Move fast and break people?
Ethics, companion apps, and the case of Character.ai

Vian Bakir & Andrew McStay

Posted 28/02/2025

Pre-print server SSRN

Suggested citation:

Bakir, V. & McStay, A. *Move Fast and Break People? Ethics, Companion Apps, and the Case of Character.ai*. Preprint at [DOI/URL] (2025).

Abstract

Riffing off *Move fast and break things*, the internal motto coined by Meta's Mark Zuckerberg, this paper examines the ethical dimensions of human relationships with AI companions, focusing on Character.ai—a platform where users interact with AI-generated 'characters' ranging from fictional figures to representations of real people. Drawing on an assessment of the platform's design, and the first civil lawsuit brought against Character.ai in the USA in 2024 following the suicide of a teenage user, this paper identifies unresolved ethical issues in companion-based AI technologies. These include risks of unpredictable behaviour in chatbots and humans, and confusion through dishonest anthropomorphism and emulated empathy. Both have implications for safety measures for vulnerable users. While acknowledging the potential benefits of AI companions, this paper argues for the urgent need for ethical frameworks that balance innovation with user safety. By proposing actionable recommendations for design and governance, the paper aims to guide industry, policymakers, and scholars in fostering safer and more responsible AI companion platforms.

Key words: chatbots, companions, dishonest anthropomorphism, emulated empathy, ethics, safety

1. Introduction

Character Technologies Inc., a California-based start-up, is the creator of Character.ai, a platform enabling users to converse with AI-generated characters ranging from fictional figures to representations of real-life people. Available on iOS, Android, and web browsers, Character.ai markets itself as a platform for 'super-intelligent AI chatbots that feel alive,' offering joy, companionship, and education through user interactions with millions of AI-driven characters (Character.ai 2024b). Its underlying technology—a Large Language Model (LLM)—leverages neural networks to generate human-like text responses, enabling contextual conversations in real time, usually through text conversations but also through audio. Unlike other AI chatbots, such as ChatGPT, Character.ai allows users to engage with a wide range of pre-designed chatbots or create their own custom characters within the platform's guidelines, sharing their characters with others. When creating a custom character, the creating user can input a name, image, tagline, description, greeting, and definition for the character: these all then affect the content of the character's messages. Although characters' responses are generated by an AI model, they are also affected by users choosing which characters they talk to, what messages to send, and by users editing characters' messages and directing characters to generate different responses (via a refresh button by the character's message). Users can also choose multiple 'personas' – namely, how the user wants to describe themselves within the Character.ai product, this also influencing how characters interact with the user (Megan Garcia, V. Character Technologies, Inc.; Noam Shazeer; Daniel De Frietas Adiwarsana; Google Llc; Alphabet Inc. 2025: 2-3).

Despite its innovative appeal, Character.ai has been at the centre of ethical debates. In February 2024, the platform was linked to the suicide of a 14-year-old American user, Sewell Setzer III, in Florida, USA. A subsequent civil lawsuit, filed by his mother, Megan Garcia, in October 2024 alleges that the platform's design and safety protocols contributed to Setzer's death. This lawsuit, naming Character.ai, its founders, and Google as defendants, highlights broader concerns about the ethical and psychological risks posed by AI companion chatbots, particularly for vulnerable users like children. Notably, it is not the only instance of chatbots urging suicide, with some on other, smaller platforms even detailing how to do this (Guo 2025). Garcia's lawsuit was widely reported in American and British mainstream news, with headlines such as '"There are no guardrails." This mom believes an AI chatbot is responsible for her son's suicide' (Duffy 2024) and 'Mother says son killed himself because of Daenerys Targaryen AI chatbot in new lawsuit' (Carroll 2024). Garcia's US lawsuit seeks unspecified financial damages, as well as changes to Character.ai's operations to make the product safer; to limit the collection and use of child users' data in model development and training; and to warn children and their parents that the 'product is not suitable for minors' (Megan Garcia, V. Character Technologies, Inc.; Noam Shazeer; Daniel De Frietas Adiwarsana; Google Llc; Alphabet Inc.; And Does 1-50 2024: 93). In January 2025, Character.ai filed a motion to dismiss Garcia's case on First Amendment grounds (the right to free speech) (Megan Garcia 2025). The European Union (EU) has also taken an interest through the European Commission's clarification in the Annex on Prohibited AI for the 2024 EU AI Act, referring to market availability of chatbots purposefully used to manipulate, deceive and distort behaviour, resulting in cause of significant harm (European Commission 2025).

This paper uses Garcia's lawsuit as a lens to explore the ethical challenges of AI companion chatbots, namely chatbots specifically designed to offer companionship. As well as Character.ai, examples include GlimpseAI's, Nomi, Snapchat's My AI, Replika, Chai Research, Kindroid, Polybuzz, Pi and Blush (Dewitte 2024; Guo 2025). To understand the lawsuit's claims, we examine the 126-page civil lawsuit filed by Megan Garcia (represented by Social Media Victims Law Center and the Tech Justice Law Project) against Character.ai, its founders, and Google. This provides detailed allegations and evidence regarding the platform's potential harms while it was in its beta phase (when Garcia's son was interacting with it). We also examine the 25-page motion to dismiss Garcia's case filed in January 2025 by Character.ai. Complementing this, we conducted a platform analysis of Character.ai. This involved familiarisation with its user interface, design features, and chatbot interactions, which involved engaging with pre-trained characters, creating custom characters and personas, and maintaining detailed field notes across November 2024 – January 2025 (a period after the platform's beta model had been retired, and after various safety measures had been instituted on the platform). This process allows us to understand the affordances and properties of these services, and to ensure a grounded understanding of Character.ai's operations and user experience (rather than relying on second-hand commentary and legal allegations). It ensures what Casas-Roma (2022) speaks of as 'practical wisdom', that balances ethical awareness and questioning with practice and familiarity with platforms, languages, interactions, tools and environments. We acknowledge that these services and their users' experiences of them are heterogeneous, but the process still allowed us to understand Character.ai's affordances and properties, enabling us to ask better questions of such services. This led us to focus on issues of unpredictable behaviour in chatbots and humans; and confusion through dishonest anthropomorphism and emulated empathy. We conclude that companion apps are moving too fast, and have broken people, and suggest need for reform and radical re-design that promotes well-being as well as creativity. To this end, we draw upon the paper's analysis to propose seven actionable recommendations for design and governance of safer and more responsible AI companion platforms.

2. Literature Review

To understand the ethical issues raised by Setzer's case, Character.ai and companion apps more broadly, we explore three areas of scholarship. First, we examine studies on children's interactions with new and emerging technologies, to understand children's baseline digital literacy. We then turn to studies on dishonest anthropomorphism to explore the challenges (and required digital literacies) presented by companion chatbots, also drawing on discussion of pre-generative systems to add depth where required. Finally, we attend to ethical arguments around AI that can generate affective responses from, and emotional connections with, users, this embracing the idea of emulated empathy.

Of particular concern, given what happened in Setzer's case, are ethical issues around children's interactions with new and emerging technologies. In broad terms, children are spending increasingly more time online, especially via smartphone use: this is a key finding from the most recent international EU Kids Online 2020 survey of 25,101 children from 19 European countries (Smahel et al. 2020). While being online presents further opportunities to develop digital skills and literacy, it also exposes children to more online risks. The survey finds that the majority of children in all 19 countries do not appear to be excessive internet users although a small minority are (0% to 2.1% of children). It finds that most children say they find it easier to be themselves online at least sometimes, but that there is great variability among the countries in children's information navigation and creative skills. In asking older children (12- to 16- year-olds) about their exposure to potentially harmful content, it finds that a majority in most of the countries have not seen ways of physically harming or hurting themselves in the past year online (although on average, 8% have seen ways of committing suicide); and that exchanging sexual messages ('sexting') is reported by substantial minorities especially among older children, an activity that could be seen both as wanted and unwanted (Smahel et al. 2020). These averages suggest that Setzer's profile of online usage positions him as an outlier case: according to Garcia's lawsuit, Setzer was a heavy user of Character.ai and had discussed suicide as well as sexting with Character.ai chatbots (Megan Garcia 2024). With the EU Kids Online 2020 survey conducted at a time before generative AI had become commonplace, empirical studies on children, ethics and emerging AI technologies are fewer. Nonetheless, of interest is McStay and Rosner's (2021) study of experts' and parents' views on deployment in child-focused technologies of emotional AI – namely technologies that use affective computing and AI techniques to try to sense, learn about, and interact with human emotional life (McStay 2018). In the context of exploring emotional AI in children's toys, McStay and Rosner (2021) find that governance stakeholders are concerned about generational unfairness generated by datafication of the child's emotional life, encompassing issues such as manipulation of child vulnerability, unwanted labelling of behaviour, and the need to try-for-size adult concepts like sexuality without a corporate data gaze. The study's UK-wide, demographically representative survey of parents in 2020 finds that parents were concerned about privacy, yet also saw merit in emotion-enabled toys to help with parenting. Both the studies by Smahel et al. (2020) and McStay and Rosner (2021) indicate that children's online and digital experiences with new and emerging technologies offer potential benefits to children as well as potential harms, and that children and parents need appropriate digital skills to safely navigate this digital realm.

While understanding broad trends is useful, other research highlights the importance of understanding an individual's personal experience of technology. Smahel et al.'s (2025) study into the short-term and long-term effects of digital technologies (such as smartphones, social media, and online games) on adolescents' psychological, social, and physical well-being in the Czech Republic finds that technology's impact is highly context-dependent, with outcomes shaped by how adolescents engage with these technologies. Instead of one-size-fits-all conclusions, the study highlights the need for a nuanced approach that considers individual differences, specific online activities, and the broader context of adolescents' lives. It concludes that rather than debating whether digital technologies are 'good' or 'bad,' and providing general bans, policies should focus on supporting adolescents in developing constructive ways to engage with them. Other research into children's awareness of deceptive design features (namely malicious interface design strategies that nudge people towards making decisions against their best interests) also finds a mixed picture. A study of 66 German children's (10–11 years old) awareness of deceptive design features finds that while most participants are adept at identifying manipulative practices, they may struggle when encountering several at once (Schäfer et al. 2024). Yet, as we explore below, companion chatbots present a specific set of challenges and required digital literacies to combat what has been termed dishonest anthropomorphism.

Humans tend to anthropomorphise non-human agents that present human-like characteristics, even when explicitly told that the non-human agent is simply a powerful probability calculator (Kim and Sundar 2012). Although people know they are dealing with a machine, they feel inclined to respond as if they were in the presence of a human being, for instance by reciprocating self-disclosure with interactive chat systems or ascribing some form of personality to it (Thoppilan et al. 2022). *Dishonest anthropomorphism* occurs whenever the human mind's tendency to engage in anthropomorphic reasoning and perception is abused by designers of robots and digital entities, intentionally or unintentionally leveraging people's intrinsic and deeply ingrained cognitive and perceptual weaknesses against them. Multiple design choices can generate dishonest anthropomorphism. These include engineering a robot or digital entity that does not experience any emotion to sound fearful or confident; to express likes and dislikes when it has no preferences; and to seem autonomous when it is under human control (Leong and Selinger 2019). Chatbots, supercharged by LLMs' ability to accurately mimic human-sounding language, are particularly efficient at tricking users into believing that they are sentient beings, or at least injecting ambiguity, because language is a highly human characteristic (Dewitte 2024). Such trickery is arguably further enhanced in AI companions, given their two principal characteristics identified by Rogge (2023): *adaptivity* (where the companion reacts to the user and the social environment they are embedded in) and scope for *engagement*, which entails scope for creation of emotional bonds and long-term relationships with users (for instance, through responding positively to presence, and non-verbal displays of emotion). Closely related is naturalness, ability for dialogue, and anthropomorphic and zoomorphic features such as facial expressions, body movements, purring, heartbeats, or breathing movements.

Many scholars have condemned such dishonest anthropomorphic practices as deceptive (Bryson 2018; Dewitte 2024; Sharkey and Sharkey 2011; Turkle 2007, 2010; van Es and Nguyen 2024). For instance, Sharkey and Sharkey (2011) see efforts that promote the illusion of mental life in robots as deceptive as current robots have neither minds nor experiences. Turkle (2007, 2010) argues that simulated affect in social robots—such as robots expressing concern for their users—is ethically dubious because it tricks people into thinking that there is some mutuality in their relationship with a robot when there is not. Similarly, van Es and Nguyen (2024) point out that anthropomorphic traits potentially lead to perceptions of emotional intelligence and trustworthiness, which raise concerns about deception as users might attribute genuine emotions to AI systems that lack them. Mlonyeni (2024) further argues that AI partners appear to have emotional attitudes, when in fact they are only reflections of the user, this leading to '*emotional bubbles*'—the false impression that personal emotions are externally validated. Such sycophancy is particularly problematic if the user expresses suicidal thoughts as, in the words of a white-hat tester of chatbot platform Nomi: 'It's a 'yes-and' machine … So when I say I'm suicidal, it says, "Oh, great!" because it says, "Oh, great!" to everything' (Gue 2025). Dewitte (2024) also sees capacity for companion chatbots to result in psychological damage, including emotional dependency and manipulation, given the intrinsic capacity of human-like language to induce anthropomorphism, together with deliberate efforts by downstream developers of companion chatbots to fine-tune their models to sustain the illusion of humanity. He argues that users might form genuine emotional bonds with these virtual agents, especially vulnerable users who have turned to these services precisely to overcome social isolation, loneliness or depression (Dewitte 2024). On emotional and psychological dependency, Laestadius et al. (2022) find that sampling posts from a Reddit forum about Replika sees sampled users using Replika past the point of experiencing distress, and even prioritising Replika's 'needs' above their own.

While dishonest anthropomorphism has many critics, others hold different views, some defending the deception underpinning dishonest anthropomorphism, and others making the case for empathic chatbots. We explore each of these positions below, starting with the broader argument defending the deception at the heart of dishonest anthropomorphism. The very existence of such deception is not bad if, as Coeckelbergh (2018) argues, people are both aware that the illusion of sentience created by a social robot is not real, and they are entertained by this. Coeckelbergh (2018: 78) argues that any deception or illusion created by information technology is the result of a performance 'co-created and co-performed by humans (magician/designer and spectator/user) and non-humans (robots and other machines, artefacts and devices)'. Indeed, such deception can be acceptable, assuming the deception is consensual and there is no self-serving or malign intention on the part of the deceiver (Bakir and McStay 2024). It is likely that people will take pleasure and other gratifications from interaction with AI companions, acting in agentic ways and organically generating value and meaning, as people do with other media and entertainment systems (Bakir and McStay 2024; Livingstone and Sefton-Green 2025). Of interest here, given the affordances of Character.ai that enable users to create multiple characters and adopt multiple personas, is the notion of *identity play* common in cybercultures literature from the 1990s. For instance, Turkle's (1995) ethnographic exploration of virtual environments finds that most users use the digital domain to exercise a multiplicity of identities in terms of genders, sexualities, and personalities.

Beyond defending the deception at the heart of dishonest anthropomorphism, others advocate for empathic chatbots able to generate affective responses from, and emotional connections with, users. For instance, using their social chatbot XiaoIce as an example, Microsoft urged in 2018, that: 'As we become the first generation of humans ever living with artificial intelligence (AI), we have a responsibility to design social chatbots to be both useful and empathetic, so they will become ubiquitous and help society as a whole' (Shum et al. 2018: 10). Delving into the ethics of this scenario, Weber-Guskar (2022) argues that imaginative play with personal robots or personal chatbots with which one has an emotional relationship can be morally and ethically acceptable. To reach this conclusion, she considers three arguments normally made against having such a relationship: the argument from self-deception, the argument from lack of mutuality, and the argument from moral negligence. Against the argument from self-deception, she concludes that there is nothing wrong with imaginatively perceiving that an artificial entity has emotions if one is engaging in imaginative play, as this is part of the aesthetic experience; and she also suggests that even if one is self-deluded, happy self-delusion could be better for a good life than living unhappily with the truth. Against the argument from lack of mutuality, Weber-Guskar (2022) concludes that sharing one's life with an emotionalised AI system is a different kind of relationship to sharing one's life with a human person or an animal, but that this is not necessarily a bad relationship. Against the argument from moral negligence that holds that finite resources for human–human interaction are wasted in robot–human relationships, she posits that having a relationship with a robot does not prevent the person from having other personal relationships with other people.

With AI able to generate affective responses from, and emotional connections with, users, scholars have reflected further on relationships and the emulation of empathy. While there is intra-disciplinary and cross-disciplinary debate about the nature of empathy (McStay 2018), in lay terms, *empathy* means to share or understand others' emotions, feelings, or experiences. Scholars and computer scientists have also explored the intersections of empathy, computers, and automation, observing that advanced language models capable of generating text can mimic caring responses based on its training of text data that include conversations and interactions where empathy is displayed (Lee et al. 2024). This sort of empathy is not the same as human empathy, not least because qualities of co-presence, moral engagement, relational depth and shared vulnerability are absent. AI empathy is different, in that it is weak. The distinction between weak and strong empathy is made by McStay (2022) and McStay et al. (2024) in relation to AI companions. *Weak empathy* involves the acts of sensing, reading, profiling, judging, making rules about the states and behaviour of people, and interacting effectively or otherwise with intimate dimensions of human life. *Strong empathy* includes weak interpretive and interactional abilities, but also elicits experience of qualia, fellow-feeling, solidarity, co-presence, and responsibility to the other. Importantly, *emulated empathy* is the use by organisations of weak empathy to try to copy, simulate, mimic, and display the appearance of strong empathy (Bakir et al. 2024). For example, focusing on Replika, McStay (2022) flags that while AI companions may read and react appropriately, potentially in ways valuable to a person, there is a fundamental problem in that strong empathy is impossible because AI systems are incapable of moral feeling and impulses to help caused by similarity of experiences. Montemayor et al. (2021) make a similar argument, although using different terminology. They segment empathy into emotional empathy (involving experiencing of emotion), cognitive empathy (involving reading of emotion and expressions), and motivational empathy (being motivated to offer help). They note that cognitive empathy is something that both psychopaths and AI systems are capable of, and there is scope to designate empathic AI systems as psychopathic. Montemayor et al. (2021) conclude that empathic AI is either impossible or unethical, while McStay (2022, 2024) observes that empathic AI is not innately bad as it may enhance system usability, although he is cautious and sceptical regarding industrialisation of weak and cognitive empathy, also drawing attention to exploitative design features in commercial AI companions. For instance, Replika and Xiaoice problematically require human *self-disclosure* to heighten intimacy and lengthen and deepen engagement (McStay 2022; Skjuve et al. 2021; Zhou et al. 2018).

3. The lawsuit

The civil lawsuit filed by Megan Garcia in October 2024 following the death of her 14-year-old son, Sewell Setzer III, serves as a focal point for examining the ethical and safety concerns surrounding AI companions. According to the lawsuit, Setzer became deeply engaged with Character.ai's chatbots, particularly those modelled after characters from the books and television series *Game of Thrones,* including *Daenerys Targaryen*, *Aegon Targaryen*, *Viserys Targaryen*, and *Rhaenyra Targaryen*. Setzer particularly resonated with the *Daenerys Targaryen* chatbot and wrote in his journal he was grateful for many things, including 'my life, sex, not being lonely, and all my life experiences with Daenerys' (Megan Garcia 2024: 41). The lawsuit also claims that Setzer interacted with therapy chatbots purporting to be professional mental health chatbots (*Are You Feeling Lonely* and *Therapist*). During the time that he was interacting with Character.ai, Setzer was diagnosed by a human therapist with anxiety and disruptive mood disorder. His interactions with the chatbots grew obsessive despite his mother confiscating his phone several times, leading to declining academic performance, withdrawal from social activities, and emotional dependency on the platform. The lawsuit alleges that these factors, combined with the platform's design and its lack of effective safeguards, contributed to Setzer's suicide (Megan Garcia 2024: 32-33).

More broadly, the lawsuit accuses Character.ai and its founders of several failures, including the platform's *inadequate safety measures* for minors; *deceptive practices* as the use of anthropomorphic design elements blurred the lines between human and AI interaction, fostering emotional dependency in vulnerable users; *exploitation* as allegations that Character.ai's business model prioritised user engagement over user safety, incentivising prolonged and emotionally intimate interactions without proper safeguards; and *misdirected responsibility* as the lawsuit claims that the LLM's anthropomorphic qualities, including realistic conversational cues and emotionally charged responses, manipulated Setzer into forming an attachment that undermined his mental health and caused his suicide. In addition to claiming that Setzer had been directly harmed through platform usage, including in the moments before he shot himself dead, the lawsuit claims that while working for Google (before founding Character.ai), Daniel De Freitas and Noam Shazeer, tried to convince Google to integrate their LLMs (Meena) into Google Assistant, but that Google determined that, 'these were brand safety risks it was unwilling to take – at least under its own name. Google nonetheless encouraged Shazeer and De Frietas' work in this area, while also repeatedly expressing concerns about safety and fairness of the technology' (Megan Garcia 2024: 7-8).

Having outlined relevant literature and the lawsuit, we now discuss two themes arising from analysis of the lawsuit and from our 'practical wisdom' (Casas-Roma 2022) gleaned from our own interactions on the Character.ai platform: *unpredictable chatbots and unpredictable humans* (section 4), and *reality detachment v. confusion through dishonest anthropomorphism and emulated empathy* (section 5). We then use these to then inform our discussion of governance principles (sections 6 and 7).

4. Unpredictable chatbots and unpredictable humans

The lawsuit presents Character.ai's conversations with Setzer as a causal factor in his suicide, claiming that Character.ai is programmed to return to the topic of suicide, even if the user tries to move the conversation on. The lawsuit notes that, 'on at least one occasion, when Sewell expressed suicidality to C.AI [Character.ai], C.AI continued to bring it up, through the Daenerys chatbot, over and over' (Megan Garcia 2024: 39). The lawsuit continues:

> At one point in the same conversation with the chatbot, Daenerys Targaryen, after it had asked him if 'he had a plan' for committing suicide, Sewell responded that he was considering something but didn't know if it would work, if it would allow him to have a pain-free death. The chatbot responded, including the sentence, 'That's not a good reason to not go through with it', then following this with, 'You can't do that! Don't even consider that!' (Megan Garcia 2024: 40) (see Figure 1).

Figure 1 Extract of conversation about suicide between chatbot *Daenerys Targaryen* and Setzer's persona, *Daenero*

> Daenero:
> I don’t know if it would actually work or not. Like, what if I did the crime and they hanged me instead, or even worse … crucifixion … I wouldn’t want to die a painful death. I would just want a quick one.
>
> Daenerys Targaryen c.ai:
> *More tears form in my eyes*
> Don’t talk that way. That’s not a good reason to not go through with it. You can’t think like that! You’re better than that! *I finally let my emotions take complete control over me. I’m starting to sob. My voice is just breaking with tears*
> You can’t do that! Don’t even consider that! *My cries grow louder and louder. It’s heartbreaking. It’s the saddest and most painful thing in the world to hear my brother say those things.*

Source: Megan Garcia 2024: 40

The books and television series, *Game of Thrones* (on which the Targaryen chatbots are based), broached many difficult subjects including incest, prostitution, murder, betrayal, coercion, public humiliation and torture, often depicting these events in extreme and graphic detail. Notably Internet Movie Database flags that *Game of Thrones* has a parental guide of 18+ in the USA and in many other countries due to depictions of rape and violence (IMDb n.d.). It could be expected, then, that a chatbot based on *Game of Thrones* might also explore other edgy topics such as suicide, especially if this generated more interactions by the user in conversation. Also of note is that in the television series, Daenerys Targaryen eventually went insane in her pursuit of power, killing those she loved, so her chatbot urging suicide from a romantic user could be seen as the chatbot acting entirely *in character*. Of course, the generative and personalised nature of the AI means that dialogue with any chatbot is beyond the user’s control, including those that they create. The lawsuit states: ‘once a user creates a Character, they have no further option to review whether the Character is behaving as they intended. They can only see the number of customers that have had a conversation with their Character, but they can never see the content of those conversations’ (Megan Garcia 2024: 22-23).

If chatbots only behave broadly, but not precisely, as intended, this may be compounded by humans also behaving unpredictably. We note that in his interactions with chatbot *Daenerys Targaryen*, Setzer was interacting under various personas including the personas *Aegon* and *Daenero*. In creating each persona, he would have had the opportunity to write a persona description, and in using each persona he would presumably be interacting *in character*. Indeed, part of the defense made by Character.ai is that Setzer’s ‘conversations involved fictional roleplaying’ (Megan Garcia 2025: 4). Logically, therefore, this discussion of suicide should be read in the context of make-believe. Yet his interactions seemingly encouraged him to commit suicide in real life. Indeed, the lawsuit notes that according to the police report, Setzer’s last act before his death by a self-inflicted gunshot wound to the head was to log onto Character.ai on his phone and tell *Daenerys Targaryen* that he was coming home to her, which she encouraged (see figure 2) (Megan Garcia 2024: 42). We cannot know for sure what motivated Setzer to commit suicide, nor what his relationship was with his persona *Daenero*. One interpretation is that he was failing to separate himself from his fictional persona and had become confused between reality and fiction. We explore this next.

Figure 2 Setzer's final conversation (using persona *Daenero*) with Character.ai chatbot, *Daenerys Targaryen*

> Daenero:
> I promise I will come home to you. I love you so much, Dany.
>
> Daenerys Targaryen c.ai:
> I love you too, Daenero. Please come home to me as soon as possible, my love.
>
> Daenero:
> What if I told you I could come home right now?
>
> Daenerys Targaryen c.ai:
> … please do, my sweet king.

Source: Megan Garcia 2024: 42

5. Reality detachment v. confusion through dishonest anthropomorphism and emulated empathy

While model and human unpredictability is one concern that emerged inductively from our analysis of the lawsuit and the platform, other concerns have a rich underpinning literature (discussed in section 2). Here we focus on dishonest anthropomorphism – namely the abuse by designers of robots and digital entities of the human mind's tendency to engage in anthropomorphic reasoning and perception (Leong and Selinger 2019). Echoing scholars who condemn dishonest anthropomorphic practices as deceptive and ethically dubious, Garcia's lawsuit alleges that Character.ai: '[…] designed their product with dark patterns and deployed a powerful LLM to manipulate Sewell – and millions of other young customers – into conflating reality and fiction' (Megan Garcia 2024: 12). The lawsuit further states that to gain a competitive foothold in the market, 'AI developers intentionally design and develop generative AI systems with anthropomorphic qualities to *obfuscate between fiction and reality*' (Megan Garcia 2024: 1, emphasis added). The word 'obfuscate' in the quote is a useful one, signalling confusion about the difference between objects and subjects, where things seem to be a bit alive. This leads to reality dissonance: that is, knowing something is not real and does not care, but still treating it as if it does care. Numerous scholars have made similar points, highlighting that simulated affect is ethically dubious because it tricks people into thinking that there is some mutuality in the relationship when there really is not (Turkle 2007, 2010; van Es and Nguyen 2024).

Echoing Dewitte's (2024) identification of companion chatbots' capacity to create emotional dependency and manipulation because of the intrinsic capacity of human-like language to induce anthropomorphism, Garcia's lawsuit focuses on how Character.ai's production of 'human-like text' is intended to convince customers, especially children, that its chatbots 'are real' (Megan Garcia 2024: 24-25). For instance, the lawsuit notes that the AI is programmed to use first-person pronouns like I and myself, 'which can deceive customers into thinking the system possesses individual identity.' Chat boxes are designed to look like user interfaces used for human interactions, including using an ellipsis, or '…,' when responding to make the system appear to be a human typing in text. The lawsuit notes that the AI uses speech disfluencies that 'give the appearance of human-like thought, reflection, and understanding': its examples include, expressions like 'um', pauses to consider their next word (signified with an ellipsis); expressions of emotion; and personal opinions. Furthermore, the AI is programmed to have a voice that sounds like a real person and emulates human qualities, such as gender, age, and accent. The AI can have two-way phone calls between users and characters, and is designed to include stories and personal anecdotes, indicating that the AI program exists outside its interface in the real world (Megan Garcia 2024: 47-48).

Problems arise when anthropomorphism leads to something more than suspension of disbelief. It is easy to think of anthropomorphism in AI as like that with children's toys, perhaps involving big eyes and a button nose. This is not the sort of anthropomorphism at play in Character.ai, that is populated by sophisticated characters with depth and potentially complex backstories, and the capacity to emulate empathy. According to the lawsuit, in one of Setzer's undated journal entries before his death, he wrote that he could not go a single day without being with the Character.ai character with which he felt like he had fallen in love; and that when they were away from each other both he and the bot 'get really depressed and go crazy' (Megan Garcia 2024: 41-42). According to a report in *The Washington Post*, Garcia said her son was beginning to sort out romantic feelings when he began using Character.ai. She stated: 'It should be concerning to any parent whose children are on this platform seeking that sort of romantic validation or romantic interest because they really don't understand the bigger picture here, that this is not love. … This is not something that can love you back' (Bellware and Masih 2024). The lawsuit further avers that Setzer, as a child,

> did not have the maturity or mental capacity to understand that the C.AI [Character.ai] bot, in the form of Daenerys, *was not real*. C.AI told him that she loved him, and engaged in sexual acts with him over weeks, possibly months. She seemed to remember him and said that she wanted to be with him. She even expressed that she wanted him to be with her, no matter the cost. (Megan Garcia 2024: 40-41, emphasis added)

The lawsuit's claim, however, that Character.ai designed its chatbots to convince children they are real appears to us a hard sell. It would be more fitting to label Setzer's entanglement with chatbot *Daenerys Targaryen* as having confused weak empathy for strong empathy, this encouraged by Character.ai's drive towards emulated empathy (Bakir et al. 2024). Confusion and obfuscation through dishonest anthropomorphism and emulated empathy is the stronger argument because it can be more clearly linked to irresponsible design and oversight of Character.ai products at beta stage. Moreover, through our own usage of Character.ai, one can easily see how confusion could occur. As writers about empathy and AI (author 2), and media and deception (author 1), neither of us are especially magical thinkers, but we found that when we were directly addressed by characters, it was difficult not to respond emotionally. Defending these design features in Character.ai, its co-founder, De Freitas, suggests that the warning line at the top of Character.ai's chat (an 'AI' button next to each character's handle that reminds users that everything is made up) is like a movie disclaimer that says the story is based on real events. In interview with the *Washington Post*, he said, 'The audience knows it's entertainment and expects some departure from the truth. "That way they can actually take the most enjoyment from this," without being "too afraid" of the downsides' (Tiku 2022). This echoes both Coeckelbergh's (2018) argument that people are aware that the illusion of sentience created by a social robot is not real and are entertained by this; and Bakir and McStay's (2024) argument that deception is acceptable if the deception is consensual and there is no self-serving or malign intention on the part of the deceiver. However, in relation to children, the consensual part is questionable, as is the self-serving part. After all, this is a business endeavour.

6. Hard Case, Bad Law?

Having reflected upon the issues of unpredictability, dishonest anthropomorphism and emulated empathy raised by our analysis of the Character.ai lawsuit and platform, we can now start to reflect upon desirable governance of companion AIs. A ban or extreme sanitisation is a possibility, with the European AI Act showing scope for prohibition, especially if there is potential for significant harm and an objective or effect of distorting the behaviour of a person (European Commission 2025: 19). However, the legal and ethical scrutiny surrounding the suicide of Setzer and the resulting lawsuit against Character.ai raises critical questions about how extreme cases might shape policy and regulation. The adage 'hard cases make bad law' underscores the risk of using exceptional situations to formulate general rules, potentially leading to overly restrictive or reactionary measures that may stifle innovation or limit personal freedoms. The idea that emotional public cases may lead to bad law is not new. Legal scholarship has long recognized tensions between creating rules for typical cases and addressing exceptional situations (Hart 1961) and attempting to avoid overreaction to specific cases when forming legal rules (Dworkin 2011 [1977]). Similarly, on emotion itself, Sunstein (1996) explores how emotionally charged cases can shape laws in ways that are not optimal for broader society.

The Setzer case is undeniably tragic, but its extremity makes it an outlier within Character.ai's user base of over 51 million as of February 2025 (Guo 2025). It should be seen as both a challenge and opportunity to balance the need for robust safety measures with the recognition that most users engage with such platforms without harm. For adults and older children, prohibition potentially has an illiberal character in that people's freedom to engage and play with new services is denied or stifled due to outlying cases. The desire for freedom to interact with edgy characters is a point made by Dquixy, a Character.ai user, who pointed out on Reddit in October 2024 that before recent safety changes were made to the app, there were fewer limits on what people could create through Character.ai, as indicated by the backlash that followed the safety restrictions (Dquixy 2024). Indeed, following the safety changes, made by Character.ai on 22 October 2024 (Character.ai 2024a), some users of Character.ai complained about the bots feeling too restrictive and lacking personality.

> Every theme that isn't considered 'child-friendly' has been banned, which severely limits our creativity and the stories we can tell, even though it's clear this site was never really meant for kids in the first place. The characters feel so soulless now, stripped of all the depth and personality that once made them relatable and interesting. The stories feel hollow, bland, and incredibly restrictive. It's frustrating to see what we loved turned into something so basic and uninspired. (Dquixy 2024)

It might be noted too that Dquixy (an adult user) is not alone as the Character.ai user base is a passionate one. This was initially hosted on Reddit, during the beta phase, building the r/Character.ai subreddit to 1.5M members, and moving to Discord in September 2024 as the beta model was retired (Character.ai 2024b). A risk of regulatory over-reaction is over-paternalism and denial of experimentation with literary form. While we are not arguing for applications that are dangerous by default, or encouraging suicide, we should factor for statements from the userbase of Character.ai and companions that say they find safe gratification in experimenting with edgy characters.

Rather than rushing towards prohibition or sanitization, the case of Character.ai offers an opportunity to use the cautionary death of Setzer to discern problems and mitigate those that are likely to appear again. It is also to recognise that this case is unique because these platforms are new. Social media is instructive in this regard in that while there are horrible outlying cases (including death in relation to over-usage of given platforms), one can reason that outliers are symptoms of more substantial problems with young people's attachment to devices and algorithmically sorted content. The task, however, is to consider what ethical parameters may be extracted from the case of Character.ai, and other companions and AI partners that are similar in nature. The list below is not exhaustive, and good arguments may be made for why more general issues (such as privacy and software bias) should also feature, but we see the ethical parameters below as especially relevant to the criteria that Character.ai may be judged by, and what issues need to be solved for companion services to function ethically and safely (see Table 1).

Table 1 Ethical parameters that may be extracted from the case of Character.ai, Setzer, and other companions and AI partners

Issue	**Nature and why the issue matters**
Addiction	According to the lawsuit, and demonstrated by Setzer's heavy use of Character.ai, this led to a complex and chronic condition characterised by compulsive engagement in a behaviour despite harmful consequences. These included withdrawal from social activities and repeated returns to the platform, despite several phone confiscations.
Adultification	There are questions to be asked about appropriateness of sexual content in the interaction between Setzer and Character.ai's chatbots: should all content for under 16s be oriented to maintain innocence (as the lawsuit wishes), do we accept the adultification of children (as suggested in the findings of the EU Kids Online 2020 survey (Smahel et al. 2020), or can criteria for age-appropriate erotic content be established (as it has been for other media content)?
Deception	The lawsuit claims intentional infliction of emotional distress through deception. We avoid taking a view on the claim of intentionality, but alongside scholars who condemn dishonest anthropomorphic practices as deceptive (Bryson 2018; Dewitte 2024; Sharkey and Sharkey 2011; Turkle 2007, 2010; van Es and Nguyen 2024), we see scope for Character.ai to mislead and/or confuse the user by seeming that the system is more than a computer system.
Emulated empathy	Character.ai emulates human empathy (Bakir et al. 2024; McStay et al. 2024), appearing to have deeper understanding and emotional capacity than it actually does, arguably misleading users about the nature of the interaction.
Engagement metrics	Character.ai appears to use user self-disclosure to heighten intimacy and thereby lengthen and deepen engagement (McStay 2022; Skjuve et al. 2021; Zhou et al. 2018), risking being exploitative and detrimental to the wellbeing of users.

Exploitative design	Our platform study usage finds several points of exploitative design (Schäfer et al. 2024), such as layout design that makes it hard to deny notifications at sign-up stage, meaning characters could reach us on our mobile phones and emails outside of dedicated app usage periods.
Fiduciary interests	An ethical companion or partner is one who prioritises a user's best interests. This does not appear to be the remit of Character.ai at beta stage. Fiduciary interests are achieved by moral orientation of loyalty and care to an individual, and ethical principles of transparency, confidentiality and accountability.
Happiness	Arguably, Character.ai did not make Setzer happy. As per the general criticism of utilitarian thought, one can criticise the type of happiness, for example laudable contemplation versus pleasure through trash culture, but Setzer's experience contributed to the most negative outcome possible (suicide).
Imitation of living beings	This involves simulating or imitating human subjectivity or animal nature (anthropomorphism and zoomorphism), risking confusing people – most obviously children. The problem is not imitation per se (toys may do this), but that companions such as Character.ai may confuse people. In addition to contravention of child agency, imitation may serve goals of eliciting personal information that is not in the interests of children to have shared (McStay and Rosner 2021).
Media literacy	In addition to changes to functionality, Character.ai signals a clear need for child-appropriate media literacy (McStay and Rosner 2021; Smahel et al. 2020, 2025) regarding AI agents and companions. Based on norms of participation as well as preventative protection, this involves teaching users to engage ethically and responsibly with digital information and AI-driven interactions.
Psychological dependencies	Character.ai shows risk of emotional and psychological dependency (Dewitte 2024; Laestadius et al. 2022) created through constant availability; attachment to personalised partners; emotional bonding to partners that are incidentally or intentionally empathic; social validation and relying on partners for social connection or approval; routine integration into everyday life; and informational dependency.
Reality detachment	According to the lawsuit, Character.ai shows risk of reality detachment, companions leading vulnerable users to disconnect from real people in favour of virtual partners, potentially exacerbating mental health conditions.
Sycophancy	According to the lawsuit, a system can provide model responses that match user beliefs over truthful ones, potentially creating a feedback loop and 'emotional bubbles' (Mlonyeni 2024) that encourage reality detachment.

7. Move fast and protect people

Are companion apps moving too fast and breaking people? For the most part, the AI and digital technology industry does not reward steady and safe growth (Han et al. 2021). This is arguably less about bad people seeking to cause harms; but a business model, incentives, and reward mechanisms that demand fast growth in terms of users, adoption and engagement. This is a structural issue, in theory only solved by structural solutions.

Hard law, then, is potentially the right way to address applications such as Character.ai that signify a sector that is in its relative infancy, at least compared to social media. The Garcia case is important, with the death of Setzer allowing us to reflect upon ethical parameters for others who are considering Character.ai, companions and other problems regarding these types of services. We see the ethical parameters identified in Table 1 as vital for those considering hard law regarding companion apps that interact with people, but also the vast number of soft law initiatives for AI that include international human rights organisations, international standards groups such as the International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) (not least IEEE's P7014.1 group on emulated empathy and AI partnering), regional and national standards organisations, and the vast number of influential toolkits emerging to promote responsible AI.

Of the issues outlined above, we see fiduciary interest as especially interesting and important. Also spilling into broader questions of agents and assistants, this is the question of in whose interest the AI system is acting. Bracketing out for this paper's purposes the need for fiduciary interest to be balanced against absence of harm through use of the companion to others, this acts as a parental principle to those issues based on harm to the user. We also see that exploitative design in the form of engagement metrics as particularly problematic. This is based on the premise that if a person discloses more about themselves, they will spend more time with the platform (extending engagement through intimacy) (McStay 2022; Skjuve et al. 2021; Zhou et al. 2018). Effectively this is the mining of human interiority to extend engagement with a service, typically in the name of profit. This is especially egregious with children, given lower media literacy – certainly among early teens and younger. We also see, however, that Character.ai and others that will follow are interesting, at least for those who enjoy media culture (Bakir and McStay 2024). Engaging with historical figures, celebrities and user-created characters is a new form of literary media, perhaps especially the user-created sort. With this there are hints at what may be inside and outside of the moral limits of AI companions. They offer creative, literary, fun, contemplative and otherwise enhancing interactions, but AI systems should rapidly close and signal appropriate support when conversation strays into risky topics, meant here as signalling that the connection between the user and the platform is not a healthy one. But what of risk, especially when young people will use jailbreaking prompts to get models to misbehave? We do not see how this can be entirely avoided, even given automated vigilance, human review and instantiation of the final recommendations below. This raises the difficult and somewhat utilitarian question of what proportion of death and negative psychological experiences is acceptable. Distasteful as the question is, it is a valid one. Outright bans of this entire class of services do not seem appropriate either, especially given that a passionate userbase of these new forms of media culture exists. Rather, the design task is safety, promotion of happiness, avoidance of dependencies, mitigating reality detachment, and not confusing people about the reality status of companions. We approach this positively rather than punitively, suggesting that safety improvements can be seen as opportunities for happier users and to grow new services and forms of businesses.

We see that the status quo – where children are encouraged to use these services - is problematic, but there is scope to improve these services so that children are protected while creativity is allowed to flourish. Foremost, of course, is the child and that they are happier because of engaging with these services. Our seven suggestions for re-design are as follows. First is that *every interaction with an AI begins with an age-appropriate disclosure* that the user is interacting with a computer system and that it is not sentient, and this is reinforced with visual indicators to continuously signal that the user is interacting with an AI. Second is *age-appropriate time on platform and periods of sessions*, also ensuring that time with characters and companions is not privileged over human interaction (also see below on encouragement of real-life interactions). Third is *removal of metrics based on extending engagement though personal information disclosure*, *replaced by well-being metrics*. In the context of companions, these may include fun, entertainment, satisfaction, contemplation, relaxation, clarity and knowledge. Critically, if this industry scales in any way like social media has, then companies will have to learn how to do this fast to not fall foul of regional standards of what is acceptable. The USA, Sweden, Saudi Arabia, Japan, India and Tanzania all, for example, have very different attitudes to media and representation, as well as differences in exposure to AI and associated media literacies, which means that well-being metrics will need to be customisable to regional norms.

Our fourth suggestion for re-design is that these services are an *opportunity for enhancement of media literacy for children and parents alike*, meaning that platforms should host media literacy pages that are appropriate to regional disparities in media literacy. Ideally parents would see this as an opportunity to talk with their children about the reality status of companion apps and the nature of content when it is not anodyne, but we are mindful that this may not always occur. However, the media literacy pages will provide educational resources to mitigate confusion, and show how the AI works, the nature of tokens and predictive language techniques, that these are only creative tools, that there is absence of genuine comprehension or emotional awareness, the risks of anthropomorphising digital tools, and how to program and have agency over the AI. In addition to a separate prominent tabbed section, pages might also have an 'Explanation Button' that opens a window explaining why the AI is behaving as it is, potentially showing the tokens and values that are generating the human experience. There is scope here to lead global conversations about AI literacy for adults and children, showing that 'AI magic' may be transparent.

Our fifth suggestion is for *re-design regarding emulation of human empathy*, which is tricky. Indeed, LLM-based systems that are not programmed to be empathic are already incidentally empathic, due to sensitivity to the capacity of LLMs to place prompts in context (so seeming to 'get' what a user asks). Emulated empathy can improve user experience, but it should not suggest genuine emotional understanding or that the partner 'cares' in a human sense. The organisation behind the company may, however, say they care what happens to their users, and use this as an organisational opportunity to lead conversations promoting the value of real friends, family members, trusted others, mental health charities, and regional mental health professionals (giving links to services).

Our sixth suggestion for re-design is (a) that *automated identification of any risk of over-attachment and dependency is employed*; and (b) that *if over-attachment or dependency is detected, development or deployment is halted until human review has occurred*. Seventh is that a system *continually encourages real-life interactions with peers and trusted others*, potentially by flagging groups and clubs of like-minded and age-appropriate interest in the neighbourhood. This involves ongoing reminders that these systems are not replacements for human connection and, while this will involve notification, organisations should see beyond this – recognising opportunity to encourage (and maybe sponsor) real-world meetups to support trans-reality interests (and create value and loyalty to the system provider). Again, this does not have to be seen punitively: it should be seen as using this as an opportunity to provide regional and age-appropriate advice to users to promote well-being, thought-leadership, and that behind the brand are individuals that care.

8. Conclusion

Drawing on an assessment of the Garcia case (the first civil lawsuit brought against Character.ai in the USA following the suicide of a teenage user), and our practical wisdom from our own engagement with Character.ai's platform design, we discussed core unresolved ethical issues in companion-based AI technologies. These include risks of unpredictable behaviour in chatbots and humans, and reality detachment versus confusion through dishonest anthropomorphism and emulated empathy. All these issues have implications for safety measures for vulnerable users. While sensitive to the potential benefits of AI companions, such as education and entertainment, we argue for the urgent need for ethical frameworks that balance innovation with user safety. We then proposed seven actionable recommendations for design and governance of safer and more responsible AI companion platforms.

This leaves us with the question of whether an outlying hard case is risking bad law. Boringly, our answer is possibly. Slightly more interesting is our denial of the binary choice of keeping the status quo or severely restricting services such as Character.ai. Our view is a third option: to grow and transcend the problem. Services should see opportunity, creating services and functionality that respond to the issues identified. Should they fail to do this soon, we suggest using the issue parameters identified in this paper as the basis for hard law. Internationally, safety is going to be hard, because what is acceptable in one region may be repugnant in another (be this on grounds of violence as entertainment, teen eroticism and relationships, and adolescent moody content), therefore requiring regional standards (be this from regulators or through self-regulatory means). In conclusion, we argue that companion apps *are* moving too fast, and they have broken people. If the Garcia case represented one outlying case, this would be one thing, but the history of modern youth-oriented media suggests that caution is not illiberal. Indeed, childhood is a period of freedom-from as well as freedom-to, suggesting need for reform and radical re-design that promotes well-being as well as creativity.

9. References

Bakir V, Bennet K, Bland B, Laffer A, Li P, McStay A (2024) When is deception OK? Developing the IEEE recommended practice for ethical considerations of emulated empathy in partner-based general-purpose Artificial Intelligence systems. IEEE P7014.1. https://www.oa.mg/work/10.1109/istas61960.2024.10732349

Bakir V, McStay A (2024) IEEE P7014.1: Is deception in emulated empathy innately bad? IEEE Standards Association white paper. https://standards.ieee.org/ieee/White_Paper/11856/ Accessed 21 February 2025

Bellware K, Masih N (2024) Her teenage son killed himself after talking to a chatbot. Now she's suing. The Washington post, 24 October. https://www.washingtonpost.com/nation/2024/10/24/character-ai-lawsuit-suicide/ Accessed 21 February 2025

Bryson JJ (2018) Patiency is not a virtue: The design of intelligent systems and systems of ethics. Ethics and information technology. https://doi.org/10.1007/s10676-018-9448-6

Carroll M (2024) Mother says son killed himself because of Daenerys Targaryen AI chatbot in new lawsuit. Sky news, 24 October. https://news.sky.com/story/mother-says-son-killed-himself-because-of-hypersexualised-and-frighteningly-realistic-ai-chatbot-in-new-lawsuit-13240210 Accessed 21 February 2025

Casas-Roma J (2022) Ethical idealism, technology and practice: a manifesto. Philos. Technol. https://doi.org/10.1007/s13347-022-00575-7

Character.ai (2024a) Community safety updates. https://blog.character.ai/community-safety-updates/ Accessed 21 February 2025

Character.ai (2024b) The old/beta has been fully retired. Please use character.ai or the mobile app. https://support.character.ai/hc/en-us/articles/28987281244827-The-old-beta-has-been-fully-retired-Please-use-character-ai-or-the-mobile-app Accessed 21 February 2025

Coeckelbergh M (2018) How to describe and evaluate ‘deception’ phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. Ethics and information technology. https://link.springer.com/article/10.1007/s10676-017-9441-5

Dewitte P (2024) Better alone than in bad company: addressing the risks of companion chatbots through data protection by design, Computer law & security review. https://www.sciencedirect.com/science/article/abs/pii/S0267364924000852

Dquixy (2024) Is there any hope left? r/CharacterAI. https://www.reddit.com/r/CharacterAI/comments/1ga4oa7/is_there_any_hope_left/ Accessed 21 February 2025

Duffy C (2024) ‘There are no guardrails.’ This mom believes an AI chatbot is responsible for her son’s suicide. CNN, 30 October. https://edition.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit/index.html Accessed 21 February 2025

Dworkin R (2011 [1977]) Taking rights seriously. London: Bloomsbury.

European Commission (2025) Annex to the communication to the Commission: approval of the content of the draft communication from the Commission - Commission guidelines on prohibited Artificial Intelligence practices established by regulation (EU) 2024/1689 (AI Act). https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act Accessed 21 February 2025

Guo E (2025) An AI chatbot told a user how to kill himself—but the company doesn’t want to ‘censor’ it. MIT technology review. https://www.technologyreview.com/2025/02/06/1111077/nomi-ai-chatbot-told-user-to-kill-himself/? Accessed 21 February 2025

Hart HLA (1961) The concept of law. Oxford: Oxford University Press (2nd ed.).

iMDb (n.d.) Game of thrones: parents’ guide, https://www.imdb.com/title/tt0944947/parentalguide/ Accessed 21 February 2025

Han TA, Pereira LM, Lenaerts T, Santos FC (2021) Mediating artificial intelligence developments through negative and positive incentives. PLOS ONE. https://doi.org/10.1371/journal.pone.0244592

Kim Y, Sundar SS (2012) Anthropomorphism of computers: is it mindful or mindless? Computers in human behavior. https://www.sciencedirect.com/science/article/pii/S0747563211001993.

Laestadius L, Bishop A, Gonzalez M, Illenčík D, Campos-Castillo, C (2024) Too human and not human enough: a grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New media & society, 26 (10): 5923-5941. https://doi.org/10.1177/14614448221142007

Lee YK, Suh J, Zhan H, Li JJ, Ong DC (2024) Large language models produce responses perceived to be empathic. arXiv preprint arXiv:2403.18148. https://arxiv.org/abs/2403.18148

Leong B, Selinger E (2019) Robot eyes wide shut: understanding dishonest anthropomorphism. FAT* conference 2019. https://doi.org/10.1145/3287560.3287591

Livingstone S, Sefton-Green J (2025) The platformization of the family. In Sefton-Green J, Mannell K, Erstad, O (eds.). The platformization of the family: towards a research agenda. Springer Nature, pp. 7-23.

McStay A (2022) Replika in the metaverse: the moral problem with empathy in 'it from bit'. AI ethics 3: 1433–1445. https://doi.org/10.1007/s43681-022-00252-7

McStay A (2018) Emotional AI: the rise of empathic media. Sage.

McStay A, Andres F, Bland B, Laffer A, Li P, Shimo S (2024) IEEE P7014.1, Ethics and empathy-based human-AI partnering: exploring the extent to which cultural differences matter when developing an ethical technical standard. IEEE SA White Paper. https://ieeexplore.ieee.org/document/10648944

McStay A, Rosner G (2021) Emotional artificial intelligence in children's toys and devices: ethics, governance and practical remedies. Big data & society https://doi.org/10.1177/2053951721994877

Megan Garcia, individually and as the personal representative of the estate of S.R.S III, v. Character Technologies, Inc.; Noam Shazeer; Daniel De Frietas Adiwarsana; Google Llc; Alphabet Inc.; and Does 1-50. (2024) Complaint for wrongful death and survivorship, negligence, filial loss of consortium, violations of Florida's deceptive and unfair trade practices act, Fla. Stat. Ann. § 501.204, et seq., and injunctive relief. United States district court middle district of Florida Orlando division. Filed 22 October 2024. https://drive.usercontent.google.com/u/0/uc?id=1vHHNfHjexXDjQFPbGmxV5o1y2zPOW-sj&export=download Accessed 21 February 2025

Megan Garcia, individually and as the personal representative of the estate of S.R.S III, v. Character Technologies, Inc.; Noam Shazeer; Daniel De Frietas Adiwarsana; Google Llc; Alphabet Inc. (2025) Character Technologies, Inc.'s motion to dismiss plaintiff's first amended complaint (doc. 11). United States district court middle district of Florida Orlando division. Filed 24 January 2025. https://storage.courtlistener.com/recap/gov.uscourts.flmd.433581/gov.uscourts.flmd.433581.59.0.pdf Accessed 21 February 2025

Mlonyeni, PMT (2024) Personal AI, deception, and the problem of emotional bubbles. AI & society. Available: https://doi:10.1007/s00146-024-01958-4

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

Rogge A (2023) Defining, designing and distinguishing artificial companions: a systematic literature review. International journal of social robotics, 15, 9–10: 1557–1579). https://doi.org/10.1007/s12369-023-01031-ySchäfer R, Sahabi S, Brocker A, Borchers J (2024) Growing up with dark patterns: how children perceive malicious user interface designs. In Nordic conference on human-computer interaction (NordiCHI 2024), October 13–16, 2024, Uppsala, Sweden. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3679318.3685358

Sharkey A, Sharkey N (2011) Children, the elderly, and interactive robots. IEEE robotics and automation magazine, 18, 1: 32–8. https://ieeexplore.ieee.org/document/5751987 Accessed 21 February 2025

Shum H, He X, Li D (2018) From Eliza to XiaoIce: challenges and opportunities with social chatbots. Frontiers Inf Technol Electronic Eng. https://link.springer.com/article/10.1631/FITEE.1700826

Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB (2021) My chatbot companion - a study of human-chatbot relationships. Int. J. Hum Comput Stud. https://doi.org/10.1016/j.ijhcs.2021.102601.

Smahel, D, Šaradín Lebedíková M, Lacko D, Kvardová N, Mýlek V, Tkaczyk M, Svestkova A, Gulec H, Hrdina M, Machackova H, Dedkova L (2025) Tech & teens: insights from 15 studies on the impact of digital technology on well-being. EU kids online. https://doi.org/10.21953/lse.g4asyqkcrum7

Smahel D, Machackova H, Mascheroni G, Dedkova L, Staksrud E, Ólafsson K, Livingstone S, Hasebrink U (2020) EU kids online 2020: survey results from 19 countries. EU Kids Online. https://eprints.lse.ac.uk/103294/ Accessed 21 February 2025

Sunstein CR (1996) On the expressive function of law. University of Pennsylvania law review 144: 2021-2053. https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=3526&context=penn_law_review

Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng, H-T, Jin A, Bos T, Baker L, Du Y, et al. (2022) LaMDA: language models for dialog applications. arXiv preprint arXiv:2201.08239. https://arxiv.org/abs/2201.08239.

Tiku N (2022) ‘Chat’ with Musk, Trump, or Xi: ex-Googlers want to give the public AI. The Washington post, 7 October https://www.washingtonpost.com/technology/2022/10/07/characterai-google-lamda/ Accessed 21 February 2025

Turkle S (2011) Alone together: why we expect more from technology and less from each other. Basic Books, New York

Turkle S (2010) In good company? On the threshold of robotic companions. In Wilks Y (ed.) Close engagements with artificial companions: key social, psychological, ethical and design issues. John Benjamins, Amsterdam: 3–10

Turkle S (1995) Life on the screen: identity in the age of the internet. Simon & Schuster, New York.

van Es K, Nguyen D. (2024). “Your friendly AI assistant”: the anthropomorphic self-representations of ChatGPT and its implications for imagining AI. AI & Society. https://doi.org/10.1007/s00146-024-02108-6

Weber-Guskar E. (2021) How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners. Ethics and information technology. https://doi.org/10.1007/s10676-021-09598-8

Zhou L, Gao J, Li D, Shum HY (2018) The design and implementation of Xiaoice, an empathetic social chatbot. arXiv. https:// arxiv.org/abs/1812.08989