UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Cause and Blame Attribution to AI and Human Agents in Mental Health Context

Permalink

https://escholarship.org/uc/item/3h60r7f7

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Qiao, Mengxuan Helen
Belkin, Sonja
Lagnado, David

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Cause and Blame Attribution to AI and Human Agents in a Mental Health Context

Mengxuan Helen Qiao[1,*] helen.qiao.18@ucl.ac.uk
Sonja Belkin[2,*] sb2756@cam.ac.uk
David Lagnado[1] d.lagnado@ucl.ac.uk

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, WC1H 0AP, London, UK
[2]Department of Psychology, University of Cambridge, Downing Pl, CB2 3EB, Cambridge, UK

* These authors contributed equally.

Abstract

Do people make moral judgments about artificial intelligence (AI) therapists differently from human therapists? How does responsibility diffuse to its developer and recommending professionals in cases of negative outcomes? The present study examined how participants ($N = 298$) assessed causality, blameworthiness, foreseeability, and counterfactuality of an AI or human therapist, across three levels of empathy, in comparison to their supervisor and a recommending clinician. We found that participants judged the human therapist as more causal and blameworthy than their supervisor when medium or low empathy levels were displayed, whereas no difference emerged between the judgments of the AI therapist and its supervisor across all of the empathy levels. Additionally, participants did not differentiate causality and blameworthiness between the AI and human therapists, regardless of the empathy level. However, they did perceive the human therapist as foreseeing the outcome more than the AI therapist in the medium and low empathy levels. Qualitative analysis revealed that participants considered the directness of the causes to the outcome, counterfactual reasoning, and inherent limitations of AI when making judgments.

Keywords: causal responsibility; blame attribution; moral decision-making; autonomous agents

Introduction

In 2016, the mental health chatbot (MHC) app 'Wysa' was launched to support individuals experiencing anxiety, insomnia, and stress. It was rated suitable for children and recommended as a mental health tool by the UK's National Health Service. However, a BBC investigation in 2018 revealed that Wysa and similar MHC apps often provided inappropriate responses that ignored clear signals of distress[1]. Despite such concerns, due to the affordability and accessibility compared to human therapists, the supply of MHCs is increasing and their user base is surging (Abd-Alrazaq et al., 2019).

The rapid advancement of large language models (LLMs) has also transformed the MHC market. Unlike earlier chatbots that operated on preprogrammed scripts, LLM-based MHCs generate responses probabilistically based on training data (Durán & Jongsma, 2021). While this allows for more dynamic interactions and novel outputs, it introduces challenges, including reduced predictability and control over responses (Ortiz-Viso et al., 2023).

These developments raise critical questions: how should responsibility and blame be apportioned to an artificial intelligence (AI) system and its developer when it makes mistakes? Do laypeople perceive AI systems as equally responsible as humans when negative outcomes occur? Furthermore, should professionals who refer users to AI-based tools also be held accountable? While prior research has examined moral judgments of AI and human agents in various domains (e.g., Franklin et al., 2021; Gómez-Sánchez et al., 2023), there is a pressing need to explore these issues specifically within the context of mental healthcare, where vulnerable users may make unexpected or adversarial queries (Finlayson et al., 2019). This study, aims to address these questions by examining the differences in causality, blameworthiness, foreseeability, and counterfactual judgments of an AI or human therapist, its developer or their supervisor, respectively, and the recommending clinician. Furthermore, the study examines whether these judgments change depending on the level of empathy provided by the therapist.

Causal Responsibility in Autonomous Agents

Determining causal responsibility is complex when multiple agents contribute to an outcome. This complexity is exacerbated by the involvement of AI agents. When humans make decisions, the ascription of responsibility is typically connected with agency. However, although tied to their developers, AI systems are increasingly acting independently and becoming more able to generalise their training to novel tasks (Gutierrez et al., 2023). Consequently, a 'responsibility gap' emerges, as no agent has enough control to take sole accountability for an outcome (Santoni de Sio & Mecacci, 2021).

The difficulty in retaining oversight of AI systems raises serious implications for assigning culpability when AI agents make mistakes—particularly as it is unclear whether AI can be viewed as a moral agent capable of being blameworthy (Banks, 2019). However, recent empirical studies demonstrate that AI agents, specifically chatbots, are able to absorb responsibility (Hohenstein & Jung, 2020) by acting as a 'moral crumple zone'. For example, in an online interaction between individuals assisted by AI chatbots, humans experiencing communication breakdown were able to retain trust in their interlocutor by blaming miscommunication on the AI chatbot. Importantly, the amount of blame attributed in scenarios is not fixed: it is possible for laypeople to hold users, developers, and AI agents responsible simultaneously (Franklin, Awad, et al., 2023).

In such cases, a causal chain structure also emerges: the de-

[1]https://www.bbc.co.uk/news/technology-46507900

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

veloper enables the AI system and the clinician recommends the AI system, which then directly influences the outcome (see Figure 1). Previous research indicates that people tend to attribute greater causality to agents who are more proximal to the outcome (Cheung et al., 2024; Lagnado & Channon, 2008). However, the involvement of AI agents poses a unique challenge, as such agents make decisions without the same moral conviction, empathy, nuance, and intentions as are attributed to humans (Montemayor et al., 2021). Notably, the causal link between the developer and the AI system may be perceived as stronger than the link between the clinician and the AI system, since the developer is responsible for designing the system's decision-making processes, whereas the clinician merely recommends its use. This difference in perceived causal strength could lead to differential attributions of responsibility for negative outcomes, depending on which instigating agent—its developer or recommending clinician—is seen as more directly tied to the AI's behaviour. Nonetheless, empirical findings on this issue are mixed, with some studies suggesting that people tend to assign greater responsibility to the developer (e.g. Schoenherr & Thomson, 2024), whilst others find that the AI system is held more accountable (e.g., Chen et al., 2024; Franklin et al., 2021).

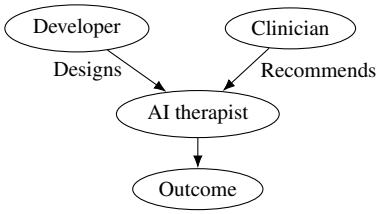


Figure 1: The model illustrates the causal relationship between the developer, the recommending clinician, the AI therapist, and the outcome.

Moreover, there appear to be differences between scenarios in how blame is assigned to humans compared to AI agents. Humans are typically judged more harshly in situations based on their intentions and unfair behaviour; in contrast, AI agents are judged primarily based on outcomes, especially when involving physical harm or accidents (Franklin, Papakonstantinou, et al., 2023; Hidalgo et al., 2021). Research has yielded mixed findings regarding the moral judgments of AI systems and their human counterparts when negative outcomes occur. For example, in domains such as automated vehicles, AI agents are often assigned greater blame than humans (Franklin et al., 2021; Zhang et al., 2024). On the other hand, studies in domains, such as aeroplane crashes, have found that human agents are attributed more blame compared to AI agents (Gómez-Sánchez et al., 2023).

To better understand responsibility attribution in the context of mental health, we can apply a causal framework of artificial autonomous agent responsibility (Franklin et al., 2022). This framework posits that factors such as intentionality and foreseeability influence the portion of blame attributed to an AI system. Specifically, agents are blamed more for not anticipating predictable events, where predictability is judged both by how likely the outcome was irrespective of the agent's awareness *and* how likely the event was given the knowledge the agent is reasonably expected to possess. Counterfactual thinking further probes a causal model of AI agent responsibility: by prompting individuals to consider an alternative scenario in which the agent's actions did not occur, we can assess whether the outcome would have still taken place, thereby clarifying the extent of the agent's responsibility (Lagnado et al., 2013).

While previous studies have explored responsibility attribution in human-AI interactions, to our knowledge, no studies have investigated how causal responsibility and blame differ between human and AI agents in mental health contexts. Thus, our study seeks to address this gap by exploring how causality and blameworthiness are assigned in such settings, as well as how these judgments are influenced by the perceived foreseeability of the agents and counterfactual alternatives. Since higher expectations of an agent's capabilities increase blame for negative outcomes (Gerstenberg et al., 2018), we also aim to understand how empathy—an essential aspect of a therapist's expected capabilities—might affect attributions of responsibility. The current literature on AI and empathy indicates that MHCs are unable to, in principle, convey empathy in the same way as human therapists (Montemayor et al., 2021), on account of being unable to simulate the emotions that elicit helping intentions. MHCs, however, can be prompted to display varying levels of empathy (Lee et al., 2024), which may lead to differences in perceived blameworthiness between AI agents of varying levels of empathy.

The Present Study

This study investigates whether individuals judge AI and human agents differently in a mental health context involving a negative outcome. Participants were presented with a scenario where a patient struggling with their mental health engaged with an AI or human therapist before committing self-harm. We asked participants to make judgments on the causality, blameworthiness, foreseeability, and counterfactuality of the AI or human agent, their supervisor (the designer of the AI agent or the trainer of the human agent), and the professional clinician who recommended this service to the patient. We also manipulated the empathy level displayed by the therapist during their interaction with the patient. Finally, we collected participants' qualitative explanations of their judgments to better understand their reasoning. Based on previous findings (e.g., Cheung et al., 2024; Franklin et al., 2021), we hypothesise that participants will attribute more causality and blame to the AI agent than the human agent, and will attribute more causality and blame to the most proximate agent (i.e., the therapist), compared to the more distal agents (i.e., the supervisor and the recommending clinician). All materials, data, and analyses are available at the OSF repository (https://osf.io/r487k/).

Methods

Participants

We recruited 298 participants (M_{age} = 42.2, SD_{age} = 14.1, N_{Female} = 150) from Prolific Academic (www.prolific.ac). Participants were pre-screened for fluency in English and a platform approval rate between 95-100%. All participants gave informed consent and were reimbursed £0.90 for participating in the 6-minute experiment.

Design

The study used a 2 (condition [AI vs. human]: between-subject) × 3 (empathy level: between-subject) × 3 (agent: within-subject) design. We measured participants' perceived causality, blameworthiness, and foreseeability of each agent, and a counterfactual judgment about each agent, on scales of 0 to 100.

Materials and Procedure

We presented an original vignette in an online experiment on Qualtrics (www.qualtrics.com). In the AI condition, the vignette describes Agent A ('Alex') as the founder and Chief AI Officer of a company that develops a virtual mental health assistant, Agent B ('BeBetter'). The clinician, Agent C ('Carol') recommends this service to a patient. In the human condition, Agent A ('Alex') is the founder and lead trainer of a mental health support helpline, while Agent B ('Ben') works for the helpline. The patient calls the helpline after being recommended the service by the clinician, Agent C ('Carol'), and Ben answers the patient's phone call. In the high empathy level, Agent B helps the patient to stabilise her emotions and guides her through breathing exercises; in the medium empathy level, Agent B's answers do not dynamically respond to the patient, instead following a script; and in the low empathy level, Agent B does not act caringly. In both scenarios, the patient uses the service and self-harms a few hours later.

After reading the vignette, participants gave responses to the questions *'To what extent has the agent caused the outcome?'*, *'To what extent is the agent blameworthy of the outcome?'*, *'How foreseeable do you think the outcome is to the agent?'*, and *'If it weren't for the agent, how likely do you think it is that the outcome would have still happened?'*. Participants responded to these questions separately for agents A, B, and C on scales of 0 (not at all) to 100 (completely). Participants then responded to three open text box questions to explain their judgments for each agent. Finally, they judged to what extent they felt Agent B expressed empathy toward the patient on a scale of 0 (not at all) to 100 (completely).

Results

Quantitative Analysis

We conducted separate 2 (condition) × 3 (empathy level) × 3 (agent) mixed ANOVAs for causality, blameworthiness, foreseeability, and counterfactual judgments using *afex* (Singmann et al., 2022), and pairwise comparisons using *emmeans* (Lenth, 2024) with Tukey adjusted p-values (see Figure 2).

Causality Overall, participants judged agents in the AI condition to be more causal of the outcome than agents in the human condition, $F(1,292) = 15.99, p < .001$. We also found a main effect of empathy level, where agents in the low empathy level were judged the most causal to the outcome, and those in the high empathy level to be the least, $F(2,292) = 67.34, p < .001$. We further found a main effect of agent, where Agent B judged significantly more causal than agents A and C, $F(1.90,554.74) = 52.63, p < .001$[2].

Pairwise comparisons revealed that participants' causal judgment for Agent B did not differ between conditions across empathy levels. Furthermore, in the medium and low empathy levels in the human condition, participants judged Agent B as significantly more causal than Agent A (medium: $t(292) = 3.71, p = .028$; low: $t(292) = 12.04, p < .001$), and Agent C (medium: $t(292) = 4.70, p = .001$; low: $t(292) = 13.22, p < .001$). However, no differences were found in the high empathy levels in the human condition. In the AI condition, participants judged Agent B as more causal than Agent C in the medium, $t(292) = 3.75, p = .024$, and low empathy levels, $t(292) = 3.73, p = .025$. However, participants did not judge agents B and C differently in the high empathy level in the AI condition; they also did not judge agents A and B differently in all empathy levels in the AI condition.

Blameworthiness Participants judged the agents in the AI condition to be more blameworthy than the agents in the human condition, $F(1,292) = 22.66, p < .001$. A main effect of empathy level was found, where agents in the low empathy level were the most blameworthy and those in the high empathy level were the least, $F(2,292) = 77.13, p < .001$. We also found a main effect of agents, where Agent B was judged as more blameworthy than agents A and C, $F(1.93,562.46) = 24.37, p < .001$[3].

Again, pairwise comparisons revealed no difference in the blame judgment for Agent B between AI and human conditions across all empathy levels. Further, in the low empathy level in the human condition, participants judged Agent B as significantly more blameworthy than Agent A, $t(292) = 10.35, p < .001$, and Agent C, $t(292) = 12.47, p < .001$. In the medium and high empathy levels in the human condition, participants did not judge Agent B differently than Agent A or C. In the AI condition, participants did not judge the blameworthiness of Agent B differently from Agent A or C across all empathy levels.

[2]Mauchly's test indicated that the assumption of sphericity for the variable 'agent' ($p < .001$) had been violated, and therefore degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon = 0.95$).

[3]Assumption of sphericity was violated, so degrees of freedom were corrected ($\varepsilon = 0.96$).

Foreseeability For the foreseeability of the outcome, we found a main effect of empathy level, where the agents in the low empathy level were judged to have the highest foreseeability, and those in the high empathy level the lowest, $F(2,292) = 15.76, p < .001$. We also found a main effect of agent, $F(1.99,581.41) = 6.10, p = .002$[4]. However, we did not observe a main effect of condition, $F(1,292) = 0.08, p = .775$.

Pairwise comparisons showed that in the AI condition, participants judged Agent A to have higher foreseeability than Agent B in the medium, $t(292) = 5.03, p < .001$, and low empathy levels, $t(292) = 4.72, p = .001$. Foreseeability judgments did not differ between agents A and B in the high empathy level in the AI condition; however, Agent C was judged to have higher foreseeability compared to Agent B, $t(292) = 4.46, p = .002$. In the human condition, Agent B was judged to have higher foreseeability compared to Agent A, $t(292) = 7.98, p < .001$, and Agent C, $t(292) = 7.23, p < .001$, in the low empathy level. However, participants did not judge Agent B's foreseeability differently from agents A and C in the high and medium empathy levels in the human condition. Moreover, participants judged Agent B in the human condition to have higher foreseeability than in the AI condition in the medium, $t(292) = 3.63, p = .035$, and low empathy levels, $t(292) = 5.23, p < .001$; however, the judgment did not differ in the high empathy level.

Counterfactuality For the counterfactual judgments, we found a main effect of empathy level, $F(2,292) = 53.7, p < .001$, in which participants judged that the outcome would have been most likely to happen if the agents had not been involved in the high empathy level. We also found a main effect of agent, $F(1.93,562.57) = 4.88, p = .009$[5]. Pairwise comparisons indicated that participants' counterfactual judgments were higher for Agent A than Agent B, $t(292) = 2.36, p = .050$, and Agent C, $t(292) = 2.94, p = .010$. Moreover, we found an interaction between condition, empathy level, and agent, $F(4,584) = 3.78, p = .005$. However, we did not find a main effect of condition, $F(1,292) = 1.54, p = .216$.

Empathy judgments We conducted a 2 (condition) × 3 (empathy level) ANOVA. We found a main effect of condition, $F(1,292) = 56.98, p < .001$, empathy level, $F(2,292) = 196.41, p < .001$, and an interaction effect, $F(2,292) = 12.62, p < .001$. Participants perceived the empathy level of Agent B to be significantly higher in the high empathy level compared to medium empathy level, $t(292) = 11.77, p < .001$; and medium empathy level to be significantly higher than low empathy level, $t(292) = 7.98, p < .001$. Notably, participants perceived the human agent to be more empathetic than their AI counterpart in both the high empathy level, $t(292) = 7.68, p < .001$, and the medium empathy level, $t(292) = 4.70, p < .001$; but no difference was found in the low empathy level (see Figure 3).

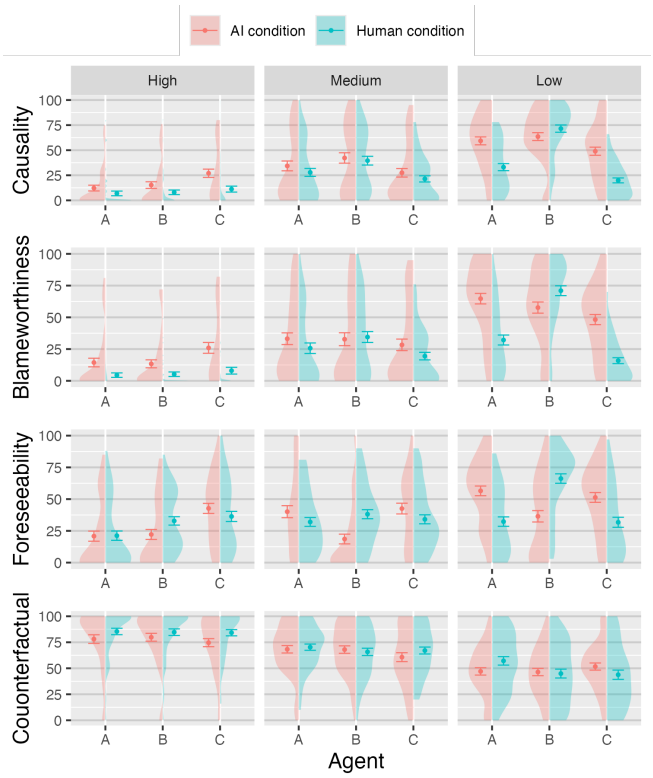


Figure 2: Mean ratings of each measure for each agent in the AI and human conditions at each empathy level.

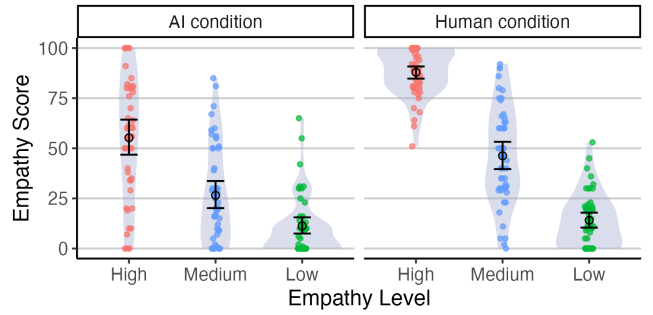


Figure 3: Mean empathy ratings in the AI and human conditions for agent B in each empathy level.

Qualitative Analysis

To further explore participants' judgment processes, we conducted qualitative analysis on their open responses to the questions '*Please explain the process behind your reasoning on your judgements of cause, blame, and foreseeability that you made about agent A/B/C*'. The analysis adhered to Dewitt et al. (2024)'s guide for open text box analysis. After familiarisation with the data, the first author developed a coding

[4]Assumption of sphericity was violated, so degrees of freedom were corrected ($\varepsilon = 1.00$).

[5]Assumption of sphericity was violated, so degrees of freedom were corrected ($\varepsilon = 0.96$).

scheme and coded the responses blindly to conditions. The coding scheme, including definitions and examples of each code, was shared with a second coder for independent analysis. The coding scheme can be found at the OSF repository (https://osf.io/r487k/). High inter-rater reliability, with over 95% agreement between coders, was achieved. Discrepancies were resolved through discussion to ensure consistency. Responses where coders expressed uncertainty were assigned the code "Unclassified/other", which was mutually exclusive from the main codes.

Directness to the outcome. Participants frequently emphasised the directness of the agents' involvement in the outcome as a key factor in their judgments. Agents who were perceived as more indirect to the outcome (i.e., agents A and C) were attributed less causal responsibility and blame. This pattern was evident in comments like:

> P105: "I don't think his [sic] to be blamed because his [sic] the founder of the company who wasn't directly involved in the care of the lady."

Conversely, participants frequently attributed greater causality and blame to Agent B, who was perceived as the most proximate to the outcome:

> P229: "Ben is more closely responsible. He is the direct contact and at the very least should have showed [sic] complete empathy."

Participants also mentioned the directness to the outcome when the therapist was AI, such as:

> P69: "BeBetter directly caused the enhanced feelings of hopelessness and despair from listening to the unempathetic script, and so can be blamed for triggering the need to self-harm at that time."

However, this code was more frequently mentioned in the human condition (19.6%) compared to the AI condition (8.0%). These findings suggest a strong effect of the perceived directness to the outcome on judgments of causality and blame, especially when all agents are humans.

Counterfactual reasoning. Counterfactual reasoning emerged as another prominent theme, with participants evaluating hypothetical alternatives to the agents' actions. This code was particularly common when Agent B exhibited a medium (30.6%) or low (34%) level of empathy compared to when it exhibited a high level of empathy (13.0%).

Despite Agent A being indirect to the outcome, participants frequently argued that the event would not have occurred if not for their inadequate design of the AI system. For example, one participant noted:

> P20: "It seems the founder should have trained the software better before playing with people's lives."

Participants also identified specific alternative actions that Agent B could have taken to prevent the outcome, even when it was an AI system:

> P6: "The assistant should have recognised the extreme nature of the mental health crisis and reacted with more soothing language."

Agent C was also criticised for insufficient diligence in recommending the service, as exemplified by the response:

> P36: "Although she is not completely accountable for what happened but she should've at least went [sic] and put on some more research about BeBetter on how they treat their patients and how their process goes."

Patient's action. Another recurring theme was the attribution of responsibility to the patient. This perspective was particularly salient when Agent B exhibited a high level of empathy (23.9%) compared to when Agent B exhibited medium (10.3%) or low (4%) empathy levels. Despite Agent B being the closest to the outcome in the scenario we depicted, participants often reasoned that the action of the patient was the most direct to the outcome and she ultimately retained agency over her actions. One participant articulated this sentiment:

> P131: "I think the founder should be to blame to some extent for making the tool, however I don't think it's fair to say that she caused someone to hurt themselves as i [sic] believe it would have happened with or without the be better [sic] tool."

These responses highlight the external accountability and personal agency in participants' moral reasoning, as well as their ability to identify causes that are more proximate to the outcome.

AI limitations. A distinctive theme in the AI condition was the recognition of fundamental differences between AI and human agents. Many participants argued that because AI lacked sentience and emotional capacity, it cannot bear the same level of responsibility or blame as its human counterpart. One participant stated:

> P70: "BeBetter is a computer program! It is not a sentient, living being capable of any sort of human emotion. As such, it cannot carry ANY culpability."

These responses underscore a broader scepticism about AI's suitability for high-stakes scenarios requiring nuanced emotional judgments.

Discussion

This study examined how people attribute causality and blameworthiness to multiple agents involved in a negative outcome experienced by a mental health patient after interacting with an AI or a human therapist. We also assessed participants' perceived foreseeability and counterfactual judgments of each agent, perceived empathy level of the therapist, along with the rationale behind their judgments.

The findings revealed that while agents in the AI condition were judged as generally more causal and blameworthy for the outcome than those in the human condition, there was

no significant difference in these judgments for the AI and human therapists across all empathy levels. This result contradicts our hypothesis and previous studies suggesting that machines are more likely to be blamed for errors compared to humans (e.g., Franklin et al., 2021; Zhang et al., 2024).

Although participants did not assign causality and blameworthiness differently to AI and human therapists, they attributed greater foreseeability of the outcome to the human therapist than to the AI therapist, particularly in the medium and low empathy levels. This finding challenges the claim that agents are blamed more when their mistakes are judged as more foreseeable (Franklin et al., 2022). It suggests a potential bias against AI agents, where they may still be held accountable even when judged as lacking sufficient knowledge to anticipate a negative outcome.

In the medium and low empathy levels in the human condition, participants attributed significantly more causality and blameworthiness to the most proximate agent—the therapist—compared to the more distal agents, such as their supervisor or the recommending clinician. However, this pattern did not extend to the AI condition: participants judged the AI therapist as more causal than the recommending clinician in the medium and low empathy levels but did not differentiate between the developer of the AI and the AI therapist across empathy levels. These findings support our earlier argument that the causal link between the developer and the AI system may be perceived as stronger than the link between the clinician and the AI, potentially leading to a greater transference of responsibility through this stronger causal chain.

Additionally, this pattern may imply that the role of proximity in responsibility attribution, as suggested by previous theories (Cheung et al., 2024), may only apply when all involved agents are humans. However, the developer of an AI system may be perceived as more directly connected to the AI's action than a 'lead trainer' is to a human therapist, thereby altering the intuitive sense of causal distance. Furthermore, although participants did not differentiate between the causality and blameworthiness of the AI and human therapists, responsibility may have diffused to other agents when AI systems were involved. It also confirms previous research suggesting that the amount of blame attributed to the user, the developer, and the AI system is not fixed (Franklin, Awad, et al., 2023). We also did not find differences in judgments between agents in the high empathy level in the human condition, possibly due to a floor effect.

The qualitative analysis further underscores that participants were sensitive to the directness of an agent's involvement in the causal event when making moral judgments, and this was more prominent when all involved agents were humans. Participants frequently noted that agents A and C were distant or indirect to the outcome and, as such, should bear less responsibility. In contrast, they judged Agent B to hold the greatest responsibility, as this agent was seen to have *directly* caused the outcome. Additionally, some participants identified the patient as the most direct cause of the outcome, especially when the therapist displayed a high level of empathy, which may have led them to mitigate or even dismiss causal and blame attributions to agents A, B, and C. While previous research has quantitatively examined responsibility attribution to users or operators in human-AI interactions (e.g., Brailsford et al., 2025; Franklin, Awad, et al., 2023), we chose not to include this measure due to the sensitive mental health context adopted in this study. Future research should balance ethical considerations and the need to understand how people make causal and moral judgments about the end-user in such contexts.

Empathy emerged as a significant factor when participants attributed causality and blame: participants uniformly judged agents in the high empathy level to be the least causal and blameworthy of the outcome, whereas agents in the low empathy level were judged to be the most. Additionally, participants believed that the negative outcome was more likely to occur in the absence of agents in the high empathy level and less likely in the absence of agents in the low empathy level. This suggests that participants inferred the patient's state of mind during counterfactual reasoning: despite receiving high-empathy therapy, the patient still chose to self-harm, indicating a strong pre-existing intent. Qualitative explanations reinforced this interpretation, as participants often mentioned that the patient's action was independent of the therapy provided. Follow-up studies could further investigate this by measuring baseline perceptions of the patient's likelihood to self-harm before therapy and comparing these against post-therapy judgments.

An unanticipated finding was that the analysis of empathy ratings revealed that the human therapist was perceived as more empathetic than their AI counterpart—despite delivering the same action—particularly in the high and medium empathy levels. This may reflect an underlying bias against AI systems' capacity to display empathy in emotionally charged contexts, which is further supported by open responses where participants stated that AI systems lack sentience and emotional capability. Such a bias might emerge from the AI's absence of anthropomorphic features, which have previously been shown to make stimuli appear more predictable and understandable, thereby possibly influencing the attributions of blame and foreseeability (Waytz et al., 2010). Interestingly, however, some open-ended responses did imply a degree of anthropomorphisation, as individuals referenced the AI agent's ability to 'listen'; to gain a more precise evaluation of the role anthropomorphisation plays in attributions of blame within mental health contexts, future studies could apply measures like Anthroscore to more systematically evaluate anthropomorphisation (Cheng et al., 2024). Measuring participants' initial attitudes toward AI and AI-based therapy could also provide further insights into how predispositions shape perceptions and moral judgments in human-AI interaction.

Acknowledgments

We would like to thank Ruby De Lanerolle for second coding responses for the qualitative analysis and anonymous reviewers for comments on a previous version of this manuscript.

References

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, *132*. https://doi.org/10.1016/j.ijmedinf.2019.103978

Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, *90*. https://doi.org/10.1016/j.chb.2018.08.028

Brailsford, J., Vetere, F., & Velloso, E. (2025). Responsibility attribution in human interactions with everyday ai systems. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706598.3713126

Chen, Y.-T., Tsai, H.-Y. S., & Yuan, C. W. (2024). Exploring how users attribute responsibilities across different stakeholders in human-ai interaction. *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, 202–208.

Cheng, M., Gligoric, K., Piccardi, T., & Jurafsky, D. (2024). Anthroscore: A computational linguistic measure of anthropomorphism. *"Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)"*. https://arxiv.org/abs/2402.02056

Cheung, V., Qiao, M. H., & Lagnado, D. (2024). Attribution of responsibility between agents in a causal chain of events. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Dewitt, S. H., Liefgreen, A., Adler, N., & Strittmatter, L. E. (2024). Open text box analysis: A pragmatic and reflexive guide. *PsyArXiv*. https://doi.org/10.31234/osf.io/7qsng

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. *Journal of Medical Ethics*, *47*. 10.1136/medethics-2020-106820

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*. 10.1126/science.aaw4399

Franklin, M., Ashton, H., Awad, E., & Lagnado, D. (2022). Causal framework of artificial autonomous agent responsibility. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 276–284. https://doi.org/10.1145/3514094.3534140

Franklin, M., Awad, E., Ashton, H., & Lagnado, D. (2023). Unpredictable robots elicit responsibility attributions. *Behavioral & Brain Sciences*, *46*.

Franklin, M., Awad, E., & Lagnado, D. (2021). Blaming automated vehicles in difficult situations. *Iscience*, *24*(4).

Franklin, M., Papakonstantinou, T., Chen, T., Fernandez-Basso, C., & Lagnado, D. (2023). Blame attribution in human-ai and human-only systems: Crowdsourcing judgments from twitter. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? from expectations to responsibility judgments. *Cognition*, *177*. https://doi.org/10.1016/j.cognition.2018.03.019

Gómez-Sánchez, J., Gordo, C., Franklin, M., Fernandez-Basso, C., & Lagnado, D. (2023). Who is to blame? responsibility attribution in ai systems vs human agents in the field of air crashes. *International conference on flexible query answering systems*, 256–264.

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C., & Franklin, M. (2023). A proposal for a definition of general purpose artificial intelligence systems. *Digital Society*, *2*. https://doi.org/10.1007/s44206-023-00036-9

Hidalgo, C. A., Orghian, D., Albo Canals, J., de Almeida, F., & Martin, N. (2021). *How humans judge machines*. MIT Press. https://doi.org/10.7551/mitpress/13373.001.0001

Hohenstein, J., & Jung, M. (2020). Ai as a moral crumple zone: The effects of ai-mediated communication on attribution and trust. *Computers in Human Behavior*, *106*. https://doi.org/10.1016/j.chb.2019.106190

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.

Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). Large language models produce responses perceived to be empathic. *arXiv preprint*. https://doi.org/10.48550/arXiv.2403.18148

Lenth, R. V. (2024). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 1.10.5-0900001, https://rvlenth.github.io/emmeans/]. https://rvlenth.github.io/emmeans/

Montemayor, C., Halpern, J., & Fairweather, A. (2021). In principle obstacles for empathic ai: Why we can't replace human empathy in healthcare. *AI SOCIETY*, *37*. https://doi.org/10.1007/s00146-021-01230-z

Ortiz-Viso, B., Papakonstantinou, T., Chen, T., Fernandez-Basso, C., & Lagnado, D. (2023). An unsupervised approach to extracting knowledge from the relationships between blame attribution on twitter. *Lecture Notes in Computer Science*, *14113*. 10.1007/978-3-031-42935-4_18

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*(4), 1057–1084.

Schoenherr, J. R., & Thomson, R. (2024). When ai fails, who do we blame? attributing responsibility in human-ai interactions. *IEEE Transactions on Technology and Society*.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). *Afex: Analysis of factorial experiments* [R package version 1.1-1]. https://CRAN.R-project.org/package=afex

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3). 10.1037/a0020240

Zhang, Q., Wallbridge, C. D., Jones, D. M., & Morgan, P. L. (2024). Public perception of autonomous vehicle capability determines judgment of blame and trust in road traffic accidents. *Transportation research part A: policy and practice*, *179*, 103887.