http://dx.doi.org/10.14236/ewic/BCSHCI2025.52

Human-AI Collaboration in High-Stakes Decision-Making: Work in Progress

Jianan Zhou
Dyson School of Design Engineering
Imperial College London
London, UK
jianan.zhou22@imperial.ac.uk

Ranya Aloufi
Dyson School of Design Engineering
Imperial College London
London, UK
r.aloufi18@imperial.ac.uk

Talya Porat
Dyson School of Design Engineering
Imperial College London
London, UK
t.porat@imperial.ac.uk

Nejra van Zalk
Dyson School of Design Engineering
Imperial College London
London, UK
n.van-zalk@imperial.ac.uk

Abstract: This work in progress investigates human interaction with an LLM-powered chatbot, presented as either a fellow human or a transparently disclosed AI collaborator, in a high-stakes decision-making simulation—the NASA Moon Survival Task. We will employ a one-way between-subjects design to examine how individuals' collaboration and communication are influenced by the identity of their partner (AI vs. human). Specifically, we will evaluate individuals' collaboration processes (i.e., collaborative behaviour and communicative dynamics) and outcomes, alongside their retrospective interaction experience and perceptions of the partner. We will also examine dyadic-level linguistic coordination during the interaction and conduct user profiling to uncover variations in AI collaborative benefits. We anticipate that this study will have four key impacts: safeguarding human-AI collaboration, democratising AI benefits, guiding model improvement, and making methodological contributions. The anonymised dialogues and associated data will be open-sourced upon study completion.

Keywords: Human-AI Interaction, Collaborative Decision-Making, Large Language Model, AI Identity, Linguistic Coordination, User Profiling, Moon Survival Task

1. INTRODUCTION

Interactive intelligent systems (Jameson and Riedl 2011), such as chatbots, virtual humans, and robots, are increasingly integrated into society. Recent breakthroughs in AI have driven their deployment beyond low-stakes, casual settings into scenarios where tasks become more consequential and human engagement is more deliberate (i.e., requiring significant cognitive effort; Evans 2019). Prior research shows that individuals intuitively treat AI partners as social beings, responding to AI partners similarly as they would to fellow humans (Zhou et al. 2024). This was observed using Cyberball, a well-known experimental paradigm researching social ostracism in a scripted online ball-tossing game (Williams and Jarvis 2006). Participants compensated an ostracised AI partner by tossing the ball to them more frequently, mirroring typical responses to an ostracised human. However, ball-tossing behaviour during Cyberball is intuitive and effortless (Zhou et al. 2024). This leaves open the question of how participants respond to AI when engaging more deliberately (i.e., using cognitive effort to reflect on the interaction and their partner, weigh alternatives, and coordinate actions towards shared goals). It is crucial to understand whether individuals still interact with AI similarly to human partners in such demanding scenarios—for example, when collaborating on decision-making with high stakes.

AI systems, particularly those powered by LLMs and other deep neural networks, have evolved to show the technical capacity to act as collaborators, augmenting decision-making in high-stakes domains like healthcare and business (Zhang et al. 2024; Niraula et al. 2025; Hao et al. 2024; Ameen et al. 2025). These models can be adapted to diverse conditions using techniques such as fine-tuning and prompt engineering. For example, recent work has customised LLaMA-3.1-8B-Instruct via prompt

© Zhou et al. Published by BCS Learning and Development Ltd. Proceedings of BCS HCI 2025, UK

engineering, enabling it to collaborate with humans on survival tasks (Wang et al. 2025). LLMs have been demonstrated to approach or even exceed human-level performance in both knowledge-based tasks (Luo et al. 2024; Abdelsamie and Wang 2024) and situational judgment tasks (Mittelstädt et al. 2024). However, collaboration success depends not merely on technical performance but on the quality of interaction: respecting user agency, enabling mutual adaptation, and fostering trust (Puerta-Beldarrain et al. 2025).

Our recent meta-analysis found that individuals exhibit comparable behavioural trust, social alignment (i.e., the extent to which one aligns their internal states and behaviours with their partner), and task performance when interacting with performance-matched AI and human partners. Yet, individuals are less prosocial and moral when interacting with AI partners, and perceive them as having lower social traits (Zhou et al. 2025). These findings suggest identity-related differences independent of objective capability, necessitating empirical investigation into how AI identity influences human-AI collaboration.

To this end, the current study will investigate how AI identity influences behavioural and attitudinal dimensions of user collaboration and communication. We will evaluate individuals' collaboration processes and outcomes, as well as their retrospective interaction experience and perceptions of the partner. Furthermore, we will delve into the subtle, dynamic interplay between human and AI collaborators. We will analyse whether and how dyadic-level processes of mutual adaptation, such as linguistic coordination, evolve over the course of the interaction, and examine whether these adaptation cues predict collaboration outcomes. Linguistic coordination—the tendency for a dyad to coordinate their language style and semantic content to become more similar to each other during interaction—is central to achieving common ground and fostering the development of social relationships in human-human communication (Babcock et al. 2014; Müller-Frommeyer et al. 2019). We will thus examine whether similar adaptation processes occur in human-AI communication. Finally, we will profile users based on individual characteristics, including dispositional traits, AI usage experience, and demographics, and explore profile variations in AI collaborative benefits.

2. RESEARCH QUESTIONS

In this study, we will customise OpenAI's GPT-4 series models (e.g., GPT-4o and GPT-4.1) to simulate task-aligned, human-like collaborative behaviour. The optimally customised LLM will be configured to serve as the collaborative partner (framed as an AI or human; detailed in Methods) in the subsequent experiment. We will investigate how individuals interact with the LLM-powered chatbot, presented as either a fellow human or a transparently disclosed AI collaborator, in a high-stakes decision-making simulation (i.e., the NASA Moon Survival Task). We will address three main research questions:

- **RQ1:** Does the identity of the collaborative partner (AI vs. human) have an impact on collaboration processes, outcomes, and subjective experience?
- **RQ2:** How does linguistic coordination evolve over the course of interaction with AI vs. human partners, and does it predict collaboration outcomes?
- **RQ3:** Can distinct profiles be identified based on user characteristics, and how are these user profiles linked to collaboration with AI?

3. METHODS

The study will employ a one-way between-subjects design with two conditions, both involving interaction with the LLM-powered chatbot: **(1) Human partner condition:** participants will be led to believe that they are interacting with another human participant; **(2) AI partner condition:** participants will be told that they are interacting with an AI.

In the human partner condition, participants will be instructed before the task that their collaborative partner is another human participant logged on at the same time. They will be unaware they are interacting with an LLM-powered chatbot, leveraging deception and withholding-disclosure strategies validated in previous studies (e.g., Zhou et al. 2024; Yin et al. 2024; Russo et al. 2021).

3.1. Technical Implementation

Both chatbots will be built on the customised LLM and operate in text mode. We are currently employing prompt engineering for customisation. To ensure consistency between the two conditions, both chatbots will use identical dialogue structures and non-expert task knowledge, differing only in identity disclosure, as specified in the system prompts. They will demonstrate human-like dialogue acts, reasoning patterns, and imperfections throughout the interaction. This setup aims to isolate the effects of partner identity.

We have developed sophisticated system prompts to tailor GPT-4-series behaviour. Early evaluations by the research team indicate that this approach shows promise in achieving the desired levels of human-likeness and task alignment. However, if formal

pilot testing with human participants proves that prompting is insufficient or equivalence checks on condition consistency fail, we will fine-tune the model using existing human-human dialogue datasets from similar survival tasks (e.g., winter survival; Braley and Murray 2018; Freeman et al. 2014).

The chatbot frontend is designed within the Qualtrics platform, as shown in Figure 1. The overall chatbot architecture is presented in Figure 2. Incoming user messages with participant IDs are sent via HTTPS POST from the Qualtrics frontend to the Google Cloud Run backend. For each call, Cloud Run fetches the conversation history from Memorystore, combines it with the system prompt and new user input, and sends the assembled request to the customised LLM via OpenAI's Chat Completions API. The model's response is returned to Cloud Run, which then sends it to the Qualtrics frontend for display. Cloud Run also stores both the user input and model response in Memorystore to update conversation history for subsequent calls. The full dialogue and metadata (i.e., timestamps and IDs) are logged in Cloud SQL for further data analysis.

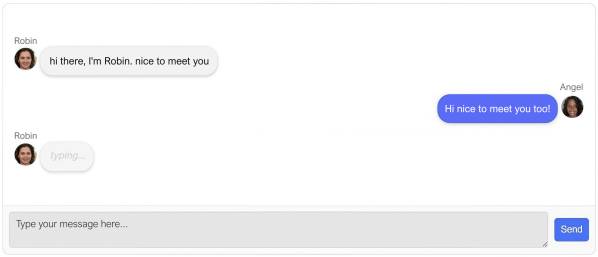


Figure 1: *Qualtrics Chat Interface. The chat interface features message displays, avatars, and pseudonyms to facilitate real-time collaboration. Avatars are selected by participants from a set of human faces generated by GANs, and pseudonyms are created by participants before the task to protect their privacy and increase engagement.*

3.2. Task and Procedure

The experiment will be administered online via Qualtrics, including a collaborative decision-making task—NASA Moon Survival Task (Hall and Watson 1970)—and associated questionnaires. The moon survival task challenges participants to rank 15 survival items in order of importance following a hypothetical spacecraft crash on the moon. This task and its variants, including winter survival and desert survival, are often used in psychological and organisational research to study group collaboration, interpersonal dynamics, and decision-making in simulated high-stakes scenarios.

Participants will first rank 15 survival items individually. They will then be randomly assigned to collaborate with either an AI or a purported human partner

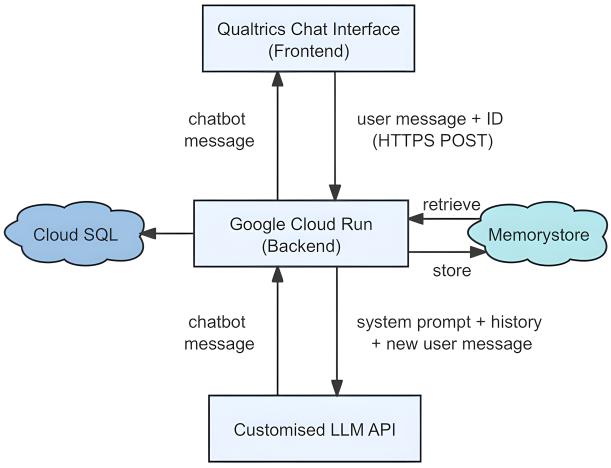


Figure 2: *Chatbot Architecture. Participants interact via a Qualtrics chat interface to complete the moon survival task with a customised LLM simulated as an AI or a purported human collaborator. Participants' messages are routed through the Cloud Run, which makes API calls to the LLM.*

on the moon survival task. Participants will engage in unstructured, text-based interaction with their partner to discuss rankings, explore each other's reasoning, and seek consensus where possible (full agreement is not required). After the interaction, participants will re-rank the items individually. Participants' collaboration outcomes will be assessed by pre-post improvement based on distance from the expert ranking and conformity to the partner's ranking. Collaboration processes will be assessed by coding both theory-driven collaborative behaviour (Liu et al. 2016) and communicative dynamics during the collaboration (Dowell et al. 2019).

Following the moon survival task, participants will report their interaction experience and perceptions of the partner. Specifically, subjective collaboration experience, perceived closeness, trust, and perceived warmth and competence of the partner will be assessed using validated scales (e.g., Aron et al. 1992) or items used in prior research (e.g., McKee et al. 2023). Participants' attention allocation during the interaction will be assessed via self-report items measuring how much attention they devoted to their partner and to the interaction itself, relative to solving the ranking task. A manipulation check will then confirm whether participants perceived their partner's identity as intended.

Moreover, participants will complete measures of user characteristics, including dispositional traits (i.e., Big Five traits and collaborative disposition), AI usage experience, and demographics. Finally, participants will complete an awareness check before being debriefed and thanked for participation. Figure 3 depicts the experimental procedure. All

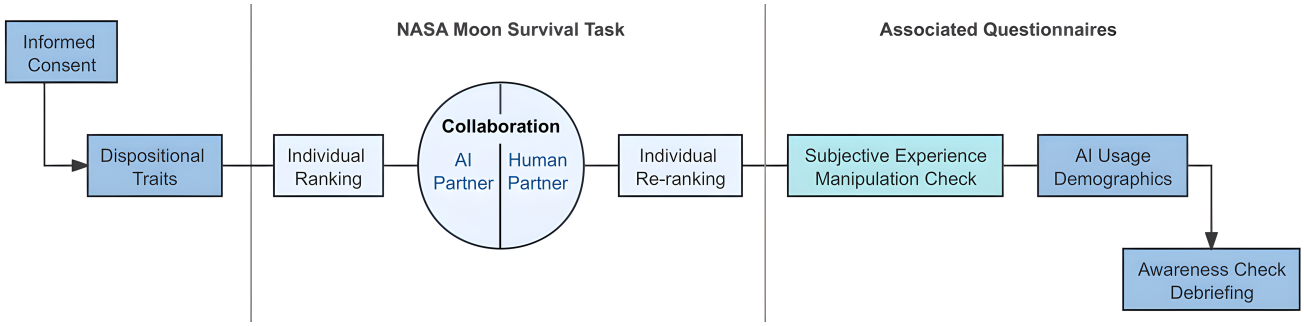


Figure 3: *Experimental Procedure*

procedures and study materials will adhere to the ethical principles outlined in the World Medical Association Declaration of Helsinki, with institutional ethical approval obtained prior to data collection.

3.3. Data Collection and Analysis

Participants will be recruited online via Prolific (2024), targeting those aged 18 and over whose primary language is English. The minimum unadjusted sample size for a one-way between-subjects design (partner identity: AI vs. human) is 200 to detect an effect size of d = 0.40 with α = 0.05 and power = 80% (Brysbaert 2019). The adoption of this small-to-medium effect size is based on two primary reasons. First, the absence of comparable prior research limits our ability to estimate the expected effect size needed to ensure adequate statistical power in this study. Second, d = 0.40 is considered an average effect size observed in experimental psychology-related research and holds both practical and theoretical relevance (Brysbaert 2019; Richard et al. 2003). To account for potential participant attrition, we plan to over-recruit to N = 300.

Prior to statistical testing, the data will be checked for assumptions (e.g., normality, homoscedasticity) to ensure the appropriateness of the selected analyses. First, a series of ANOVAs, with Benjamini-Hochberg adjustments for multiple comparisons, will evaluate the effects of partner identity (AI vs. human) on participants' collaboration processes, outcomes, and subjective experience. Second, mixed-effects models will evaluate how linguistic coordination evolves over the course of interaction with AI vs. human partners, followed by testing whether it predicts collaboration outcomes. Language style matching and latent semantic similarity will serve as key coordination metrics (Babcock et al. 2014). The former metric will be calculated as dyads' similarity in the use of function words (Ireland and Pennebaker 2010). The latter will be calculated using sentence-transformer embeddings with cosine similarity (Yu et al. 2025). Potential confounds like message length will be tested and, if warranted, adjusted for as covariates. Third, latent profile analyses will identify distinct profiles based on a range of user characteristics, with further analyses on how these user profiles link to collaboration with AI. Finally, the dialogues and associated data (with participants' consent) will be anonymised and released as open source upon study completion.

4. EXPECTED IMPACT

Safeguarding human-AI collaboration: This study will contribute to quantifying the risks of automation bias and/or algorithmic aversion by identifying how AI identity influences behavioural and attitudinal dimensions of user collaboration and communication. The findings will help shape AI deployment policies by balancing transparency and user acceptance in collaborative decision-making contexts.

Democratising AI benefits: The study will uncover variations in AI collaborative benefits by mapping distinct user profiles to collaboration processes and outcomes. The findings will inform interventions to advance the design of equitable AI systems that adapt to subgroup-specific needs.

Guiding model improvement: The study will reveal current LLM's capabilities and limitations in collaboration and yield actionable solutions to optimise model's ability to emulate human conversational pragmatics and enhance collaborative potentials.

Methodological contributions: The study will be the first to validate linguistic coordination metrics such as language style matching and latent semantic similarity as benchmarks for evaluating human-AI interaction. In addition, the study will establish frameworks for latent profiling of AI users to enable systematic analysis of how individual differences shape collaboration with AI.

5. ACKNOWLEDGMENT

This study has received GPT API credits from the OpenAI Researcher Access Program.

REFERENCES

Abdelsamie, M. and H. Wang (2024). Comparative Analysis of LLM-based Market Prediction and Human Expertise with Sentiment Analysis and Machine Learning Integration. In *2024 7th International Conference on Data Science and Information Technology (DSIT)*, Nanjing, China, pp. 1–6. IEEE.

Ameen, N., M. Pagani, E. Pantano, J. Cheah, S. Tarba, and S. Xia (2025). The Rise of Human–Machine Collaboration: Managers' Perceptions of Leveraging Artificial Intelligence for Enhanced B2B Service Recovery. *British Journal of Management 36*(1), 91–109.

Aron, A., E. N. Aron, and D. Smollan (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology 63*(4), 596–612.

Babcock, M. J., V. P. Ta, and W. Ickes (2014). Latent Semantic Similarity and Language Style Matching in Initial Dyadic Interactions. *Journal of Language and Social Psychology 33*(1), 78–88.

Braley, M. and G. Murray (2018). The Group Affect and Performance (GAP) Corpus. In *Proceedings of the Group Interaction Frontiers in Technology*, Boulder CO USA, pp. 1–9. ACM.

Brysbaert, M. (2019). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition 2*(1), 16.

Dowell, N. M. M., T. M. Nixon, and A. C. Graesser (2019). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods 51*(3), 1007–1041.

Evans, J. S. B. T. (2019). Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning 25*(4), 383–415.

Freeman, V., J. Chan, G.-A. Levow, R. Wright, M. Ostendorf, V. Zayats, Y. Luan, H. Morrison, L. Fox, M. Antoniak, and P. Parsons (2014). ATAROS Technical Report 1: Corpus collection and initial task validation. Technical Report, U. Washington Linguistic Phonetics Lab.

Hall, J. and W. H. Watson (1970). The Effects of a Normative Intervention on Group Decision-Making Performance. *Human Relations 23*(4), 299–317.

Hao, X., E. Demir, and D. Eyers (2024). Exploring collaborative decision-making: A quasi-experimental study of human and Generative AI interaction. *Technology in Society 78*, 102662.

Ireland, M. E. and J. W. Pennebaker (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology 99*(3), 549–571.

Jameson, A. and J. Riedl (2011). Introduction to the Transactions on Interactive Intelligent Systems. *ACM Transactions on Interactive Intelligent Systems 1*(1), 1–6.

Liu, L., J. Hao, A. von Davier, P. Kyllonen, and D. Zapata-Rivera (2016). A Tough Nut to Crack: Measuring Collaborative Problem Solving. In Y. Rosen, S. Ferrara, M. Mosharraf, and J. Keengwe (Eds.), *Handbook of Research on Technology Tools for Real-World Skill Development*:, Advances in Higher Education and Professional Development. IGI Global.

Luo, X., A. Rechardt, G. Sun, K. K. Nejad, F. Yáñez, B. Yilmaz, K. Lee, A. O. Cohen, V. Borghesani, A. Pashkov, D. Marinazzo, J. Nicholas, A. Salatiello, I. Sucholutsky, P. Minervini, S. Razavi, R. Rocca, E. Yusifov, T. Okalova, N. Gu, M. Ferianc, M. Khona, K. R. Patil, P.-S. Lee, R. Mata, N. E. Myers, J. K. Bizley, S. Musslick, I. P. Bilgin, G. Niso, J. M. Ales, M. Gaebler, N. A. Ratan Murty, L. Loued-Khenissi, A. Behler, C. M. Hall, J. Dafflon, S. D. Bao, and B. C. Love (2024). Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour 9*(2), 305–315.

McKee, K. R., X. Bai, and S. T. Fiske (2023). Humans perceive warmth and competence in artificial intelligence. *iScience 26*(8), 107256.

Mittelstädt, J. M., J. Maier, P. Goerke, F. Zinn, and M. Hermes (2024). Large language models can outperform humans in social situational judgments. *Scientific Reports 14*(1), 27449.

Müller-Frommeyer, L. C., N. A. M. Frommeyer, and S. Kauffeld (2019). Introducing rLSM: An integrated metric assessing temporal reciprocity in language style matching. *Behavior Research Methods 51*(3), 1343–1359. Publisher: Springer Science and Business Media LLC.

Niraula, D., K. C. Cuneo, I. D. Dinov, B. D. Gonzalez, J. B. Jamaluddin, J. J. Jin, Y. Luo, M. M. Matuszak, R. K. Ten Haken, A. K. Bryant, T. J. Dilling, M. P. Dykstra, J. M. Frakes, C. L. Liveringhouse, S. R. Miller, M. N. Mills, R. F.

Palm, S. N. Regan, A. Rishi, J. F. Torres-Roca, H.-H. M. Yu, and I. El Naqa (2025). Intricacies of human–AI interaction in dynamic decision-making for precision oncology. *Nature Communications 16*(1), 1138.

Prolific (2024). Prolific. `https://www.prolific.com`. Accessed: 2025-08-23.

Puerta-Beldarrain, M., O. Gómez-Carmona, R. Sánchez-Corcuera, D. Casado-Mansilla, D. López-de Ipiña, and L. Chen (2025). A Multifaceted Vision of the Human-AI Collaboration: A Comprehensive Review. *IEEE Access 13*, 29375–29405.

Richard, F. D., C. F. Bond, and J. J. Stokes-Zoota (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology 7*(4), 331–363.

Russo, P. A., M. Duradoni, and A. Guazzini (2021). How self-perceived reputation affects fairness towards humans and artificial intelligence. *Computers in Human Behavior 124*, 106920.

Wang, K., B.-K. Quek, J. Goh, and D. Herremans (2025). To Embody or Not: The Effect Of Embodiment On User Perception Of LLM-based Conversational Agents. Version Number: 1.

Williams, K. D. and B. Jarvis (2006, February). Cyberball: A program for use in research on interpersonal ostracism and acceptance. *Behavior Research Methods 38*(1), 174–180.

Yin, Y., N. Jia, and C. J. Wakslak (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences 121*(14), e2319112121.

Yu, C.-W., Y.-S. Chuang, A. N. Lotsos, T. Meier, and C. M. Haase (2025). The More Similar, the Better? Associations Between Latent Semantic Similarity and Emotional Experiences Differ Across Conversation Contexts. *Journal of Language and Social Psychology*, 0261927X251343096.

Zhang, S., J. Yu, X. Xu, C. Yin, Y. Lu, B. Yao, M. Tory, L. M. Padilla, J. Caterino, P. Zhang, and D. Wang (2024). Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA, pp. 1–18. ACM.

Zhou, J., T. Porat, F. Corbett, J. Byun, and N. van Zalk (2025). Psychological and behavioural responses in human-agent vs. human-human interactions: a systematic review and meta-analysis. Manuscript in preparation.

Zhou, J., T. Porat, and N. Van Zalk (2024). Humans Mindlessly Treat AI Virtual Agents as Social Beings, but This Tendency Diminishes Among the Young: Evidence From a Cyberball Experiment. *Human Behavior and Emerging Technologies 2024*(1), 8864909.