Unveiling Human-AI Interaction and Subjective Perceptions About Artificial Intelligent Agents

Mathyas Giudici(✉), Federica Liguori, Andrea Tocchetti, and Marco Brambilla

Politecnico di Milano, DEIB, 20133 Milano, Italy
{mathyas.giudici,federica.liguori,andrea.tocchetti, marco.brambilla}@polimi.it

Abstract. This work focuses on human-AI interactions, employing a crowd-based methodology to collect and assess the reactions and perceptions of a human audience to a dialogue between a human and an artificial intelligent agent. The study is conducted through a live streaming platform where human streamers broadcast interviews to a custom-made GPT voice interface. The questions extracted from the dialogues were categorized based on emotional and cognitive criteria. Our method covers thematic, emotional, and sentiment analyses of the comments platform users shared during the interview. This work aims to contribute to Human-Computer Interaction (HCI) and Human-Centered AI, emphasizing the need for a paradigm shift in AI research from focusing on technological development to considering its impact on human beings.

Keywords: Artificial Intelligence · Human-AI Interaction · Human-Centered AI · Crowdsourcing · Human-Computer Interaction

1 Introduction

Recent artificial intelligence (AI) developments influence people's work and daily lives. However, the development of AI systems has been predominantly driven by a technology-centered design approach. Recently, AI development has taken a broader perspective on the problem: technological enhancement meets ethical and human factors design. *Human-centered AI* [5] embodies an approach where AI and machine learning systems are designed with a keen awareness that they are part of a broader context, including various stakeholders and focusing on ensuring fairness, maintaining accountability, enhancing interpretability, and upholding transparency. *Human factors design* is crucial in ensuring that AI solutions are explainable, understandable, useful, and usable to humans, also considering Human-Computer Interaction, user perception, and technology acceptance [1]. More recent research has broadened the scope by introducing trust, encompassing *cognitive* and *emotional* (support) aspects, thus enabling broader user engagement analyses [4]. Crowdsourcing approaches have emerged as a valuable tool for gathering human knowledge relevant to technology development in

K. Stefanidis et al. (Eds.): ICWE 2024, LNCS 14629, pp. 414–418, 2024.
https://doi.org/10.1007/978-3-031-62362-2_36

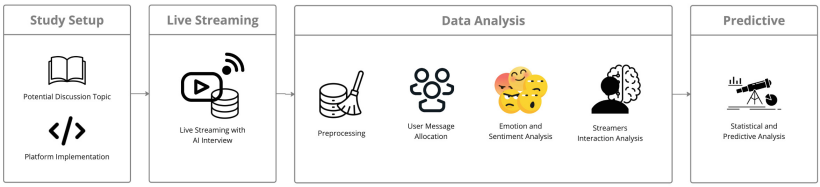

Fig. 1. Experimental Approach

AI interpretability and explainability [2], also using gamification techniques to gain enhanced user participation [6]. This work evaluates AI models from a human-centered perspective, investigating the emotions of humans towards AI as a proxy to emotional trust in AI models [3]. Pursuing such an objective, we developed a new approach using an online streaming platform, where live sessions in an interview style were broadcast by volunteer human streamers who asked questions to an AI generative model (GPT). We analyzed the audience's reaction using a crowdsourcing method to gather data from user comments posted on the platform in real time as responses to the ongoing interviews. The interviews focused on thoughtfully crafted questions derived from the fundamental notion of trust. Our study, therefore, aims to better understand human factors in human-AI interactions by directly observing users in an interactive digital setting.

2 Method

Our research evaluates the user's interaction with responses to the generative AI model OpenAI GPT3.5. Volunteer streamers from a popular streaming platform interviewed the AI model and involved in real time the audience from their community in the interaction. The communication between the streamers and the AI was oral (without any visual clues or avatar representing the AI), using a web application delivering Speech-To-Text and Text-To-Speech services between streamers and the API of the model. At the same time, the audience comments were typed into the chat UI. Our system recorded interview utterances and live-users comments. Figure 1 shows the phases of the approach:

Setup. Streamers were selected, and a set of potential discussion topics was provided to them. The proposed questions were categorized into two groups: (i) *Cognitive* - highlighting the AI's capabilities and skills, the tasks and how well it can perform them; (ii) *Emotional* - stressing ethical issues, abstract reasoning capabilities, or subjects requiring the AI to take a stance.

Preprocessing. All the texts (from streamer-AI interviews and users' live comments) were translated from Italian to English using DeepL APIs and the data was properly formatted.

User Message Allocation. Live sessions were segmented into 15-second slots timeframes to gain a more detailed overview of the temporal distribution of

Table 1. Temporal Analysis Metrics (left) and Strong Reactions Summary (right)

Metric	Value
Min AI Response Time	1.568 s
Max AI Response Time	28.313 s
Avg AI Response Time	12.14 s
Comments before AI Response	10.9%
Comments after AI Response	89.1%

Metric	Count
Total Strong Reactions	1297
> Positive Reactions (PR)	853
> Negative Reactions (NR)	444
PR Unique Comments	459
NR Unique Comments	124
PN- NR Unique Comments	226

the collected messages and then obtain a proper association between the AI's responses and user comments.

Emotion Analysis. Emotion of users' comments was analyzed using a pre-trained model[1] based on the 'roberta-base' model and trained for multi-label classification on 28 emotion labels (along with their score). We selected only the *strong reactions* (i.e., with a minimum score of 0.1), excluding the *neutral* label.

Sentiment Analysis. User sentiment was assessed through the 'twitter-roberta-base-sentiment'[2] model. The output of the model for each user message consisted of *neutral*, *positive*, and *negative* labels and their corresponding scores.

Streamers' Interaction Analysis: A topic analysis on the interviews' content was performed. The first part of the process entailed classifying the interactions into cognitive or emotional categories (as defined above). In the second part, we implemented keyword extraction and clustering methods using BERT embeddings and LDA method (although the latter didn't prove useful) (Table 1).

3 Results and Discussion

User Messages Analysis. The results showed that the majority of comments (57.7%) held a neutral sentiment, followed by positive comments (21.6%), which were slightly more prevalent than negative ones (20.7%). Instead, the emotions extracted were categorized into positive or negative (see Supplementary Material[3]). We observed that the number of positive reactions was almost double that of negative reactions. The predominant positive emotions identified from users' comments were curiosity, approval, and amusement, while the strong negative reactions were predominantly linked to emotions of confusion, disapproval, and annoyance. Such results suggest a positive and curious reaction in the online streaming community when the streamer interacts with AI.

[1] huggingface.co/SamLowe/roberta-base-go_emotions.
[2] huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment.
[3] Supplementary Material: 10.5281/zenodo.10819620.

Streamer Interactions Analysis. This phase involved analyzing the topics touched upon in streamers' interviews. The first part entailed classifying the 311 interactions into cognitive or emotional categories, finally labeling 46.2% of the interactions as Emotional and 53.8% as Cognitive. A comprehensive description of the touched topics is provided in the Supplementary Material (See footnote 3).

Statistical and Predictive Analysis. Statistical metrics show that cognitive ($M = 0.428$, $SD = 0.271$) and emotional ($M = 0.421$, $SD = 0.273$) interactions generated similar emotional intensity in user comments. The number of generated strong reactions was also similar in cognitive (M=8.05, SD=9.51) and emotional ($M = 7.30$, $SD = 8.34$) interactions, with a slightly higher mean - but not statistically significant difference - in the case of cognitive ones. Such results allow us to speculate that a novel AI model does not generate dominant reactions in the context of a live-streaming community. Finally, we conducted a predictive statistical analysis using linear regression on the number of high reactions for each interaction (80% train - 20% test). The results indicated that the number of comments was the only significant parameter (with a p-value < 0.001) explaining the predicted variable, while the topic discussed in the interview does not influence the prediction of reactions.

4 Conclusion and Future Works

We presented and evaluated a potential method for analyzing user reactions to AI using a crowd-based approach in a live-streaming setting, emphasizing the importance of collecting such reactions in a context as unbiased as possible (thus avoiding explicit requests or polls by researchers). We showed the advantage of using a new practice to grab spontaneous reactions and emotions of human-AI interaction, paving the way for future research in the field. Still, the usage of a live streaming platform allowed a diverse range of users to participate in the study, offering a familiar environment in which users can interact. However, the lack of a large and normalized dataset limited the application of advanced statistical and predictive methods able to elicit complex patterns, which would have been beneficial in understanding more of the ways users perceive and interact with AI. Future work may aim to overcome these limitations by: (1) conducting more live-streaming (or podcast) sessions based on English communities to reach a broader target audience; (2) include users from diverse cultural backgrounds; and (3) testing and comparing the reactions other AI models.

References

1. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q., 319–340 (1989)
2. Estellés-Arolas, E., González-Ladrón-de Guevara, F.: Towards an integrated crowdsourcing definition. J. Inf. Sci. **38**(2), 189–200 (2012)
3. Glikson, E., Woolley, A.W.: Human trust in artificial intelligence: review of empirical research. Acad. Manag. Ann. **14**(2), 627–660 (2020)
4. McAllister, D.J.: Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. Acad. Manag. J. **38**(1), 24–59 (1995)
5. Riedl, M.O.: Human-centered artificial intelligence and machine learning. Hum. Behav. Emerg. Technol. **1**(1), 33–36 (2019)
6. Tocchetti, A., Corti, L., Brambilla, M., Celino, I.: EXP-crowd: a gamified crowdsourcing framework for explainability. Front. Artif. Intell. **5**, 826499 (2022)