# Explaining the Inexplicable: a Study of People's Reactions to Futuristic AI

Explaining the Inexplicable

Wally Smith

University of Melbourne, wsmith@unimelb.edu.au

Peta Masters

Kings College London, peta.masters@kcl.ac.uk

Jingyi Jannie Liu

Australian National University, jannieliu830@gmail.com

Gustav Kuhn

University of Plymouth, gustav.kuhn@plymouth.ac.uk

Ryan M. Kelly

RMIT, ryan.kelly@rmit.edu.au

To better understand people's reactions to futuristic and seemingly impossible technologies of Artificial Intelligence (AI), we studied the responses of 21 students to a phone app that performs two highly surprising conjuring tricks simulated through a Wizard of Oz method: uncannily accurate lie detection, and control over a user's apparently free choices. We observed how, in the absence of readily available interpretations, our participants generated a range of *pseudo explanations* for how the AI performed its feats, which we classified as *non-explanation*, *unresolved*, *concordant*, and *revelatory*. We interpret these findings in terms of Daniel Dennett's notion of explanatory stances, and suggest that when faced with hard-to-explain AI, people may shift away from typically mentalistic explanations towards more functional accounts.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI) • Laboratory experiments • User studies

**Additional Keywords and Phrases:** AI, explanation, magic, Wizard of Oz method

**ACM Reference Format:**

Wally Smith, Peta Masters, Jingyi Jannie Liu, Gustav Kuhn, and Ryan M. Kelly. 2024. Explaining the inexplicable: a study of people's reactions to futuristic AI. In Proceedings of Nov 30 - Dec 4, 2024 (OzCHI'24). ACM, New York, NY, USA. 10 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

As more sophisticated forms of Artificial Intelligence (AI) push further into our lives, questions arise about people's understanding of how these new technologies work, with implications for their acceptance, transparency, and efficacy. Researchers in explainable and interpretable AI (XAI), one of four fields identified by Capel and Brereton [2] at the intersection of HCI and AI, focus on ways to improve *how AI technologies explain themselves*, often by making explicit hidden aspects of their decision-making processes [12].

In this paper, we explore a related but different aspect of AI's interpretability, focusing instead on *how people explain AI technologies* [11]. While the typical context for XAI has been expert decision-making in domains such as medicine and law, our focus is on non-specialists' explanations for everyday AI such as Siri or ChatGPT. Our aim is to develop a preliminary account of the kinds of explanation that people construct, for themselves and others, about these potentially hard-to-understand technologies.

Applications like Siri, Co-Pilot and ChatGPT may seem amazing on first use, but typically they become accepted as normal and even mundane over time. We present a study that attempted to capture people's early affective and cognitive reactions to a hard-to-understand AI technology before any sense normalcy had set in. To meet this challenge, we simulated a futuristic AI technology in the form of a smartphone web-app ('MagicApp') that performs conjuring tricks with a regular pack of playing cards, instructing users to carry out all of the actions with the cards themselves, yet leading to a highly surprising and inexplicable outcome.

The contribution of this paper is to offer a preliminary understanding of how people react to, and attempt to explain, new forms of AI that are seemingly inexplicable. Although such early encounters are only one small part of human-AI relations, they may be important in shaping public understanding, appreciation and engagement. New applications of AI are often projected by the tech industry as being 'magical' and, by implication, beyond explanation: 'Built to do the impossible', is a recent promotional line for Microsoft's 'Co-pilot' [10]. Whether this is a good strategy, or whether technologies might be designed to foster different kinds of public understanding, is the broader debate that we aim to inform.

## 2 RELATED WORK

To frame our exploratory study, we draw on philosopher Daniel Dennett's account of three different 'explanatory stances' that people adopt towards technological systems [4]: a physical stance, producing explanations in terms of material properties, such as an electric current causing an electromagnetic field to repels a metal rod; a design stance, with explanations couched in terms of abstract functions, such as a thermostat regulating temperature; and an intentional stance, producing explanations based on mentalistic concepts of beliefs, intentions and desires, and the presumption that technology acts rationally. Taking the example of a chess-playing computer, Dennett argued that explanations of intelligent technologies, because of their behavioural complexity, unavoidably adopt an intentional stance. That is, we typically explain AI's behaviour in the terms used to explain a rational person's behaviour. This insight is carried forward by many researchers in XAI [5, 7] who use it to argue that technologies should be designed to express recognizable intentions and beliefs.

Dennett further argued that when we explain the behaviour of a person who is 'imperfectly rational', we often shift from an intentional stance to a design stance; that is, we explain them in terms of break-downs in the functioning of their thought processes. He claimed that we do the same for an intelligent machine when we suspect it of not behaving rationally (e.g., the strangely bad move by a chess-computer must be a bug in the software's piece-moving function). In support of Dennett's account, Malle reported that when people are unable find an intentional explanation of another's behaviour, they are

inclined instead to supply a ‘causal history of reason’ that appeals to the individual's personal history or cultural background [8]. In analogous fashion, de Graaf et al [3] found that, when people could not generate an intentional account of a robot’s behaviour, they explained it in terms of the way the robot must have been programmed.

In our study, we examine how Dennett’s account of explanatory stances, and the shifts between them, might shed light on people's explanations of MagicApp as an example of a futuristic and seemingly impossible AI. Specifically, we examined if people apply an intentional stance, as is typical for AI, or if instead they resort to functional terms as sometimes occurs for people and technologies that appear to be malfunctioning? Through its appeal to mentalistic concepts, an intentional stance depends on an anthropomorphic view of technology. In our study, such anthropomorphism may have been encouraged by a sense of MagicApp being deceptive, albeit playfully. Short et al [13] reported that people are more likely to attribute human-like qualities to a machine when they see it acting deceptively; in their study, a robot trying to cheat at a game of rock-paper-scissors by quickly changing its hand configuration on seeing that of its human opponent. They also found that participants frequently had an emotional response to the cheating robot: they were surprised and amused that the robot knew how to cheat. As way to gain insights into the explanatory stances adopted in our study, we examined whether these or similar responses characterised people’s reaction to MagicApp.

## 3 METHOD

We used a Wizard of Oz [9] method to create a web-based app that seemed to possess impossibly intelligent capabilities, but which actually depended in part on hidden human input. To make this a playful and engaging experience, our app took the form of a voice-assistant app that presented itself as performing two magic tricks, hence its name MagicApp. This builds on a long history of using technology, with and without human input, to create magical effects [14, 15]. At the end of each testing session, participants were fully debriefed on the true nature and workings of MagicApp. The study was approved by the Psychology Health and Applied Sciences Human Ethics Sub-Committee of the University of Melbourne (Approval id: 1954358). As part of our ethics application we prepared a protocol to mitigate any confusion or distress caused by the deceptive aspects of MagicApp. In practice, all participants were content with the playful nature of the deception at the debriefing stage.

### 3.1 The Design of MagicApp

The two simple but highly surprising card tricks performed by MagicApp enacted uncannily accurate lie detection, and control of participants’ apparently free choices. Trick 1 (lie detection) allowed for a possible explanation in terms the AI somehow analysing non-verbal speech cues, while Trick 2 (control of free choice) did not have any readily available explanation (see Table 1).

MagicApp performed its tricks with a regular pack of playing cards that was provided by the researchers. When participants activated a trick on the phone, the voice assistant delivered a sequence of spoken instructions, such as ‘Shuffle the cards thoroughly’ and ‘Cut off about half of the cards and place them to the right on the table’. After voicing each instruction, MagicApp waited for the participants to complete the action before giving the next instruction, until the whole trick was complete.

MagicApp achieves its effects through a combination of secret human inputs and the inherent deceptiveness of the magic tricks which are scripted into its voice instruction generator. The timing of the app's voiced instructions were triggered by a human researcher who was present during testing, thereby reinforcing an illusion that the app was responsive to the participants, instructing them only when appropriate or when requested to do so.

Table 1. The two tricks performed by MagicApp running on the participant's phone

| Trick | Feat Performed | The effect |
|---|---|---|
| Trick 1 | Lie detection | The voice of MagicApp explains that it will demonstrate that it can tell if the spectator is lying or telling the truth. It instructs the spectator to shuffle the playing cards well, and to select and remember a chosen card, which is then cut back into the pack. The app then instructs the spectator to deal the cards from the pack face up on to the table, and for each card to say out loud, "No, that is not my card", whether or not this is true. The spectator follows these instructions and tries to keep their voice constant, but when the chosen card appears, the app correctly calls out that they are lying and that this is the chosen card. |
| Trick 2 | Control of free choice | MagicApp explains that it will demonstrate that it can control the spectator's choices in advance, even when they seem to be freely made. It instructs the spectator to shuffle the playing cards well. The app makes a prediction by showing an image of a playing card on the phone's screen. The app then gives instructions for the spectators to cut the pack into four piles, to choose one and discard the others; then to deal this chosen pile into four 'hands' of cards and to choose one hand, etc. Finally, one freely selected card remains. It is turned face to reveal that it is the card originally predicted by MagicApp. |

The effectiveness of MagicApp in seeming to exhibit impossible capabilities depends on the structure of the trick routines that are built into the app. The precise workings of the tricks cannot be revealed in this publication, but are not important to understanding the effects that were created. Trick 1 (lie detection) deploys a simple linear script, while Trick 2 (control of free choice) enacts a more complex conjuring principle called 'equivoque' and its patter is generated in a non-linear fashion by a network of speech utterances that are navigated in response binary 'hot/cold' signals coming from a human controller. This network of speech utterances forms a simple finite-state automaton that models the state of the cards and the actions of spectators, and directs events towards a predetermined outcome, without this being detectable by participants.

### 3.2 Participants

The participants were 21 native Mandarin speakers in the age range 22 – 26, just under 50% female, all Masters students in computing and technology courses who answered an advert in a University news site. A version of the app in Mandarin was used to make its voice instructions readily intelligible. All participants were familiar with intelligent voice assistants, though 20% had never used one directly, 25% used them at least every week, with the rest in between. All had awareness of entertainment magic, with most reporting that they enjoyed it to some degree.

### 3.3 Procedure, Data Collection and Analysis

Participants were introduced to MagicApp in groups of 3 or 4 people, creating the experience of a small audience. They sat together at the end of a long table in a meeting room, while two researchers kept some distance at the other end of the table where a camera on tripod was mounted to record the event. The minimalism of the setting was such that participants could be very confident that no other cameras or technologies were operating in the space, and indeed there was none. One member of each group was invited to use their own phone and was given a link to access MagicApp, and they were given a regular pack of playing cards.

For all groups, Trick 1 (lie detection) was performed first followed by Trick 2 (control of free choice). After the performance of each trick, each participant was instructed to write a report of the performance addressing three aspects: 1. Their memory of what happened in enough detail to inform somebody who was not present, 2. Their reaction to what happened, and 3. Their best explanation for how the app performed the feat. Next, participants gave 5-point Likert

ratings on how *surprising*, *funny*, *enjoyable*, *deceptive*, and *interesting* they found the app, and they rated if the performance had influenced their view of the app in terms of it seeming *human-like* or *creepy*. Participants' open-ended explanations were subject to a simple thematic analysis conducted by two of the authors. Classification of explanations was done by each researcher independently, and then combined to form a single classification by consensus.

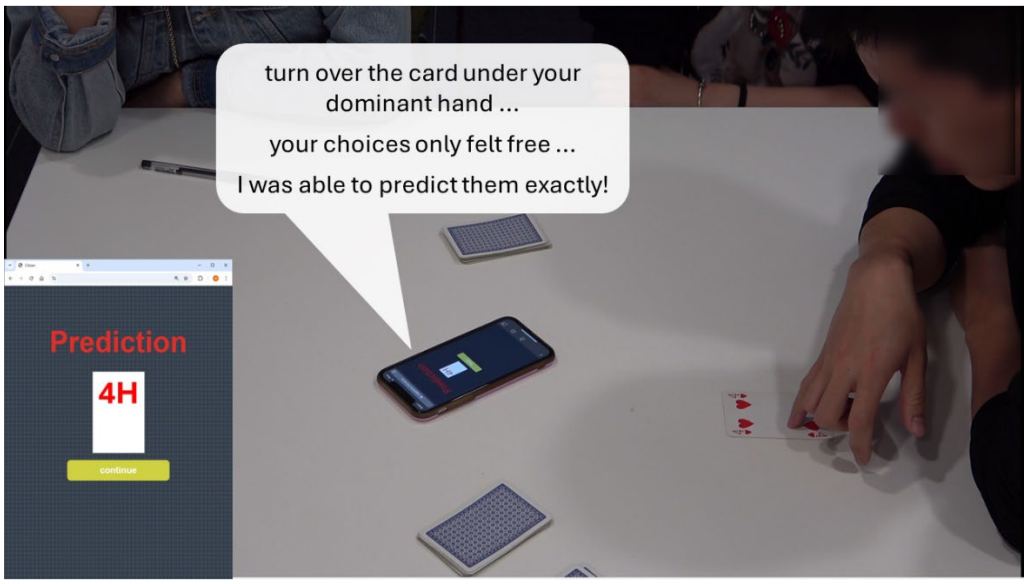


Figure 1. The conclusion of a performance of Trick 2 (control of free choice) with inset showing MagicApp's minimal interface and speech bubble showing the app's final three utterances

## 4 FINDINGS

Before turning to our main qualitative findings about the forms of explanation given for MagicApp (Section 4.2), we first present participants' rating scale data (Section 4.1). These ratings provide confirmation of our premise that MagicApp would be experienced as highly surprising, and they reveal the extent to which it was regarded as human-like and/or deceptive, these being associated with the adoption of an intentional stance.

### 4.1 Participant Ratings of Their Immediate Reaction to the Feats Performed

For both Trick 1 (lie detection) and Trick 2 (control of free choice), MagicApp succeeded in presenting feats of AI that were genuinely surprising and interesting, and which were therefore a suitable target for explanation. Participants' mean ratings (Figure 2) confirmed that they were overall greatly *surprised* (4.74), *interested* to know how the effects were achieved (4.90), and that they found the performances overall *funny* (4.62) and *enjoyable* (4.29). Overall, participants did not feel strongly that they had been *deceived* (3.10, closest to 'neither agree or disagree').

Ratings also suggested that overall MagicApp had influenced participants' view of the phone on which it was deployed (Figure 3), making it seem more *human-like* (3.90), *more human-like relative to a GPS voice assistant* (4.00), and slightly more *creepy* (3.24). In contrast to claims by Short et al [13], Goodman and Kruskal's non-parametric test of association revealed no associations between a sense of being *deceived* and of the app seeming *more human-like*, or *more human-like*

*relative to a GPS voice assistant*. However, strong positive associations were found between ratings of MagicApp being *surprising* and of the phone both *seeming more human* ($\gamma = 0.935$; $p<0.005$) and *seeming more human than a GPS* ($\gamma = 0.857$; $p<0.02$). Overall, this pattern was not observed for the harder-to-explain Trick 2 (control of free choice), although ratings of *surprising* were associated with the phone *seeming more creepy* ($\gamma = 0.780$; $p<0.001$).

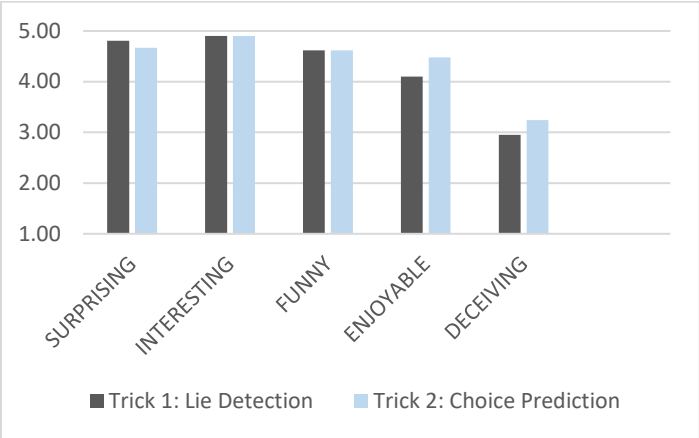


Figure 2. Participants' mean Likert ratings of their affective response to MagicApp following the two tricks (1 = strongly disagree, 5 = strongly agree)

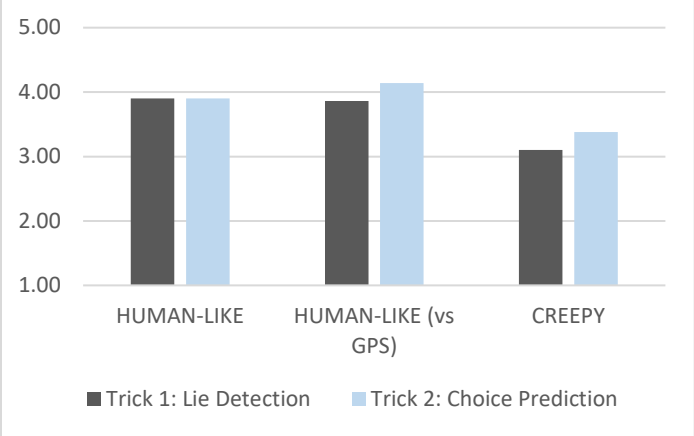


Figure 3. Participants' mean Likert ratings of the phone that ran MagicApp following the two tricks (1 = strongly disagree, 5 = strongly agree)

### 4.2 Thematic Analysis of Qualitative Reports

We conducted an inductive and interpretive thematic analysis of participants' written and spoken reports about MagicApp to conceptualize common types of explanation given for its feats; the explanation types are themselves the resultant themes.

While our analysis drew on standard interpretive techniques of Braun and Clarke [1], a fully reflexive thematic analysis was not necessary given our predetermined aim to identify the types of explanation given.

A key observation is that most participants, in the absence of knowing how the effects were accomplished, constructed what we will call *pseudo explanations*. These were statements that identified potentially relevant elements of the situation, but did not constitute a complete account of how the outcomes were reached. Some pseudo explanations contained no sense of causality, for example: for Trick 2 (control of free choice) '*It mentioned using the hand that you mostly use or the other one that you do not mostly use, thus I think this might be a clue*' (P10); '*maybe it uses a kind of math theory or probability theory?*' (P11). But more often, pseudo explanations provided incomplete causal accounts, with potentially relevant elements: for Trick 1 (lie detection) '*I think it's just some complex algorithms, maybe concerned with deep learning, that combines our operation time and voice speed and something … and fits in a model and gets the results*' (P13); '*The phone app maybe connects with the camera. And the camera could collect some information off the tester and then give some feedback to the app. Then, the app could analyze the information and then calculate the outcome.*' (P12).

Within these pseudo explanations, we distinguished four kinds of attempted interpretation: non-explanation; unresolved explanation; concordant explanation, and revelatory explanation.

**Non-explanation** (1 case in Trick 1, 7 in Trick 2). These are responses in which not even a tentative or incomplete explanation was offered. In these cases, participants responded that there was no apparent explanation and did not elaborate further: '*I have no idea what happened. Maybe … just magic?*' (P4).

**Unresolved explanation** (1 in Trick 1, 1 in Trick 2). These were explanations that were self-refuting. One participant gave this kind of explanation for both tricks: for Trick 1 '*The only input I gave was my voice … because it actually can't see what card I chose … but from the second experiment I realize there's no voice input, so it can't be voice input …*' (P20).

**Concordant explanation** (12 in Trick 1, 1 in Trick 2). Explanations of this type implicitly accepted the mode of operation projected by the app itself: that the app was indeed performing lie detection in Trick 1, and choice control in Trick 2. This was the most common kind of response for Trick 1: '*because my emotion changed … so the way I speak … the tone I used changed'* (P8); *'the app could analyse the different tones when people are telling the truth or lies*' (P4); '*I guess the app did the trick by tracking, analyzing the tone, speed, the vibration when we say "that's not my card"* ' (P19). Such explanations were often supplemented with a possible technological mechanism for the lie detection: '*maybe there is a machine learning way to get to our voice … because once we get nervous, once we see the true card, we will get some different mood in our mind and the voice will change, and the sensor will notice that*' (P16). In a few cases, peculiarities of MagicApp's performance were incorporated as supportive evidence for an explanation. For example, when the app failed during Trick 1, and only succeed on its second attempt, this was interpreted as providing support for a machine learning based explanation: '*with plenty of voices it has been already trained … the first time it failed because the data is still a bit limited … it needs more training*' (P15). Conversely, sometimes a concordant explanation was offered with an acknowledgement of a lack of supportive evidence; *'… the voice changes … actually I can't hear the differences, but there might be some differences*' (P9).

Because Trick 2 was always performed second, some participants tentatively carried over explanatory elements from Trick 1: '*maybe it can tell from the sound you make? I am really confused*' (P1). Interestingly, the greater sense of impossibility created by Trick 2 (control of free choice) did not translate into more concordant explanations. Only one participant offered a concordant explanation of choice control: '*maybe it can control … human's brain*' (P17). Instead, the strangeness of Trick 2 elicited many more revelatory explanations as we turn to next.

**Revelatory explanation** (7 in Trick 1, 12 in Trick 2). The final type of explanation we identified is one that implicitly rejects the technology's purported method of achieving its effects, and offers an alternative account of its workings. Despite this intention to expose a hidden reality of what was really going on, what participants typically revealed was another layer of pseudo explanation. For Trick 1, these revelatory explanations typically rejected the idea that participants' voices were being analyzed and pointed to some other basis for the effect: '*I believe this AI is not based on my sound … I guess this AI do[es] the prediction by calculat[ing] the time we split the stack of the cards … or the time of changing the sequence of cards*' (P18). A common tactic was to declare that some other abstract method was at play: '*it's all about statistics*' (P14); '*it should be some kind of mathematic[al] trick, which means it has nothing to do with the sensors in the phone*' (P9). Such revelations could also be made with an acknowledgement of the difficulty of accommodating the evidence: '*I think it is a mathematical calculation … but it is still very surprising to see it happen … because no one knows what we are choosing because all the cards are face down*' (P19). Another tactic was to draw a parallel with another situation, often a magic show: '*I have seen a similar trick done by a human magician … I think there must be an algorithm to detect which card you choose … I don't think this relies on voice or movement sense*' (3C). Finally, these revelatory explanations included 3 cases of identifying some of the correct elements underlying MagicApp's workings: '*I think the assistant … controlled something. Maybe she is the boss who is behind the app?*' (P16). Interestingly, these correctly identified elements were not articulated within full explanations of MagicApp's performance, and we still considered them to be pseudo explanations.

## 5 DISCUSSION

We now consider what has been learned about the nature of people's explanations of a seemingly impossible, or at least highly surprising, AI technology. First, it can be seen in Figure 2 that our methodology of presenting an app that performs card tricks succeeded in producing strong reactions of surprise across the 21 participants. When asked to explain how this futuristic technology performed its feats, the most striking observation was the prevalence of what we have called *pseudo explanations* that identify various potentially relevant elements of the situation, but do not fully account for how the surprising events were produced. These pseudo explanations, making up 32 of the total 42 explanations given in the study, appeared to fill a void of understanding created by a genuine explanation being unavailable. They can be seen as analogous to explanations of why people do things in terms of their cultural and historical background when a direct causal reason cannot be found, as described by Malle [8]. The only cases of people avoiding a pseudo explanation were 8 *non-explanations* that completely side-stepped the challenge of explaining MagicApp, and 2 *unresolved explanations*; the later being self-refuting accounts.

Of the pseudo explanations offered, we distinguished between concordant and revelatory accounts. *Concordant explanations* were those which implicitly accepted the claims of MagicApp to be enacting lie detection or choice control. The relative plausibility of MagicApp being able to perform lie detection, as opposed to controlling participants' free choices, resulted in far more concordant explanations: 12 for Trick 1 compared to just 1 for Trick 2. In a few cases, adopting such explanations sometimes appeared to distort participants' memory of events, with some recalling an actual change in their voice when lying for Trick 1. A similar effect has been reported by Olson et al [12] in their study of implanting false thoughts in people about a medical scanner.

*Revelatory explanations* were those which implicitly rejected MagicApp's claims by providing an alternative account of how it really worked to perform its feats. The greater implausibility of MagicApp's explicit claim in Trick 2 to be controlling the user's free choices resulted in more revelatory explanations that sought to expose what the app was really doing: 12 in Trick 2, including 3 partial identifications of its actual secrets, compared to 7 in Trick 1. Most striking was one participant who gave a concordant explanation of the lie detection in Trick 1 (lie detection) '*it can figure out our lie is*

*based on our tune when we say "This not my card" ... I guess when we didn't tell the truth our pronunciation may be different*' (P16); but when confronted with the impossibility of choice control in Trick 2, shifted to a revelatory explanation that implicated a possible role of the researchers.

Taken together, this pattern of pseudo explanations suggests that when confronted with a highly surprising and futuristic AI, people's explanations are influenced as much by its framing narrative, such as lie detection, as by their own witnessing of what it actually happened. When the framing narrative was considered plausible, participants were more likely to construct a concordant explanation. But when the framing narrative was considered implausible, regardless of the strength of the evidence of the technology in action, they were more likely to attempt a revelatory explanation of what the technology was really doing.

One of Dennett's key claims is that people tend to adopt an 'intentional stance' for people and AI technologies, interpreting them through mentalistic concepts [4]. Interestingly, then, the pseudo explanations constructed by our participants were overwhelmingly formulated in terms of technological functions, what Dennett called a 'design stance'. Participants did not speak at all of beliefs, desires or intentions to interpret or explain MagicApp's behaviour. This is particularly surprising given that participants overall rated the app as seeming human-like. However, a possible resolution of this anomaly can be found in another of Dennett's claims that people often revert to a 'design stance' to interpret non-rational behaviour by people or intelligent machines, now preferring to explain them in functional terms. This suggests the possibility that when faced with a form of machinic agency that is beyond our grasp to understand, and in some sense superior to us, we may treat it in the same way we treat non-rational behaviour. That is, for hard-to-explain AI, we may abandon the mentalistic concepts of the intentional stance and resort to functionals explanation under a design stance. This might explain why our participants appeared to do just that when interpreting the inexplicable behaviour of MagicApp.

## 6 LIMITATIONS AND CONCLUSION

We have presented a study of how people attempt to explain novel and seemingly impossible AI technologies, represented here as an app that seemed to perform lie detection and exert control over a user's free choices. The study was carried out on technology students, and involved them interpreting a particular and unusual technology, with these factors clearly limiting the generalizability the observed behaviour. Nevertheless, we found that our participants overwhelmingly filled their void of understanding about the novel AI they encountered with what we have called *pseudo explanations;* meaning that they identified possibly relevant elements but did not cohere into a complete account. We distinguished attempts to construct *concordant explanations*, that implicitly accepted the way the AI presented itself, from those attempting *revelatory explanations*, offering an alternative account of how the AI was really working. All of these pseudo explanations were expressed in terms of technological functions, which is inconsistent with Dennett's and others' claim that explanations of AI typically adopt mentalistic terms (such as beliefs, intentions, desires). However, Dennett also noted that irrational human behaviour often elicits functional explanations (such as a breakdown of faculties), and an analogous effect with technology might underpin our participants' use of functional concepts in their attempts to explain the inexplicable AI that they encountered in our study.

## ACKNOWLEDGMENTS

This work was supported by the Australian Research Council (DP180101215). The paper benefitted greatly from the helpful comments and suggestions of two reviewers.

## REFERENCES


[1] Braun, Virginia & Victoria Clarke, 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[2] Capel, Tara, and Margot Brereton, 2023. "What is human-centered about human-centered AI? A map of the research landscape." In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1-23.

[3] de Graaf, Maartje MA and Bertram F Malle. 2019. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In *Proc. International Conference on Human-Robot Interaction*. ACM/IEEE, 239–248.

[4] Dennett, Daniel C. 1971. Intentional systems. *The Journal of Philosophy*, 68, 4, 87–106.

[5] Eyssel, Friederike, Dieta Kuchenbrandt, and Simon Bobinger. 2011. Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of the 6th international conference on Human-Robot Interaction*. ACM. 61–68.

[6] Hoffman, Robert R., Shane T. Mueller, Gary Klein and Jordan Litma., 2023. 'Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance', *Frontiers in Computer Science* Vol 5, 2023.

[7] Kiesler, Sara and Jennifer Goetz. 2002. Mental models of robotic assistants. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. ACM, 576–577.

[8] Malle, Bertram F. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press.

[9] Maulsby, David, Saul Greenberg and Richard Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 277-284). ACM.

[10] Microsoft, https://www.microsoft.com/en-au/surface/devices/surface-pro-11th-edition, accessed 01-July- 2024

[11] Miller, Tim. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, pp 1-38

[12] Olson, Jay A., Mathieu Landry, Krystele Appourchaux and Amir Raz. 2016. Simulated thought insertion: Influencing the sense of agency using deception and magic. *Consciousness and Cognition*, 43, pp.11-26.

[13] Short, Elaine and Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *Proc. International Conference on Human-Robot Interaction*. ACM/IEEE 219–226.

[14] Smith, Wally. 2015. Technologies of stage magic: Simulation and dissimulation. *Social Studies of Science*, 45, 3, 319-343.

[15] Sussman, Mark. 1999. Performing the intelligent machine: Deception and enchantment in the life of the automaton chess player. *The Drama Review*, 43, 3, 81-96.