Ted-AI Talks: Challenging Chats with AI-Enabled Teddies

Pam Briggs, Matt Wood, Dan Jackson, and Kyle Montague
Northumbria University
Newcastle upon Tyne, UK
p.briggs@northumbria.ac.uk
matt.j.wood@northumbria.ac.uk
dan.g.jackson@northumbria.ac.uk
kyle.montague@northumbria.ac.uk

People are increasingly turning to AI agents for advice and support, but can, as a result, be misled or develop an over-reliance upon a single AI voice. In this paper we offer a provocation that takes a playful approach to multi-agent conversations, where our embodied AI agents, each with their own unique characteristics and preferences, offer diverse opinions in a social discourse setting. Such a multi-agent approach stimulates critical thinking and moves us away from a simplistic assumption that the AI agent is always right.

Agentic AI, Pluralistic Value Alignment, Natural Language Dialogue, Mental Health, Human Cognition

1. INTRODUCTION

Agentic AI is seen by many as the future of AI, with conversational agents poised to offer play, companionship, information and advice to a range of individuals, many of whom struggle to find more traditional sources of support. Increasingly this includes those seeking support for mental health issues (Feng et al., 2025), however, these one-to one human-AI relationships are not without their problems. A growing literature recognises that, whilst conversational agents can be persuasive tools that can offer valuable information and advice, they can also be used to flatter, to shore up inappropriate beliefs, or to misinform or mislead. For many, AI is best seen as a supportive rather than therapeutic tool, and recent initiatives would seek to reduce over-reliance on AI and to build personal agency and more considered decision-making (Murugesan, 2025).

With this in mind, we present a provocation in the form of argumentative teddy bears and other plushies that allow us to think more freely about multi-agent AI conversational systems that provide diverse AI generated opinions and advice. Our provocation takes the form of a set of 'cute' embodied agents (each with unique characteristics or approaches), prompted to disagree with one another and with the 'human in the loop', ensuring the human must ultimately make up their own mind as to what constitutes the best advice. We propose that such a multi-agent conversation creates greater exposure to other values, reduces sycophancy, offers better opportunities for personal reflection, and moves us away from simplistic 'AI is always right' single solution thinking.

2. BACKGROUND

There are three reasons why multi-agent conversations could prove a better agentic AI resource than single agent interactions. Firstly, over-reliance upon a single AI agent with 'one voice' can often be problematic (Bo et al., 2025), particularly in cases where people are vulnerable and likely to seize on any answer as a way out of a difficult situation, or where they are simply seeking confirmation that they've made the right decision. Secondly, there is significant psychological value in being exposed to a range of potential solutions when seeking advice and both uncertainty and disagreement between parties can be useful when weighing up options. Thirdly, Agentic AI works best when it supports, rather than replaces human decision-making and indeed systems that encourage critical thought are generally regarded as preferable. We will explore these issues in a little more detail, before going on to describe our multiagent provocation.

© Briggs et al. Published by BCS Learning and Development Ltd. Proceedings of BCS HCI 2026, UK

2.1. Over-reliance upon a single AI voice

People are increasingly turning to AI agents to ask for information and advice on relationships, health, finance and a range of practical matters. The agent offers a number of advantages: it is always there, it seems discreet and it avoids any unnecessary social awkwardness and so it isn't surprising that consultations with AI agents have been on the rise. However, people can quickly become over-reliant upon this single, readily available source of advice, which is particularly worrying when it comes to queries around mental health issues. A number of research trends have highlighted this problem, outlining serious ethical risks associated with AI counselling, including failure to understand context, mishandling of crises, inappropriate boundary violations, overconfident or inaccurate responses, and the creation of false empathy (Iftikhar et al., 2025).

AI generated advice from a single LLM can be limited, culture-blind and insensitive to the human range of values and beliefs, which means they offer advice that fails to respond appropriately to cultural context (Novis-Deutsch et al., 2025). There have been calls for AI to present pluralistic solutions that actively endorse or promote multiple value systems and some researchers have argued for a 'design justice' approach where AI systems are co-created with diverse patients and publics, centering multiple voices rather than imposing top-down solutions (Pozzi and Proost, 2025). Such papers recognize a concern that single AI voices (particularly in a therapeutic setting) can create systematic harm through cultural incompetence, algorithmic bias, and failure to recognize diverse expressions of mental health.

2.2. The value of uncertainty and disagreement

AI generated advice can prove problematic when we seek a more nuanced interaction, typical of humanhuman exchange. Good interpersonal advice relies not only upon an understanding of facts or history but also upon an understanding of the complex nature of human needs and preferences. Uncertainty in such situations becomes critical, and yet it is largely absent in an AI response.

The value of uncertainty as a resource for design was spelled out a few years ago Gaver et al. (2003) and the dangers associated with certainty and over-confidence have long been recognized by psychologists. For example, Janis's classic 'groupthink' paper (Janis, 2008) questioned the way that a team of people could collectively come to make a series of disastrous decisions when the group came too early to an unquestioning consensus. Groupthink was characterised by an illusion of invulnerability, the use of rationalisations to discount warnings, a sense of moral superiority, i.e. a belief that the group held values that were superior to others, coupled with stereotyped views of outsiders, an emphasis on consensus and a willingness to pressurise others who held contrary opinions. This ultimately led to self-censorship in members of the team who had doubts and created an illusion of unanimity.

It is easy to imagine how AI, as it is currently framed, might help to exacerbate a groupthink situation, where a group coalesces around an early AI generated 'solution' and where diverse, critical thinking is discouraged. Uncertainty and disagreement are useful tools that are largely absent in single AI responses, although they do feature in some AI co-reasoning systems designed to stimulate critical thought (Danry et al., 2023), or show pluralistic alignment (Sorensen et al., 2024), i.e. designed to recognise different perspectives and to present a range of solutions.

2.3. Encourage, rather than discourage human thought

We know that current use of GenAI has led to a selfreported drop in critical thinking and confidence in a range of workers Lee et al. (2025). However, human AI co-reasoning systems can be designed with the goal of stimulating rather than stifling critical thought (Danry et al., 2023). The authors of a recent Turing Institute paper on 'Doing AI Differently' (Hemment et al., 2025) note that, thus far, GenAI development has largely followed the assistant paradigm, with systems performing like search engines, clerks, copywriters, and illustrators, with system outputs largely conforming to supplied specifications. They argue that we should look beyond the assistant model toward more complex forms of human-AI collaboration, developing 'Human-AI ensembles' capable of strengthening our collective intelligence and enhancing human capabilities. If we translate this into the Agentic AI space, then we should see agents not as suppliers of solutions, but as entities capable of prompting deeper and more meaningful engagement with a problem.

3. PROVOCATION

3.1. Talking Ted-AIs

Imagine that a user can choose a set of characters to take part in a discussion or advice session. We suggest two or three characters as optimal (Figure 1). These characters would have a physical manifestation and we have been using plush animals (teddies, dogs, etc.) in our ongoing work, in part because we wish to use this provocation as a means of exploring the extremes and the risks of different forms of AI embodied interaction. Obviously, the physical appearance of such conversational agents



Figure 1: *Users might choose two or three AI-enabled characters (plush toys) with different personalities and opinions to host a discussion on a topic of their choice*

can vary and could indeed be made digital (avatars), however we found that the act of having a lively debate with (sometimes highly opinionated) teddies was a fun way to explore multi-agent AI.

Once the 'companions' have been selected, a user could choose a topic for discussion. These could be sensitive or highly controversial topics, but so far we've been exploring areas where individuals are likely to disagree and where good arguments can be made for different sides of a debate. Topics such as 'should I go vegan?', 'What is the best way to reduce my household energy bill?' allow for disagreement and uncertainty when it comes to personalized advice.

We then enter the debate phase, with our agents expressing their own views, doing some turn taking but also (depending on character) sometimes trying to get their point across more forcefully or sometimes sowing doubt. Again, we can think about ways in which this becomes more provocative. Should our teddies become aggressive? Should they swear at each other? Not necessarily easy to engineer in today's off-the-shelf, super-polite AI systems but such questions allow us to think about the reasonable limits of disagreement.

4. SYSTEM ARCHITECTURE

We have been exploring how to turn this provocation into a functional demonstration, using three 'bases' – positions where a character may be placed to take part in the discussion. Each base has an NFC reader and a speaker. The NFC reader is used to identify which, if any, character is present at a particular base, by reading the unique tags from the toy. The speakers are individually addressable so that the participant may correctly localize each character's voice.

A choice of character toys are presented, having their own individual personality traits, communication styles, and voices. The participant can select which characters to include in the discussion by placing the toys on the bases. Each toy contains a uniquely identifiable NFC tag, the ID of which is linked to a specific character profile in the system. Each profile is associated with: the character's name; a specific Large Language Model (LLM) to use for the character; a description of their personality, communication style, and perspective; example statements; and the speech synthesizer voice to use for the character's dialogue. The name, description and example statements are used to complete a prompt template to generate an overall initial LLM prompt for the character. Characters' utterances are synthesized into speech using their speech synthesizer voice, and played over the associated speaker.

The participant is situated in front of a directional microphone, and the audio is used to locally transcribe their speech with a speech recognizer (no sound is permanently recorded, stored, nor transmitted elsewhere). Voice activity detection is used to identify when the participant is speaking, so that they may interrupt the conversation at any time. Should the environment be too noisy to clearly detect the start of speech, an 'interrupt' button is provided, and this may also be used in a push-to-talk manner (held while talking) for even noisier environments, where the end of speech may be unclear.

We experimented with an early version of the system where we used a single LLM tasked with being a script writer, with the characters being introduced in an initial prompt, the LLM being prompted to continue the script for the non-human characters, and the human participant's input being inserted into the script. This approach, while simple, excluded the possibility of using truly separate language models for each character.

A multi-model version of the system would allow for each character to use a different language model (either locally, or from different LLM providers), along with their own customised prompt and could work as follows: The system is operated by a central controller which manages the conversation. Each character's chat begins with their unique initial prompt, and grows with the shared dialogue history from the participant and the other characters. After a contribution from any of the chat participants, the controller prompts each character to characterize their urgency (low, medium, high) to: begin speaking (at the conversation start, after participant input, or when all characters have yielded), continue or yield (if the character was the last to contribute), or to interrupt (if another character contributed last without yielding). These urgency questions are not retained in the chat history. The controller then selects the next speaker based on these urgencies, with a bias against long runs from a single character, and prompts for their next contribution. Where the participant interrupts, the controller truncates any

current character's utterance, both audibly and in the chat history, and the participant's input becomes part of the shared conversation. Parts of this system run in parallel where possible to minimize latency, such as prompting characters, and speech synthesis while the chat output is being streamed.

5. WHAT MAKES THIS AN INTERESTING PROVOCATION?

Agentic AI is clearly a critical issue for the CHI community, but current work typically assumes a single agent model. In this multi-agent exploration we tap into issues that are central to an understanding of the longer-term impact of AI agents on human wellbeing, cognition and decision-making, personal agency and pluralistic value alignment. There is, as yet, relatively little work that explores the one vs many voices in agentic AI (indeed, work on hybrid human-AI teams is still in its infancy, particularly when it comes to mental health support (Swinger et al., 2025)) but such matters are likely to become more pressing given the rapid roll out and adoption of AI agents.

Although this is a playful provocation, the underlying issues around AI over-reliance, cultural sensitivity and advice quality are extremely serious. It also allows us to explore interesting questions around the nature of embodiment in AI. Does a relatively 'powerless' and 'friendly' form make for a good opponent? Under what circumstances could we take talking teddies seriously? How disagreeable should agents be and what kinds of arguments might help rather than hinder human cognition? Might we use such systems to mirror team decision-making in order to learn more about the strategies and tactics used to undermine or empower certain individuals? As with many provocations within HCI and critical design communities, making something that looks a bit odd or unlikely can become a means to free up thinking about a particular problem and to challenge an overly simplistic adoption of the status quo.

6. ACKNOWLEDGEMENTS

This work was supported by two awards from the Engineering and Physical Sciences Research Council: The Centre for Digital Citizens - Next Stage Digital Economy Centre; EPSRC (EP/Y009800/1) and the Centre for Doctoral Training in Citizen-Centred Artificial Intelligence (CCAI) (EP/Y030729/1).

REFERENCES

Bo, J. Y., S. Wan, and A. Anderson (2025). To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

Danry, V., P. Pataranutaporn, Y. Mao, and P. 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 *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

Feng, X., L. Tian, G. W. Ho, J. Yorke, and V. Hui (2025). The effectiveness of ai chatbots in alleviating mental distress and promoting health behaviors among adolescents and young adults: Systematic review and meta-analysis. *Journal of Medical Internet Research 27*.

Gaver, W. W., J. Beaver, and S. Benford (2003). Ambiguity as a resource for design. In *Conference on Human Factors in Computing Systems - Proceedings*, pp. 233–240. Association for Computing Machinery (ACM).

Hemment, D., C. Kommers, R. Ahnert, M. Antoniak, G. Arbix, V. Belle, S. Benford, A. Brintrup, N. Bryan-Kinns, M. Bunz, et al. (2025). Doing ai differently: rethinking the foundations of ai via the humanities.

Iftikhar, Z., A. Xiao, S. Ransom, J. Huang, and H. Suresh (2025). How llm counselors violate ethical standards in mental health practice: A practitioner-informed framework. Technical report.

Janis, I. L. (2008). Groupthink. *IEEE Engineering Management Review 36*(1), 36.

Lee, H. P. H., A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, and N. Wilson (2025, 4). The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

Murugesan, S. (2025). The rise of agentic ai: Implications, concerns, and the path forward. *IEEE Intelligent Systems 40*, 8–14.

Novis-Deutsch, N., T. Elyoseph, and Z. Elyoseph (2025). How much of a pluralist is chatgpt? a comparative study of value pluralism in generative ai chatbots.

Pozzi, G. and M. D. Proost (2025). Keeping an ai on the mental health of vulnerable populations: reflections on the potential for participatory injustice. *AI and Ethics 5*, 2281–2291.

Sorensen, T., J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu,

N. Dziri, T. Althoff, and Y. Choi (2024). A roadmap to pluralistic alignment.

Swinger, N., C. M. Baseman, M. Ryu, S. Abdullah, C. W. Wiese, A. M. Sherrill, and R. I. Arriaga (2025). There's no" i" in teammait: Impacts of domain and expertise on trust in ai teammates for mental health work. *Proceedings of the ACM on Human-Computer Interaction 9*(2), 1–36.