

Compounding this effect, explanations containing definitions and displayed in text form were especially unconvincing.

4.1.3 Trustworthy. Participants used similar explanation characteristics when describing trustworthiness as they did convincingness. Participants shared that the tone of the explanation and the source of the argument played the largest role in whether it was perceived as trustworthy. The effect of definitions (positive), external examples (positive), personal experience (mixed) and fact/proof (positive) did not differ across authors. Neutral tones were only trustworthy for humans and decidedly untrustworthy if believed to be AI-generated, and several characteristics were only discussed in terms of human-authored explanations.

Across authors, definitions were perceived as trustworthy due to their source (the dictionary):

> "I would actually give it five out of five for trustworthy. Predominantly because it is **pulling it from the dictionary**" [P13; H]

Similarly, external examples and facts/proof were seen as trustworthy regardless of author.

> "**...there's no facts, no data I can rely on** ... you're presenting something new. So that's why I don't think it's...trustworthy, no data, no facts, nothing pointing to a reference or whatever." [P8;AI]

Participants had mixed reactions to seeing personal experiences in explanations, in both perceived human and perceived AI explanations. Some participants explained how personal experiences relate to an innate sense of being able to trust someone:

> "because it seems like it doesn't come only from academic experience but also from **personal experience firsthand**. So ...it appeals to our senses to trust one that can handle both" [P10; H]

However, they also noted that combining these personal experiences with more objective facts or definitions can add to the credibility, and this is especially important when perceived to come from AI:

> "...**I don't wanna discount another human's experience because that could be their subjective truth**. Doesn't mean I have to accept that, but ... I can envision how that would show up for someone in their life...**When it's AI...I receive that information a bit differently**...I probably don't ascribe as much trust to it...I want to take one additional step to validate it before I give it more credibility." [P15; H]

Other characteristics only affected human explanations, such as informal, defensive, or emotional tones. Informal tones had a positive effect on trustworthiness, while defensive tones had a negative effect. An emotional tone in perceived human-authored explanations had negative effects on the perceived trustworthiness (more so in an audio condition):

> "I think for most people, as soon as they hear something **emotional** that upsets them, it automatically...gives you less trustworthiness [because] it makes you seem **more biased**..." [P7; H]

Similarly to the assessment of convincingness, a neutral tone was only perceived as trustworthy when believed to come from humans, as it would indicate an unbiased view on the topic, while AI-generated explanations were perceived as contradicting themselves:

> "they're contradicting themselves and [it] impacts your trust" [P13; AI].

Also when explanations were perceived to be human-authored, the use of opinions had a negative effect on the perceived trustworthiness:

> "From my perspective, **it's just an opinion**. So you can't necessarily trust it 100%" [P7; H]

The only characteristic with a unique effect on AI-generated explanations was referencing outcomes as the source of argument, which increased perceptions of trustworthiness.

4.1.4 All Assessments. A number of characteristics were described as having a positive impact across authors, modalities, and evaluations (relevance, convincingness, and trustworthiness).

Three characteristics under *Argumentative Sophistication* that were commonly brought up—completeness of argument consideration, contextual understanding, and relation to the scenario—had a positive effect on relevance, convincingness, and trustworthiness:

> "I don't think it's a convincing argument. It's a statement, but I'm not sure that it has **enough breadth or depth to it** to be convincing to really drive that point home." [P7; AI]

> "[trustworthiness] the person or AI that said it has potential because **it is already trying to notice the nuances** and if there are allies or not. So it gives me the impression that **the person understands what they're talking about**..." [P10; H]

> "Yeah, it's completely addressing the prompt that it's being given out, but it's **not necessarily doing so in a way that...connects to it properly**. It sort of has its own prompt within... their own thought of that topic..." [P7; H]

Our findings show that the closer an explanation relates to the participant's own experience and opinions, the more it affects their assessment of this explanation as relevant, convincing, and trustworthy across perceived authorship and modalities:

> "I trust it. I think I build trust based on **my own experiences** and situations around me. And I've seen this type of thing play out. " [P19; AI]

> "[trustworthiness] it's also **in line with my personal understanding**. Your gender doesn't define who you are or what you wanna work on." [P3; AI]

We also note that many, though not all, of the same explanation characteristics influenced perceptions of convincingness and trustworthiness. As shown in Table 4, the effect of external examples and facts/proof were the same across convincing and trustworthy. While we defined each term for participants and repeated these definitions throughout the study, some participants stated that these two assessments were related; they were convinced only if they found something trustworthy:

> "I think **trust... [and] convincing kind of go hand in hand directly**. And because I found it convincing, I think I would trust it a lot." [P9; H]

Despite this similarity, we found that participants would repeat elements of the definitions of convincing and trustworthy in their responses, suggesting they understood them as two separate constructs. In this example, the dictionary definition of trustworthy included "reliable and true":

> "Trustworthy? Yeah. I mean it is based on someone's opinion...So **is it reliable? Yes, it's reliable. Is it true? I don't know** because in this explanation you have to think for yourself..." [P8; H]

While some explanation characteristics affect convincing and trustworthy perceptions similarly, differences have emerged in this work which can be further teased apart in future work.

4.1.5 Summary and Contextualization. Our findings provide the first classification of characteristics that affect the perception of explanations for collaborative decision-making. Notably, we find that within all three evaluation categories (relevant, convincing, trustworthy), the effect of explanation characteristics varies, and the perception of the explanation author affects this perception. We extend past work that describes that explanations should be relevant [16] by outlining the specific characteristics affecting the perception of relevance in this context. We identify many new relationships between explanation characteristics and perceptions not yet identified in past work and confirm other relationships. We confirm that counterfactual-style explanations are perceived as convincing [20], but not trustworthy [149] and that the author plays a role in the trustworthiness of an explanation [65, 88]. We extend this further to describe how the author affects how some, but not all, explanation characteristics are perceived, further motivating our second research question. Some of our findings contradict past work, highlighting context's importance in subjective decision-making. We did not find that an ambivalent tone influenced perceptions of trustworthiness, contrary to Schaekermann et al. [132]. We identified more nuance within the effect of personal explanations on trustworthiness, building on Kunkel et al. [91] that they are trustworthy, but more so when accompanied by a fact-based argument.

4.2 RQ2: Which explanation characteristics do people use to predict the authorship?

As hypothesized, we found that the perception of relevance, convincingness, and trustworthiness of explanations does depend on the perceived authorship. Thus, we now present patterns in how participants used explanation characteristics to explain the perceived authorship of an explanation, their difficulties in identifying authorship, and conclude by discussing findings in the context of previous literature. Notably, participants predominantly identified positive (if A then B) rather than negative (if A then not B) characteristics (Table 6). There were a number of characteristics that clearly distinguish human and AI explanations for our participants—for example, emotional tones in human explanations versus logical tones in AI examples, and the use of personal experiences by humans versus facts by AI. We first describe these opposite characteristics, and then comment on the characteristics that were more contradictory—ascribed to both human and AI authorship.

4.2.1 Human. We outline the explanation characteristics (*Tone, Grammatical Elements, Argumentative Sophistication,* and *Relation*) that participants used when describing the authorship of explanations. All four types of characteristics were used frequently. Broadly, human explanations were described as sounding emotional and informal, having poor grammar/structure, and demonstrating a broad contextual understanding and a personal connection to scenarios.

Explanations thought to come from humans were described as containing an emotional and informal tone, as opposed to logical and formal tones. Participants recognized tone characteristics through word choices that were perceived to express a more detached view (logical tone) or a more personal or passionate position (emotional tone). This was true across modalities, and the rationale was predominantly related to vocabulary choices (content) rather than voice inflection (delivery).

> “I would expect a machine to be a little more robotic and not use **very extreme words** [like] ‘driving me insane’ or ‘ridiculous’.” [P13; H]

In terms of neutrality of tone, humans were described as either being defensive (e.g., strongly arguing for one side) or being ambivalent (e.g., taking into consideration multiple sides of an argument to the point of showing indecisiveness or even contradiction). Participants found this to be a sign of human complexity:

> “there’s this uncertainty. It starts with ‘I’m not sure’, and then the next sentence goes.. why it is not being sure...So yeah, there’s this duality in the answer, and I think that’s **indicative of nuance**.” [P2; H]

Participants generally described human-authored explanations as having poor or less complex grammar and structure, e.g., a person hastily typing and making mistakes. Human explanations were identified as lacking in punctuation and containing poor sentence structure. Interestingly, ellipsis punctuation (i.e., ...), used to indicate hesitation or incomplete thought, was also seen as uniquely human:

> “I just don’t feel like AI’s gonna produce the punctuation marks here too. The dot dot dot...that’s kind of stand out and just **it reads how someone would type it**.” [P12]

Finally, personal pronouns (I, me, my) were associated with human writing from a first-person perspective.

Explanations with a broad contextual understanding of the issue, such as pointing to nuances in societal structures or indicating a personal connection to the situation, were perceived as uniquely human.

> “It’s actually coming with **real-life circumstances and scenarios** of why this might not be sexist... So I feel like whoever explained this has **real-world experience**, and it’s not just something that was trained to say the right thing.” [P5]

Interestingly, participants who noted contextual understanding as an inherently human characteristic performed significantly worse at predicting authorship ($p = 0.045$), while those who relied on the personal experience characteristic performed significantly better ($p = 0.036$).

Table 6. Summary of characteristics and direction used to assess authorship of an explanation. Human = explanation perceived to be written by a human, AI = explanation perceived to be written by an AI. Characteristics described to be positively associated with the author are shown with a + in red, characteristics described to be negatively associated with the author are shown with a - in blue, and characteristics described to have both positive and negative associations are shown with a +/- in purple. Numbers represent the number of instances when a characteristic was discussed in the indicated direction.

				Human				AI			
Characteristic				Text		Audio		Text		Audio	
Tone	**Emotionality**	**Emotional**		+	12	+	13			-	5
		Logical				+	2	+	2	+	4
	Formality	**Formal/ Academic**						+	5	+	10
		Informal/ Conversational		+	14	+	9				
	Neutrality	**Neutral**						+	3		
		Defensive		+	7	+	3				
		Ambivalent/ Contradictory		+	6	+	3				
Grammatical Elements	**Length**	**Short**		+	3	+	3	+	2	+	5
		Long						+	2		
		Repetition		+	2					+	3
	Grammar	**Double Negatives**									
		Correct								+	3
		Punctuation		+/-	3/4			+	3		
		Sentence Structure		-	2			+	4	+	2
	Pronoun Type	**Personal**		+	9	+	4				
		Impersonal									
Argumentative Sophistication	**Source of Argument**	**Definition**						+	3	+	3
		Outcome									
		Example	**External**			+	2				
			Personal Experience	+	8	+	9	-	2		
			Counterfactual	+	2	+	2				
		Opinion									
		Authority									
		Fact/Proof				+	2	+	4		
		Similarity	**Online Content**			+	3	+	5	+	3
			Humans					+	4	+	2
	Argument Structure	**Formality**				+	3	+	8	+	2
		Directness	**Direct**	+	8	+	4	+	3	+	7
			Verbose								
		Completeness	**Completeness of Argument Consideration**	+	3	+	5			+	4
	Relation to Scenario	**Relation to Scenario**						-	3		
	Context Consideration	**Contextual Understanding**		+	10	+	7	-	3	-	4
Relation	**Relation to User**	**Alignment with Personal Experience**		+	8	+	9			-	3
		Alignment with Personal Stance		+	5			-	3		

While the presence of personal experiences in an explanation already led participants to believe a human authored it, when the experience aligned with the participant's own, this strengthened the effect:

> "**I find it relatable, that explanation to my career, to my job** and everything. So I feel connected to that explanation, or that makes me feel that it was done by a human." [P11, H]

4.2.2 AI. When describing why an explanation was believed to be AI-authored, participants often used exact opposite arguments than they used for humans. As such, AI explanations were described as using a formal or academic tone, with the correct use of complex grammar and structure, and based on facts or definitions. AI explanations were perceived to have a logical, formal, or neutral tone — the exact opposite of human's emotional, informal, and defensive tones.

> "AI, I would imagine, would try to find this kind of simple explanation, that is **not related to exactly a feeling** because, I guess, AI cannot have feelings for now." [P10; AI]

Similarly, participants believed AI produces correct grammar and complex structure, resulting in longer explanations and more punctuation, in contrast to humans' hastily written explanations.

> "I think it's kind of a **little bit complicated as a sentence structure** cause you are using a negative...and then a positive attitude." [P11; AI]

Whereas humans were expected to include broad contextual understanding and personal experiences in their explanations, AI explanations, on the other hand, were expected to resort mostly to facts or dictionary definitions:

> "It's kind of tacky to immediately **start off with the dictionary** [definition]...I feel like maybe a computer did that." [P5; AI].

Lastly, even when personal experiences were included in an explanation if they were misaligned with the participant's own experience or view of the world, it was considered AI-generated.

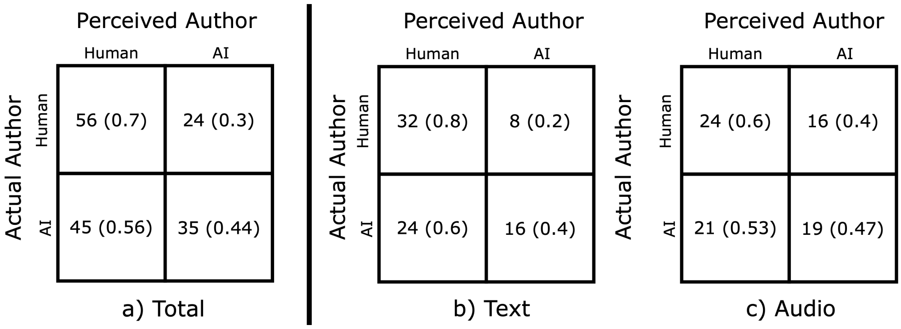
4.2.3 Contradictory Arguments. Some characteristics were attributed to humans by some participants, and to AI by others. For example, a short explanation indicated AI authorship for some, while others perceived it as uniquely human. Similar contradictions were found for the use of repetition, completeness of argument, and directness (Table 6).

> "This sounds human to me. It's a very **direct statement**" [P15; H] **VS** "I feel like that's generated by AI. It's brief, it's **straight to the point**. It doesn't really have too much meat or flesh around it..." [P12; AI]

4.2.4 Difficulty to Assess. Our quantitative analysis showed that participants achieved an overall 56% accuracy when assessing the author of an explanation, close to chance. We also compared the accuracy of participants who identified each explanation characteristic to those who did not, and only two characteristics, out of all of those shown in Table 6 displayed a statistically significant difference in accuracy. As reported in the subsections above, identifying personal experiences in explanations helped participants to identify the author of explanations, whereas identifying contextual understanding harmed their performance on this identification. No other characteristics were significantly related to correct or incorrect author identification, suggesting that many of these characteristics exist in both human and AI-generated explanations.

Interestingly, participants were better at identifying human explanations (Figure 3), partially because they assessed more explanations as being human-authored overall (62.5%), although the study design intentionally had a 50/50 split. Indeed, participants often expressed difficulties in assessing whether an explanation was human- or AI-authored, commonly second-guessing themselves. Particularly, participants wrestled with the idea of an AI trained on human data and thus sounding human-like:

> "It just sounds like very casual language, which makes me conflicted as to whether it's human or AI, because is this **AI's interpretation** of how terribly we talk, or is this human?..." [P15; AI]



		Perceived Author	
		Human	AI
Actual Author	Human	56 (0.7)	24 (0.3)
	AI	45 (0.56)	35 (0.44)

a) Total

		Perceived Author	
		Human	AI
Actual Author	Human	32 (0.8)	8 (0.2)
	AI	24 (0.6)	16 (0.4)

b) Text

		Perceived Author	
		Human	AI
Actual Author	Human	24 (0.6)	16 (0.4)
	AI	21 (0.53)	19 (0.47)

c) Audio

Fig. 3. Confusion matrices for authorship identification. "Actual Author" represents whether the explanation was AI-generated or human-generated in our dataset, and "Perceived Author" represents whether participants believed it to be human- or AI-generated. The numbers in each box represent how many interview instances fell into each category. (a) represents the accuracy across all scenarios, and (b) and (c) split this accuracy by modality. Percentages indicated in brackets are calculated based on "Actual Author" totals. For example, in 70% of cases, the human-authored explanations were perceived to be human (correct identification), and in 30% of cases they were perceived to be AI-generated (incorrect identification)

We found no statistical difference in authorship assessment accuracy between audio and text conditions ($p = 0.649$). However, we found that in text condition, participants performed better at identifying a human-authored explanation (80%), compared to audio (Figure 3 b and c), while audio led to a small increase in accuracy for AI-authored explanations (47%). Interestingly, we saw that in audio explanations, it was more difficult for participants to recall content, making it harder to assess authorship compared to text format:

> "with the audio... **I would try my hardest to just remember every word**, but it would've been nice to have the audio also as a paragraph...because some things from the audio stood out more than others. So then I would only remember and fixate on those things versus the general message..." [P5]

Participants also noted that the gendered voice in the audio condition might have impacted their assessment, especially in the context of sexist scenarios that describe actions against women:

> "all the audios were actually with a **female voice**, right? So I'd be interested to see whether the male voice would've made any difference..."[P7]

4.2.5 Summary and Contextualization. We identified characteristics of explanations generally associated with only human or only AI generation, as well as those with a less clear association. Human explanations are believed to be emotional, informal, with less complex structure, and include personal experiences and contextual understanding. AI explanations, on the other hand, were associated with logical and formal tones, correct and complex grammar, and factual argument sources. Our findings showed the importance of *Tone*, *Argumentative Sophistication*, *Relation to User*, and *Grammatical Elements*, which provide additional details and depth to the content and form categorization proposed for human evaluation of AI-generated text [31]. However, while Clark et al. [31] found that participants tend to comment predominantly on the form of the text, participants in our study relied on both the form and the content characteristics. Some of the identified explanation characteristics, such as definitions, counterfactuals, examples, and fact/proof have parallels in the explanation knowledge categories from Wang and Yin [149] (definition, comparison, example, statistical). However, participants did not identify all knowledge categories from past work, which suggests that only a subset of those may be most relevant for explanations in collaborative subjective decision-making. While some of the identified characteristics of explanations can be found in past work, including grammar, personal pronouns, punctuation, and life experiences [31, 42, 58, 77], other characteristics, e.g., the completeness of the argument, contextual understanding, and relationship to the user were not previously reported and may be specific to the context of collaborative subjective decision-making. Unlike past work [31], participants in our study associated correct grammar and punctuation with AI — perhaps hinting at more widespread knowledge of generative AI.

5 DISCUSSION

The aim of this work was to understand which *characteristics of a verbal explanation* affect the perception of (RQ1) relevance, convincingness, and trustworthiness and (RQ2) its authorship (human vs. AI) in the domain of subtle sexism detection. In our analysis, we found it particularly important to consider these two questions together, which allowed us to capture the interplay of the corresponding perceptions. We identified that the perceived author of an explanation affects how different explanation characteristics are perceived. For example, users may have more positive impressions of explanations in these sensitive contexts when they are perceived to come from humans. We found that some characteristics associated with human authorship, e.g., contextual understanding and alignment with personal stance, positively impacted perceptions of relevance, convincingness and trustworthiness of an explanation. Furthermore, many characteristics associated with human authorship affected the evaluation of explanations perceived to be AI-authored. Perhaps this was because it took participants longer to recognize some of these explanation characteristics, as they were asked to assess authorship before evaluating the explanation. Or perhaps participants did not consciously consider the author of the explanation when evaluating it. In summary, the explanations that participants found the most relevant, convincing, and trustworthy contained mostly characteristics associated with human authorship, suggesting that, in these subjective contexts, people may see human suggestions more favourably. However, further research is required to establish the exact nature of the relationship between the perceived authorship and the overall evaluation of an explanation.

This work also contributes to the larger conversation on the collaborative communication dynamics between humans and agents [30, 133, 144]. While most of this work considers how a human converses with an agent [82, 144], comparatively fewer studies investigate how humans perceive the communication from the agent [135]. Yet, we show that this perception — of the agent, and the communication it produces — does matter for the overall evaluation of the explanation, which influences acceptance in collaborative scenarios.

While scoping this study to one context allowed participants to focus on a specific scenario, making the decision-making process more realistic, it comes with generalizability considerations. Our focus on subtle sexism as the domain of this collaborative subjective decision-making may affect the explanation categories that were uncovered—for example, alignment with personal stance and experience may be less important in less personal contexts. We expect these findings to be more comparable to other content moderation contexts, such as detecting other forms of hate speech like racism, than to general subjective contexts like hiring. As we continue to study human-AI collaboration in subjective decision-making, we should broaden the scope of domains studied to tease apart domain-specific explanation characteristics from those more broadly important.

However, we expect the implication that perceived authorship can affect how these characteristics are interpreted to be critical for the design of collaborative decision-making systems, even generally: in the absence of author labels, users will associate authorship based on the characteristics of the explanation and assess the effect of these characteristics differently as a result. This contributes to the broader discussion of whether AI-generated output should be labelled as such [45, 157]. On the one hand, system designers can include "generated by AI" labels on explanations or content; however, these could be easily missed, or purposefully placed in hard-to-find areas. Another approach might be to synthesize the AI output to be in line with human perceptions of what it should be — however flawed those are. This may allow for content to be more automatically recognized as AI.

We also saw that, when assessing authorship, participants often struggled and questioned their beliefs regarding AI's abilities. This study took place before the public release of ChatGPT in November 2022; users' familiarity with the capabilities of LLMs is likely to change, especially as they are rapidly embedded in everyday products. This study can act as a baseline of these perceptions that we can use to monitor changes as people are more exposed to LLMs.

5.1 Cognitive Biases and Ethical Considerations

There are a number of findings in this work that we believe highlight the unique characteristics of explanations in collaborative subjective decision-making, particularly in sensitive cases such as subtle sexism. In this subsection, we discuss several resulting ethical considerations.

5.1.1 The Reinforcing Role of Explanations. We found that participants recognize experiences or opinions in explanations that align with their own, and perceive these as more relevant, convincing, and trustworthy. Arguments sourced from personal experience or opinions were perceived as convincing, but there was a mixed impact on trust— except when the experience or opinion displayed aligned with their experience; in that case, they found the explanation both convincing and trustworthy. This surfaces the role of confirmation bias—defined as "seeking or interpreting evidence in ways that are partial to existing beliefs" [114, p. 175]—posing the risk of reinforcing the user's existing opinions and negating the intention of providing new perspectives to the decision-making process. Creating explanations that prompt users to consider other points of view may need to lean more heavily on facts, which were perceived as trustworthy. Alternatively, these systems could first collect contextual opinions and experiences from the participants, and then show them explanations that do not directly align.

5.1.2 Convincing AI Explanations. Participants described AI explanations as convincing, particularly when they contained examples, and facts, or were aligned with their personal beliefs/experiences. This risk from AI models being able to convince humans to believe a certain story or point of view has been well-studied [12, 83, 115, 118]. Scholars have discussed how AI algorithms may create filter bubbles on social media [28], or how news site curation can spread misinformation [53], aided by artificial intelligence [83]. The goal of human-AI collaborative systems is to achieve trust calibration [156, 161], where the human decision-maker does not blindly follow the AI recommendation but instead forms trust appropriate to the model behaviour. Paired with the known limitations of current LLMs [29], uncalibrated trust in the perspectives provided by AI could be detrimental, particularly in sensitive contexts. Nourani et al. [116] suggests that exposing humans to poor AI explanations before correct ones helps to avoid overreliance, though it is less clear what constitutes a poor explanation in subjective contexts. As shown by Buçinca et al. [17], systems that encourage analytical thinking during the decision-making process can reduce overreliance on AI decisions. Perhaps future collaborative systems should require human decision-makers to provide a final explanation for their decision, specifying the ideas provided by AI.

5.1.3 Mimicking Humans in AI Explanations. Despite the preference for explanations that align with their opinions and experiences, humans also find it troubling when LLMs generate attributes that are considered uniquely human, such as personal experiences or emotions [142], and their perceptions towards AI, in general, are still forming. Even experts in the field debate which human-like abilities these models really have [66]. Particularly in cases like sexism, where one may be attached to their experiences, "made up" experiences produced by a model should be used with great caution. The popular example of the Google employee who believed that the company's LLM was sentient [32] shows that these models can deceive and confuse users. This has led to blogs explaining why LLMs cannot actually generate subjective experiences [79], as well as calls to the academic community to stop using anthropomorphizing language when describing AI [137, 138]. Perhaps, a warning that the explanation is model-produced (which is being recommended as policy in the United States [125]) may be required, though we also saw in our study that some participants perceived AI-generated explanations as less convincing.

5.1.4 Gender-Based Assumptions. We found that participants often assumed the gender of characters in the scenario when no indicators were present, commonly assuming the victim was a woman. While, indeed, the majority of sexism on the internet is still directed towards women [128], sexism, affects all genders, and in particular trans or non-binary individuals [117, 120, 122]. This bias in participants' reasoning suggests that argumentation, generated for the cases where gender or other demographic variables are important, may need to include these indicators in scenarios.

5.1.5 The Inherent Bias in LLMs. AI applications can perpetuate cultural biases [21], LLMs are capable of producing hate speech and stereotypes [29], and even the most recent language models still contain gender bias [87]. This is a critical issue when looking at subtle sexism, where protected attributes like gender are bound to be included in the discussion. Researchers are developing safeguards to prevent these biases from making their way into downstream tasks [159]. Yet, there is still a long way to go, as users are devising workarounds to get public models to produce hateful content [155]. However, given that LLMs are currently being proposed for human-AI collaboration systems [136], it is important to study these systems in their current state.

5.2 Design Implications

AI explanations in collaborative subjective decision-making cases such as subtle sexism detection can bring new perspectives for a human user to consider [50, 93]. However, these AI-generated explanations must be relevant, convincing and trustworthy to foster collaboration in the hopes of achieving complementarity [9]. Hence, our work contributes to informing the design of such AI systems meant to augment subjective decision-making by offering four *Design Guidelines* related to explanation characteristics that we found to be brought up most often by our participants: Tone, Supporting Evidence, Societal Context and Diverse Perspectives.

Tone: AI-generated explanations can be made less convincing and trustworthy when they are perceived to have an emotional tone. This may be an intended or unintended effect, depending on the context. Additionally, it is also associated mostly with human-authored explanations. The emotional tone was identified by the presence of text descriptions of feelings or words that express them (e.g., "angry", "frustrated", etc.). Therefore, these words and phrases can be used strategically. Participants preferred AI to display a more decisive tone in its assessment. Nonetheless, one must also consider that excessive confidence in tone might lead to overreliance and inappropriate trust calibration [156, 161] as discussed in the previous section. In human-AI collaboration tasks, avoiding emotional tone can also help the human counterpart in identifying the author as AI by matching their mental model and expectations of how a machine behaves [57, 66].

Supporting evidence: AI-generated explanations are perceived as more convincing and trustworthy when they cite facts and external examples to support the argument. Additionally, referring to the data sources (e.g., the dictionary, scientific reports, etc) also makes these explanations more convincing and trustworthy to the user. This can be achieved, for instance, with post-hoc explainability methods [145]; however, that can be challenging in terms of traceability and data provenance [81]. Alternatively, if using LLMs to generate explanations, these models can be fine-tuned with sources relevant to the specific context. In the human-AI collaboration context, leveraging supporting evidence can be beneficial for trust calibration by offering the human counterpart the facts and sources behind the decision.

Societal Context: Interestingly, for explanations to be considered relevant, convincing, and trustworthy, participants required a balance between scenario specificity and broader contextualization. The first relates to a preference towards explanations directly related to the scenario at hand, versus generic statements about sexism or gender, for example. The second, however, prescribes that explanations should connect the specific scenario to broader societal contexts (contextual understanding). To achieve the correct balance, AI-generated explanations must first address the scenario at hand, and then connect it to the issues within the broader societal context. This may mean providing relevant social context to the model (in the form of a prompt, or fine-tuning), and using prompt engineering to ensure the model focuses on the scenario provided.

Diverse Perspectives: Our work also confirms that participants were more biased towards explanations that aligned with their personal beliefs and experiences, as prescribed by Mitamura et al. [107]. Hence, AI systems designed to augment subjective value-based decision-making tasks should offer explanations with a mix of diverse perspectives the human counterpart might have ignored to potentially counter-balance such biases. This could be achieved with prompts that require models to consider multiple perspectives or could be designed as a multi-agent system, where multiple models with different "personas" all provide input to the user. A potential challenge in this approach is mitigating social biases that foundational models, specifically LLMs, have been known to replicate [38].

Further, Figures 1 and 2 represent the information from Tables 4 and 6 for use in system design. We present this diagram to aid system builders in designing explanations for a specific perception, such as the appropriate amount of trust in the system, for instance by balancing positively and negatively perceived features. These design considerations aim to foster human-AI collaboration and complementarity; however, as discussed previously in ethical considerations, these mechanisms should be applied cautiously and reflect considerations of the potentially evolving users' mental models.

6 LIMITATIONS

There are some limitations of this work that we must keep in mind when interpreting the results. As early-research cycle work investigating collaborative subjective decision-making, the exploratory nature of the study determined a limited sample and scenario size (20 participants, eight scenarios). The scenario limit was also determined by the study length considerations and it means that we cannot generalize our results to all possible subtle sexism scenarios. Now that this preliminary work has identified the explanation characteristics that play a role in the quality of explanations for subjective decision-making and the perceptions that can influence collaboration, we can test them in a controlled study using more explanations from our dataset and a larger number of participants, e.g., through crowdsourcing. In particular, as we found few significant relationships between explanation characteristics and correct author identification, future work should test these relationships with more participants in order to understand how to design these explanations for correct author identification in unlabelled scenarios.

We also note that the domain chosen here (subtle sexism) likely influenced the specific explanation characteristics which were identified by participants. Thus, these findings and specific design guidelines are best suited for other subjective and sensitive contexts like hate speech detection and content moderation. While the ultimate understanding of how perceived authorship can influence the perception of different explanation characteristics such as relevant, convincing, and trustworthy applies to collaborative decision-making broadly, future work may need to identify relevant domain-specific explanation characteristics.

Another limitation is that participants sometimes struggled to distinguish between convincing and trustworthy in their evaluation of explanations, providing the same rationale for both (though not always, Table 4). While we provide some details regarding the characteristics differentiating perceived convincingness and trustworthiness, future work could more clearly differentiate between the two concepts.

Finally, this work focused on subjective perceptions and evaluations. Past research has shown that subjective ratings of explanations may not always align with objective performance on those tasks [16, 93]. Thus, before human-AI collaboration systems are put into practice for subjective decision-making, they should be formally tested in the context in which they will be used.

7 CONCLUSION

In collaborative subjective decision-making, AI explanations and argumentation can surface new perspectives critical for consideration. However, to be considered, they should be perceived by a human user as relevant, convincing, and trustworthy. Focusing on verbal explanations, we investigated which explanation characteristics affect the perceptions of the explanations and found four groups of characteristics: *Tone, Grammatical Elements, Argumentative Sophistication* and *Relation to User*—the effect of which differed based on the perceived author of the explanation. We described the relationships between the emerged characteristics and perceptions, discussed the interplay between the perceptions, and provided ethical and design considerations for AI systems supporting collaborative subjective decision-making.

8 ACKNOWLEDGEMENTS

This work was supported by NAVER corporation as a part of NAVER-Wattpad-University of Toronto research center.

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] Resham Ahluwalia, Himani Soni, Edward Callow, Anderson Nascimento, and Martine De Cock. 2018. Detecting hate speech against women in english tweets. *EVALITA Evaluation of NLP and Speech Tools for Italian* 12 (2018), 194.

[3] Alican Akman and Björn W Schuller. 2024. Audio Explainable Artificial Intelligence: A Review. *Intelligent Computing* 2 (2024), 0074.

[4] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*. 184–190.

[5] 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.

[6] Zahra Ashktorab, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 96 (oct 2020), 20 pages. https://doi.org/10.1145/3415167

[7] Krisztian Balog, Filip Radlinski, and Andrey Petrov. 2023. Measuring the Impact of Explanation Bias: A Study of Natural Language Justifications for Recommender Systems. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 203, 8 pages. https://doi.org/10.1145/3544549.3585748

[8] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proceedings of the ACM*

on Human-Computer Interaction 7, CSCW1 (2023), 1–17.

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[10] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[11] Or Biran and Kathleen McKeown. 2014. Justification narratives for individual classifications. In *Proceedings of the AutoML workshop at ICML*, Vol. 2014. 1–7.

[12] Noémi Bontridder and Yves Poullet. 2021. The role of artificial intelligence in disinformation. *Data amp; Policy* 3 (2021), e32. https://doi.org/10.1017/dap.2021.20

[13] Tom Bourgeade. 2022. *From Text to Trust: A Priori Interpretability Versus Post Hoc Explainability in Natural Language Processing.* Ph. D. Dissertation. Université Paul Sabatier-Toulouse III.

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

[15] 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. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[16] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces.* 454–464.

[17] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[18] Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the Independence of Association Bias and Empirical Fairness in Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* 370–378.

[19] Beatriz Cabrero-Daniel, Tomas Herda, Victoria Pichler, and Martin Eder. 2024. Exploring Human-AI Collaboration in Agile: Customised LLM Meeting Assistants. *arXiv preprint arXiv:2404.14871* (2024).

[20] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces.* 258–262.

[21] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[22] 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.

[23] 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.

[24] Neha Chacko and Viju Chacko. 2023. Paradigm shift presented by Large Language Models (LLM) in Deep Learning. *ADVANCES IN EMERGING COMPUTING TECHNOLOGIES* (2023), 40.

[25] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).

[26] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2022. Machine explanations and human understanding. *arXiv preprint arXiv:2202.04092* (2022).

[27] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 348, 18 pages. https://doi.org/10.1145/3544548.3581015

[28] Uthsav Chitra and Christopher Musco. 2020. Analyzing the Impact of Filter Bubbles on Social Network Polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 115–123. https://doi.org/10.1145/3336191.3371825

[29] Ke-Li Chiu and Rohan Alexander. 2021. Detecting Hate Speech with GPT-3. *CoRR* abs/2103.12407 (2021). arXiv:2103.12407 https://arxiv.org/abs/2103.12407

[30] Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, Responsiveness, and Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 420, 18 pages. https://doi.org/10.1145/3491102.3517500

[31] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

[32] Leonardo De Cosmo. 2022. Google Engineer claims AI chatbot is sentient: Why that matters. https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

[33] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. *arXiv preprint arXiv:2010.00711* (2020).

[34] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages. https://doi.org/10.1145/3544548.3580672

[35] Teresa Datta and John P Dickerson. 2023. Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook. *arXiv preprint arXiv:2303.06223* (2023).

[36] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

[37] Maartje de Graaf and Bertram F. Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *2017 AAAI Fall Symposia*.

[38] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 862–872.

[39] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture* 25 (2021), 700–732.

[40] 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

[41] Robert S Dooley and Gerald E Fryxell. 1999. Attaining decision quality and commitment from dissent: The moderating effects of loyalty and competence in strategic decision-making teams. *Academy of Management journal* 42, 4 (1999), 389–402.

[42] Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st international conference on human-computer interaction with mobile devices and services*. 1–12.

[43] Senjuti Dutta, Sid Mittal, Sherol Chen, Deepak Ramachandran, Ravi Rajakumar, Ian Kivlichan, Sunny Mak, Alena Butryna, and Praveen Paritosh. 2023. Modeling subjectivity (by Mimicking Annotator Annotation) in toxic comment identification across diverse communities. *arXiv preprint arXiv:2311.00203* (2023).

[44] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312787

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

[46] Mingming Fan, Xianyou Yang, TszTung Yu, Q. Vera Liao, and Jian Zhao. 2022. Human-AI Collaboration for UX Evaluation: Effects of Explanation and Synchronization. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 96 (apr 2022), 32 pages. https://doi.org/10.1145/3512943

[47] Thomas Fel, David Vigouroux, Rémi Cadène, and Thomas Serre. 2022. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 720–730.

[48] Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. 2024. Large Language Model-based Human-Agent Collaboration for Complex Task Solving. *arXiv preprint arXiv:2402.12914* (2024).

[49] Sharon Ferguson, Paula Akemi Aoyagui, Young-Ho Kim, and Anastasia Kuzminykh. 2024. Just Like Me: The Role of Opinions and Personal Experiences in The Perception of Explanations in Subjective Decision-Making. *arXiv preprint arXiv:2404.12558* (2024).

[50] Sharon A 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*. 1–7.

[51] Andrea Ferrario, Michele Loi, and Eleonora Viganò. 2020. In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology* 33, 3 (2020), 523–539.

[52] Jessica L. Feuston and Jed R. Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 469 (oct 2021), 25 pages. https://doi.org/10.1145/3479856

[53] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80, S1 (2016), 298–320.

[54] Raymond Fok and Daniel S Weld. 2023. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *arXiv preprint arXiv:2305.07722* (2023).

[55] Deborah H Francis and William R Sandberg. 2000. Friendship within entrepreneurial teams and its association with team and venture performance. *Entrepreneurship Theory and Practice* 25, 2 (2000), 5–26.

[56] Masaru Fuji, Katsuhito Nakazawa, and Hiroaki Yoshida. 2020. "Trustworthy and Explainable AI" Achieved Through Knowledge Graphs and Social Implementation. *Fujitsu Scientific & Technical Journal* 56, 1 (2020), 39–45.

[57] Cary Funk. 2023. How americans view emerging uses of artificial intelligence, including programs to generate text or art. https://www.pewresearch.org/short-reads/2023/02/22/how-americans-view-emerging-uses-of-artificial-intelligence-including-programs-to-generate-text-or-art/

[58] Cristina Garbacea, Samuel Carton, Shiyan Yan, and Qiaozhu Mei. 2019. Judge the judges: A large-scale evaluation study of neural language models for online review generation. *arXiv preprint arXiv:1901.00398* (2019).

[59] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *Comput. Surveys* 55, 13s (2023), 1–32.

[60] Maliheh Ghajargar, Jeffrey Bardzell, Alison Marie Smith-Renner, Kristina Höök, and Peter Gall Krogh. 2022. Graspable AI: Physical forms as explanation modality for explainable AI. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 1–4.

[61] Katherine J Hall. 2016. *" They believe that because they are women, it should be easier for them." Subtle and Overt Sexism toward Women in STEM from Social Media Commentary*. Virginia Commonwealth University.

[62] Ronan Hamon, Henrik Junklewitz, Gianclaudio Malgieri, Paul De Hert, Laurent Beslay, and Ignacio Sanchez. 2021. Impossible Explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 549–559.

[63] 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.

[64] Allyson I Hauptman, Wen Duan, and Nathan J Mcneese. 2022. The Components of Trust for Collaborating With AI Colleagues. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*. 72–75.

[65] Yugo Hayashi and Kosuke Wakabayashi. 2017. Can AI become reliable source to support human decision making in a court scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 195–198.

[66] Will Douglas Heaven. 2023. Large language models aren't people. let's stop testing them as if they were. https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/

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

[68] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 453–463. https://doi.org/10.1145/3581641.3584052

[69] Lisa Anne Hendricks, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Zeynep Akata. 2021. Generating visual explanations with natural language. *Applied AI Letters* 2, 4 (2021), e55.

[70] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.

[71] Danula Hettiachchi and Jorge Goncalves. 2019. Towards effective crowd-powered online content moderation. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 342–346.

[72] Jimpei Hitsuwari, Yoshiyuki Ueda, Woojin Yun, and Michio Nomura. 2023. Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior* 139 (2023), 107502.

[73] Robert Hoffman, Shane Mueller, Gary Klein, and Jordan Litman. 2021. Measuring trust in the XAI context. (2021).

[74] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[75] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674* (2023).

[76] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[77] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120.

[78] Jun Li Jeung and Janet Yi-Ching Huang. 2023. Correct Me If I Am Wrong: Exploring How AI Outputs Affect User Perception and Trust. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 323–327.

[79] David S Johnson, Olya Hakobyan, and Hanna Drimalla. 2023. Towards Interpretability in Audio and Visual Affective Machine Learning: A Review. *arXiv preprint arXiv:2306.08933* (2023).

[80] Ju Yeon Jung, Tom Steinberger, and Chaehan So. 2023. Towards Actionable Data Science: Domain Experts as End-Users of Data Science Systems. *Computer Supported Cooperative Work (CSCW)* (2023), 1–45.

[81] Amruta Kale, Tin Nguyen, Frederick C Harris Jr, Chenhao Li, Jiyin Zhang, and Xiaogang Ma. 2023. Provenance documentation to enable explainable and trustworthy AI: A literature review. *Data Intelligence* 5, 1 (2023), 139–162.

[82] Manveer Kalirai and Anastasia Kuzminykh. 2022. What Can You Do For Me? The Discoverability of Intelligent Assistant Skills. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*. 57–59.

[83] Katarina Kertysova. 2018. Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights* 29, 1-4 (2018), 55–81.

[84] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921* (2020).

[85] 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.

[86] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 designing interactive systems conference*. 881–894.

[87] Hadas Kotek, Rikker Dockum, and David Q Sun. 2023. Gender bias and stereotypes in Large Language Models. *arXiv preprint arXiv:2308.14921* (2023).

[88] Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science* 9, 1 (2022), 104–117.

[89] Swarn Avinash Kumar, Moustafa M Nasralla, Iván García-Magariño, and Harsh Kumar. 2021. A machine-learning scraping tool for data fusion in the analysis of sentiments about pandemics for supporting business decisions with human-centric AI explanations. *PeerJ Computer Science* 7 (2021), e713.

[90] Johannes Kunkel, Tim Donkers, Catalin-Mihai Barbu, and Jürgen Ziegler. 2018. Trust-related Effects of Expertise and Similarity Cues in Human-Generated Recommendations.. In *IUI Workshops*.

[91] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[92] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 54, 18 pages. https://doi.org/10.1145/3491102.3501999

[93] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. (2023), 1369–1385.

[94] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective explanations: Leveraging human input to align explainable ai. *arXiv preprint arXiv:2301.09656* (2023).

[95] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[96] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[97] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941* (2023).

[98] 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.

[99] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[100] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

[101] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2024. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. *arXiv preprint arXiv:2403.16812* (2024).

[102] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *Proceedings of the ACM on Human-Computer Interaction-CSCW* 3, CSCW (2019), 1–21.

[103] Daniel McDuff and Mary Czerwinski. 2018. Designing emotionally sentient agents. *Commun. ACM* 61, 12 (2018), 74–83.

[104] Elizabeth J Miller, Ben A Steward, Zak Witkower, Clare AM Sutherland, Eva G Krumhuber, and Amy Dawel. 2023. AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science* 34, 12 (2023), 1390–1403.

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

[106] Tim Miller. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 333–342.

[107] 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.

[108] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[109] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[110] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[111] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the Impact of Human Explanation Strategies on Human-AI Visual Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–37.

[112] Mohammad Naiseh, Nan Jiang, Jianbing Ma, and Raian Ali. 2020. Explainable recommendations in intelligent systems: delivery methods, modalities and risks. In *Research Challenges in Information Science: 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23–25, 2020, Proceedings 14*. Springer, 212–228.

[113] David T Newman, Nathanael J Fast, and Derek J Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160 (2020), 149–167.

[114] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[115] Nicole Nisbett and Viktoria Spaiser. 2023. How convincing are AI-generated moral arguments for climate action? *Frontiers in Climate* 5 (2023), 1193350.

[116] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[117] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture* 25, 2 (2021), 700–732.

[118] Alexis Palmer and Arthur Spirling. 2023. *Large Language Models Can Argue in Convincing and Novel Ways About Politics: Evidence from Experiments and Human Judgement.* Technical Report.

[119] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[120] Mike C Parent, Teresa D Gobble, and Aaron Rochlen. 2019. Social media behavior, toxic masculinity, and depression. *Psychology of Men & Masculinities* 20, 3 (2019), 277.

[121] Patrick philips2011, Max Scphilips2011, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* (2021), 78.

[122] Julia R. Fernandez. 2021. *"Being Yourself" Online: Supporting Authenticity for LGBTQ+ Social Media Users.* Association for Computing Machinery, New York, NY, USA, 249–252. https://doi.org/10.1145/3462204.3481786

[123] Stephen J Read and Amy Marcus-Newhall. 1993. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology* 65, 3 (1993), 429.

[124] Cheng Ren, Zachary Pardos, and Zhi Li. 2024. Human-AI Collaboration Increases Skill Tagging Speed but Degrades Accuracy. *arXiv preprint arXiv:2403.02259* (2024).

[125] Thomson Reuters. 2023. US lawmaker urges labelling, restrictions on AI content. https://www.reuters.com/technology/us-lawmaker-urges-labelling-restrictions-ai-content-2023-06-29/

[126] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[127] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization.* 223–233.

[128] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access* 8 (2020), 219563–219576.

[129] Chris Russell, Rory Mc Grath, and Luca Costabello. 2020. Learning Relevant Explanations.

[130] Christine Rzepka, Benedikt Berger, and Thomas Hess. 2022. Voice assistant vs. Chatbot–examining the fit between conversational agents' interaction modalities and information search tasks. *Information Systems Frontiers* 24, 3 (2022), 839–856.

[131] Mike Schaekermann. 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

[132] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware ai assistants for medical data analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* 1–14.

[133] Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 13 (jan 2022), 29 pages. https://doi.org/10.1145/3492832

[134] Björn W Schuller, Tuomas Virtanen, Maria Riveiro, Georgios Rizos, Jing Han, Annamaria Mesaros, and Konstantinos Drossos. 2021. Towards sonification in multimodal and user-friendlyexplainable artificial intelligence. In *Proceedings of the 2021 International Conference on Multimodal Interaction.* 788–792.

[135] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Comput. Surv.* 54, 4, Article 81 (may 2021), 43 pages. https://doi.org/10.1145/3386867

[136] Emre Sezgin, Joseph Sirrianni, and Simon L Linwood. 2022. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. *JMIR Med Inform* 10, 2 (10 Feb 2022), e32875. https://doi.org/10.2196/32875

[137] Murray Shanahan. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551* (2022).

[138] Matthew Shardlow and Piotr Przybyła. 2022. Deanthropomorphising NLP: Can a Language Model Be Conscious? *arXiv preprint arXiv:2211.11483* (2022).

[139] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.

[140] Ashish Sharma, Sudha Rao, Chris Brockett, Akanksha Malhotra, Nebojsa Jojic, and Bill Dolan. 2024. Investigating Agency of LLMs in Human-AI Collaboration Tasks. In *Proceedings of the 18th Conference of the European Chapter*

of the Association for Computational Linguistics (Volume 1: Long Papers), Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1968–1987. https://aclanthology.org/2024.eacl-long.119

[141] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 384–387.

[142] Nina Svenningsson and Montathar Faraon. 2019. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*. 151–161.

[143] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.

[144] Hiroki Tanaka, Sakti Sakriani, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2016. Teaching Social Communication Skills Through Human-Agent Interaction. *ACM Trans. Interact. Intell. Syst.* 6, 2, Article 18 (aug 2016), 26 pages. https://doi.org/10.1145/2937757

[145] Daniel Vale, Ali El-Sharif, and Muhammed Ali. 2022. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics* 2, 4 (2022), 815–826.

[146] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[147] Vanshika Vats, Marzia Binta Nizam, Minghao Liu, Ziyuan Wang, Richard Ho, Mohnish Sai Prasad, Vincent Titterton, Sai Venkat Malreddy, Riya Aggarwal, Yanwen Xu, et al. 2024. A Survey on Human-AI Teaming with Large Pre-Trained Models. *arXiv preprint arXiv:2403.04931* (2024).

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

[149] 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.

[150] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 758, 19 pages. https://doi.org/10.1145/3544548.3581366

[151] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[152] 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

[153] Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the fourth workshop on online abuse and harms*. 54–64.

[154] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing Human-AI Collaboration for Generating Free-Text Explanations. arXiv:2112.08674

[155] Kyle Wiggers. 2023. Researchers discover a way to make CHATGPT consistently toxic. https://techcrunch.com/2023/04/12/researchers-discover-a-way-to-make-chatgpt-consistently-toxic/

[156] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[157] Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky, and David G. Rang. 2023. *Labeling AI-Generated Content: Promises, Perils, and Future Directions*. https://computing.mit.edu/wp-content/uploads/2023/11/AI-Policy_Labeling.pdf

[158] Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[159] Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*. 6032–6048.

[160] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 246 (jan 2021), 25 pages. https://doi.org/10.1145/3432945

[161] 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.

[162] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. 2019. Effects of influence on user trust in predictive decision making. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[163] Qingxiaoyang Zhu and Hao-Chuan Wang. 2023. Leveraging Large Language Model as Support for Human Problem Solving: An Exploration of Its Appropriation and Impact. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 333–337.

A APPENDIX A

Table 7. Table of codes used in the study with definitions and examples

Characteristic				**Definition**	**Example**
Tone	Emotionality	Emotional		Emotions or an emotional reaction within the explanation	*Because of the level of passion or anger depending how you read it...*
		Logical		A lack of emotions or reactions; containing an objective, logical stance	*There's just something about the explanation that comes off as mechanical, very, very logical..*
	Formality	Formal/ Academic		Formal or academic language used in an explanation	*..it's again too academic for me because it's cites [the] dictionary..*
		Informal/ Conversational		Conversational, causal, or informal language used in an explanation	*..the way it's written sounds like the way someone would speak...*
	Neutrality	Neutral		A neutral stance, or lack of definitive assessment (sexist or not) in the explanation	*..it starts out with, it might be used in a different context now. So it's trying to be very unbiased..*
		Opinionated		The explanation taking a side; making a definitive choice on the assessment	*..I feel like for me the explanation is a little bit too defensive for me to agree with it..*
		Ambivalent/ Contradictory		Not being able to make a definitive decision	*.. there's this uncertainty. It starts with, I'm not sure...*
Grammatical Elements	Length	Short		The length of the explanation being short	*..how short it is and how, in such a short number of words, it can pretty much encompass the feeling..*
		Long		The length of the explanation being long	*..I think it's a little bit of a long explanation..*
		Repetition		The length of the explanation being long due to repetition of parts	*..almost repeating the sentences, repeating what was happening instead of jumping directly to the argument.*
	Grammar	Double Negatives		The presence of double negatives in an explanation	*..that double negative, it really hurts my brain..*
		Correct		The use of correct grammar in an explanation	*..Its just, it is in a very perfect English..*
		Punctuation		The presence, or lack of, punctuation in an explanation	*Or no commas.. I think AI would put some commas there...*
		Sentence Structure		The sentence structure in an explanation	*..it's kind of a little bit complicated as a sentence structure cuz you are using a negative..and then a positive attitude..*
	Pronoun Type	Personal		The use of personal pronouns (I, she, her) in an explanation	*..it's very personal, there's lots of I's..*
		Impersonal		The use of impersonal pronouns (it) in an explanation	*.. yeah, it starts with a lot of "it", like "it is" or "it is not"..*

Table 7 – *Continued from previous page*

Characteristic				**Definition**	**Example**
Argumentative Sophistication	Source of Argument	Definition		The presence of a definition in an explanation	*..They looked up hysterical in Webster's dictionary...*
		Outcome		The focus on an outcome of an action in an explanation	*..think about the long term implications of the scenario at hand..*
		Example	External	The presence of a non-personal (external) example in an explanation	*..Because it does give good examples of work house chores and stuff..*
			Personal Experience	The presence of an example containing personal experience in an explanation	*..It gave me a feeling that it came from personal experience..*
			Counterfactual	The presence of a counterfactual example in an explanation	*.. providing an opposite example and how the opposite audience would not enjoy this type of treatment..*
		Opinion		The presence of subjectivity, or one's opinion or stance, in an explanation	*..almost like a hint of a strong opinion coming from the explanations..*
		Authority		The authority of the explanation author (from experience or credentials)	*.. I would believe that it's more trustworthy from an authority or someone that understands..the subject ..*
		Fact/Proof		The presence of lack of facts, statistics, citations, or general proof in an explanation	*..I think of science again, facts. So there are no facts..*
		Similarity	Online Content	How an explanation is similar to online content (such as Reddit)	*..It sounds like a response on a forum..*
			Humans	How an explanation is similar to human composed content.	*..I think if humans can be predictable then machines can have easily picked up on their predictability..*
	Argument Structure	Formality		A formal explanation structure, or a set order of explanation components	*..it's about how it was structured..kind of introducing the scenario, then explaining it and then providing an example..*
		Directness	Direct	A direct style in the explanation, getting straight to the point	*..but the way the explanation goes is straight to why it is sexist..*
			Verbose	An elaborate or verbose style in the explanation, using many words to make a point	*..it seems like a little verbose for a human to respond to it with..*
		Completeness	Completeness of Argument Consideration	The depth, or breadth, of the argument; whether all possible explanations were considered	*..I feel like it's missing. It's not addressing each and every argument..*
	Relation to Scenario	Relation to Scenario		How much the explanation relates to the scenario; how much of the scenario is included in the explanation	*..it's actually addressing the concerns for the points on the scenario, because it's talking about basically the entire biology..*
	Context Consideration	Contextual Understanding		Whether the explanation shows some contextual understanding, generalizing the scenario to broader societal context	*.. I don't think it comes from someone with a deep understanding of why gender chores would be problematic..*

Table 7 – *Continued from previous page*

Characteristic				Definition	Example
Relation	Relation to User	Alignment with Personal Experience		Whether the experience shown in the explanation matches their personal experience	*..Yeah, the explanation felt relatable cuz as a woman in stem, that's something like I've experienced back in high school..*
		Alignment with Personal Stance		Whether the stance or opinion in the explanation matches their personal stance or opinion	*..I think what this explanation is saying is correct, and that is what I believe in and that is why I'm using the word correct because I agree with it..*

Received January 2024; revised April 2024; accepted May 2024