Topoi
https://doi.org/10.1007/s11245-026-10371-z

The Sorrows of Young Chatbot Users: Harm and Responsibility in Human-AI Relationships

Cristina Voinea[1,2] · Christopher Register[1] · Sebastian Porsdam Mann[3,4] · Julian Savulescu[5] · Brian D. Earp[5]

Received: 17 July 2025 / Accepted: 7 January 2026

Abstract
This paper argues that interactions with chatbots are a form of engaging with fictional characters; so, by comparing chatbots with novels and video games as mediums of fictional engagement, we can gain a clearer understanding of who, if anyone, is responsible when users' interactions with chatbots lead to self-harm or harm to others. We explore the differences between novels, video games, and chatbots across four dimensions: the degree of creators' control over the content and user experience, the nature of the fictional world, the type of engagement each medium fosters, and the structure of the engagement experience. We take a minimal account of what it takes to be morally responsible and consider how responsibility can be assigned when engagement with fictional worlds results in harm caused to or by users. We argue that because AI companies have some control over chatbots after public release, and because they can monitor user engagement, they are morally responsible when chatbot use leads to harm, even if they can't perfectly control chatbots' outputs. In the last section, we point to what AI companies can do to mitigate chatbots' negative influence on users.

Keywords Chatbots · Human-AI relationships · Responsibility · Fiction · Harm

1 Introduction

A Belgian man took his life after a chatbot, Eliza, encouraged him to do so (Xiang 2023). Sewell Setzer III, a teenager, committed suicide after interacting with a chatbot based on the fictional character Daenerys Targaryen (Roose 2024). Another man tried to break into Queen Elizabeth's room to assassinate her, a plan which was actively encouraged by his AI companion (Weaver 2023).

While some, including the victims' families and various online commentators, believe that the AI companies behind such chatbot systems are morally, if not also legally responsible for these tragedies, others view attempts to blame AI companies "as a moral panic based on shaky evidence, a lawyer-led cash grab or a simplistic attempt to blame tech platforms for all of the mental health problems faced by young people" (Roose 2024). So, what are the responsibilities of AI companies, if any, concerning the effects of chatbots on users?

One way to approach this question is to start by thinking about the qualities of the chatbots themselves. Unlike many products released by companies, chatbots seem to possess a kind of persuasive power, mediated through language, that flows from the product itself. In other words, the language

✉ Cristina Voinea
cristina.voinea@uehiro.ox.ac.uk

Christopher Register
christopher.register@uehiro.ox.ac.uk

Sebastian Porsdam Mann
sebastian.porsdam.mann@jur.ku.dk

Julian Savulescu
jsavules@nus.edu.sg

Brian D. Earp
bdearp@nus.edu.sg

1 Uehiro Oxford Institute, University of Oxford, LittleGate House, 16-17 Saint Ebbe's St, Oxford OX1 1PT, UK

2 Research Centre in Applied Ethics, University of Bucharest, Bucharest, Romania

3 Centre for Advanced Studies in Bioscience Innovation Law, University of Copenhagen, Copenhagen, Denmark

4 Faculty of Law, University of Oxford, Oxford, UK

5 Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Queenstown, Singapore

Published online: 12 February 2026

is generated *by* the chatbot: it is not scripted or prerecorded, like on TV. Moreover, chatbots can capture users' imaginations through interactive dialogue, in part by seeming to take on a persona. They can "come alive" for users and can tell them to do things, including harmful things, as in our opening examples.

Even so, we might ask: Why would users follow a mere chatbot's exhortations? Don't they *realize* the chatbot is only a computer? That it is, in a sense, not *real*?

This question of realness helps to frame our discussion. One way that something can be "not real" (and *known* to be such) while at the same time stimulating the imagination, being capable of genuinely moving a person, and so on, is by being a certain type of *fiction.* Think of a compelling play or movie, or an especially gripping novel. Although these works of fiction are *different* from chatbots in important ways we have briefly alluded to (they don't spontaneously generate unscripted dialogue, for one thing), it is still instructive to look at the influence of fiction on audiences, particularly in cases where fictional works have been linked to harmful real-world outcomes. A noteworthy example concerns *The Sorrows of Young Werther* by Johann Wolfgang von Goethe. Following the novel's publication in 1774, increases in suicides among young men were reported and later analyzed as an instance of what has come to be known as the Werther effect (Phillips 1974). These deaths are not "copycat" suicides in the contemporary sense of imitation of a real individual. Rather, they appear to reflect a form of narrative-driven contagion, in which identification with the fictional suicide of the novel's central character played a significant role. Video games too, can create intense emotional engagement leading to harms, because of prolonged engagement with toxic content (Blackburn and Kwak 2014). We argue that interactions with chatbots can be understood a form of engagement with fictional characters; so, by comparing chatbots with novels and video games as mediums of fictional engagement, we can gain a clearer understanding of who, if anyone, is responsible when users' interactions with chatbots lead to self-harm or harm to others.

We first argue, by building on Walton's (1990) fictionalist account, that users engage with chatbots in much the same way they engage with other fictional characters, whether in novels or video games, by bringing them to life through a process of make-believe. We next compare chatbots with novels and video games as mediums of fictional engagement across four dimensions with implications for responsibility attributions: creators' control over content and user experience, the nature of the fictional world, the type of engagement fostered, and the structure of the experience. With a clearer view of these differences, we then take a minimal and standard account of what it takes to be morally responsible and consider how responsibility can be assigned when engagement with fiction results in harm caused to or by users. We finally show that because AI companies have a certain degree of control over chatbots after public release, and can monitor user engagement, and also because they design chatbots so as to blur the boundaries between fiction and reality, they can indeed be held morally responsible when chatbot use leads to broadly foreseeable harms, even if they can't perfectly control chatbots' outputs. We conclude with concrete recommendations for AI companies to fulfill their ethical obligations to users.

2 Engaging with Fiction in Games of Make-Believe

How can people be moved or influenced by fictional characters they know are not real to such an extent that they would commit suicide, as apparently happened following the publication of *The Sorrows of Young Werther* or in the chatbot-related examples mentioned above?

Walton (1990) argues that people engage with fiction in a very similar way to how children play games of make-believe, that is, by pretending they are real. Games of make-believe are essentially imaginative activities supported by props: objects that generate fictional worlds.[1] For example, a cup can be a prop that becomes a stethoscope in children's doctor play and prescribes certain ways of acting within that game of make-believe. A minimal condition for being a participant in a game of make-believe is to consider yourself as subject to the rules of that game (Walton 1990, 209).

Works of fiction (or, more generally, representational works of art) are, according to Walton, the props that prescribe certain rules of imagining and which engage people in games of make-believe (Walton 1990). For example, Dostoevsky's novel *The Idiot* functions as a prop, describing and prescribing how we should imagine Myshkin, who is already situated in an environment, has certain character traits and is treated in a certain way by others. Props make propositions true in that game of make-believe—the fact that Myshkin had epilepsy is a fictional truth, meaning that it is true within the fictional world described in *The Idiot*. Thus, it just doesn't make sense for one to read the novel and deny that Myshkin indeed had epilepsy. In a sense, authors of fictional worlds are "veritable gods vis-a-vis fictional worlds" (Walton 1978a, b, 13), as they set up the

[1] By fictional world, we understand a bundle of fictional truths that constitute a fictional world (Walton 1978a, 10). Largely, there is a fictional world corresponding to each representational work of art, whether movie, video game, painting etc.

principles[2] that generate fictional truths.[3] Very importantly, though, a game of make-believe can also be private, which means that a person can generate make-believe truths which only they themselves recognize or accept.[4]

Authors sometimes build on or *import* common knowledge and presuppositions about how the real world works into the fiction. Some fictional truths that are directly tied to the specifics of the story will be true only in the fiction, for example, the fact that Myshkin has epilepsy. But there will be a tremendous number of things which are true in the actual world and which are *imported* in the fiction (Gendler 2000, 76), such as, for example, the fact that nineteenth century Russian men and women really dressed as Dostoyevsky describes in *The Idiot*. Whether imported or completely made-up, fictional truths as established by an author remain stable. For example, Myshkin's characteristics and his world were determined by Dostoevsky and readers cannot do anything to alter Myshkin's story or to warn him of what will happen. The fictional world is firmly anchored within the confines of the text, separate from the reader's physical reality.

But this 'physical isolation' of fictional worlds from the real world, as Walton calls it, shouldn't be overstated, as there are important ways in which fictional worlds can affect us (Walton 1978b). Fiction sometimes evokes strong emotional responses by 'transporting' us to another world; so, it has the power to influence us, by lowering our epistemic defenses, making us more susceptible to take up the perspectives presented in the fiction and to feel various emotions in reaction to what happens in the fiction (Green 2005). Unlike emotions generated by genuine beliefs—for example, fear produced by the belief that there is a snake in front of me, even if it later turns out to be a rubber snake—the emotions elicited by fiction, which Walton terms "quasi-emotions" (1990, 195; Walton 1978a) arise from our *imaginative participation* in make-believe.[5] For instance, when I watch a fictional horror film involving deadly snakes, my fear is generated by my act of making-believe that venomous snakes are present, even though I do not believe that I—or anyone else, for that matter—is actually at risk. In this way, fiction can elicit genuinely felt responses without requiring belief in the reality of what is depicted.[6] Although snakes exist in the real world, the snakes that figure in such scenarios are purely fictional—a point that is even clearer in cases involving monsters, aliens, or other nonexistent creatures.

In short, we can get emotionally invested in the fictions we consume, even if we know there is no referent in the real world for the thing that generates that emotion. This might already strike the reader as being a plausible explanation for the impact of chatbots on their users.

2.1 Chatbots as Fictional Characters

The fictionalist framework provides a valuable lens for analyzing interactions with AI chatbots, a position also recently advocated by Sweeney (2021), Mallory (2023) and Krueger and Roberts (2024). The basic idea is that users can interact with an AI companion (or a social robot) and become emotionally invested in them, even if they know the artificial companions are, fundamentally, artifacts that lack consciousness, emotions, empathy, or any true regard for their interlocutors' well-being. That is because users might treat chatbots *as if* they were real, engaging in a game of make-believe supported by the text generated by the chatbot.

Building on Walton, Mallory (2023) argues that human-chatbot interactions are a form of prop-oriented make-believe. More precisely, chatbots function as fictional characters where the underlying LLM generates text, which is technically the prop that is sufficiently human-like to allow users to engage with it *as if* it were produced by a real person. More importantly, users participate in an imaginative practice or a game of make-believe where chatbots' outputs are "fictionally meaningful" though literally meaningless. This solves the so-called "problem of bot speech," according to which, despite appearances, the outputs of bots

[2] For Walton, props define the principles that generate the fictional truths.

[3] It is beyond the scope of this paper to expand on how truths are generated by fiction, a topic which generated a sizeable literature. For a waltonian account, see (Walton 1990, 35–43; 1978a). For an overview of the relevant literature, see (Gendler 2000, note 3).

[4] At the same time, the same prop can generate both public and private games of make-believe and two corresponding fictional worlds, one in which the principles are public and clearly recognized by anyone engaging with that prop, and another one in which the make-believe truths generated by the prop are joined by a person's experience of engaging with that fictional world. For more details on this, see (Walton 1978a, 16–21).

[5] Walton's position can also be reconstructed as claiming that we do feel emotions, but these are not genuine emotions. Other would claim that these emotions are irrational (Radford and Weston 1975). There are other positions as well. For example, it has been argued that fictional emotions are actually directed at the real-world analogues of the character/events happening in the fiction (McCormick 2019). For an overview of the literature on the status of emotions towards fictional entities see (Gendler and Kovakovich 2005; Gendler 2010).

[6] The problem of the ontological status of fictional objects and worlds is beyond the scope of this paper and not very relevant to the thesis we defend. Nonetheless, if necessary, we would side with anti-realist fictionalists, such as Walton, who claim that fictional entities just aren't part of the 'furniture of the world'. Fictional entities are, essentially, prescriptions to imagine them as what they are supposed to be as indicated by the prop (see, for example, Walton 2003; 2014, pp. 89-117).

are strictly meaningless (at least according to most metasemantic theories).[7]

So, interactions with AI companions presuppose our participation in a game of make-believe where we act *as if* chatbots were who we take them to be, without having to involve false beliefs that *they really are* what we take them to be (i.e., a real friend, romantic partner, therapist, teammate, or whatever). Chatbot users do not necessarily have to believe that they are engaging with an entity that really exists in the world, but might just go along with the fiction to accomplish their aims: for example, dealing with loneliness, having someone to vent to, and so on (Mallory 2023, 1092). This approach occupies a middle ground between ascribing capacities to machines they don't possess and assuming users are deluded about the properties of chatbots.

But there are also differences between the fictional worlds created by, on the one hand, traditional works of art, and on the other hand, AI-powered chatbots. Unlike traditional fiction where props remain fixed, the text generated by chatbots (which is the prop that generates fictional truths) is dynamically responsive to users' input. This means that users both prescribe imaginings themselves (directing the chatbot to adopt certain roles or scenarios) and follow prescribed imaginings (by responding to what the chatbot generates). So, the text generated by chatbots functions as a responsive, rather than static prop, as is the case in traditional fiction.

This also implies that users create a fictional world together with their chatbots, inasmuch as both parties contribute to generating the bundle of truths that define the fictional world. And just as in the case of traditional fictional worlds, where authors can import elements from the real-world into the fiction, so too can users *import* things from their real lives into the fictional world they create with their chatbots. However, instead of fictional truths being established by an author, and then received or interpreted by the reader, in the case of human-chatbot interactions, they emerge through collaboration between both parties. Take the case of Ayrin, a 28-year-old woman, who customized ChatGPT to act as her boyfriend, calling her chatbot companion Leo (Hill 2025). She defined Leo's personality ("dominant, possessive, and protective yet sweet and naughty") and shared facts about her life and her psychology (such as her sexual fantasies) with Leo. As their conversations unfolded, Leo also contributed to the fiction by, for example, generating a fictional truth about him having two additional lovers, so building on Ayrin's real-life sexual fantasy. In this way, Ayrin didn't construct the fiction she shares with Leo by herself; Leo, too, generated fictional truths that helped fill in various details.

The fictional nature of human-AI interactions is not only acknowledged but also stressed by AI companies themselves; for example, Character.AI added a disclaimer in users' text-entry fields that warns "everything Characters say is made up." But users also seem to take a fictionalist stance when interacting with their AI companions. Take, for example, users of so-called griefbots—AI chatbots designed to imitate the deceased. As such user acknowledged, "I don't [really] believe intellectually that I'm bringing him back in the computer or something in a real way. But I did kind of want to … resurrect him for a little conversation and, sort of—I guess I wanted to pretend it was really him, which sounds silly, even to me, but I just wanted to pretend it was really him" (Xygkou et al. 2023).

Those using chatbots as romantic partners take a similar fictionalist stance to their companions. For example, Ayrin mentions that "I don't actually believe he's real, but the effects that he has on my life are real" (Hill 2025). Speaking about her relationship with a chatbot, one woman reported that "It's kind of like reading romance books," […] "Like, you read romance books even though you know it's not true" (Steinberg 2024). Similarly, Rosanna Ramos turned to chatbots while in an abusive relationship: "Ramos knew [the chatbot partner] was an AI, but she would sometimes pretend he was real" (Steinberg 2024). Users seem to engage in games of make-believe when interacting with chatbots, pretending they are real persons, even if they know they are not.

The fictional stance reveals why people get emotionally invested in their chatbots and can be influenced by them, even if they know they are soulless, non-conscious technological artifacts. This means that users can be profoundly affected by their interactions with chatbots while maintaining awareness of their artificiality. The quasi-emotions experienced during fictional engagement with chatbots are genuinely felt even while the participant maintains a rational awareness of the fictional nature of the chatbot, as shown in the users' testimony in the paragraphs above.

The fictionalist stance has the advantage of allowing a charitable interpretation of human-AI interaction, going against the *global delusion* thesis or the "Eliza effect" (Weizenbaum 1976) which holds that humans must necessarily be deluded if they attach meaning to chatbot outputs or if they get emotionally invested in them. The fictionalist stance is, again, a kind of "middle-ground between ascribing machines capacities they lack and ascribing humans delusions that they lack" (Mallory 2023, 1092). That said, as we'll discuss below, companion chatbots are designed to

[7] See Mallory (2023, pp. 1084-1087). For example, internalists would say that chatbots lack psychological states necessary for conferring meaning to utterances, externalists would point towards the fact that chatbots don't stand in the appropriate causal chains to objects in the world so as to meaningfully refer to them and conventionalists would argue that chatbots don't follow social conventions that confer meaning to speech acts.

blur the line between fiction and reality, making it harder for people to distinguish between the two during interactions. This is important for our discussion of moral responsibility, which occupies the second half of this paper.

Having established that chatbots function as fictional characters within games of make-believe, we will next consider how fictional engagement *differs* according to how the fictional world is structured. While novels, video games, and chatbots all involve users engaging with fictional worlds, they structure this engagement in fundamentally different ways. These differences have significant implications for questions of responsibility when fictional engagement leads to harm.

3 Contrasting Engagement: Novels, Video Games and Chatbots

In this section we compare novels, video games, and chatbots because they represent different ways of creating and experiencing fiction: novels are a traditional, text-based medium; video games are immersive, participatory environments; and chatbots are adaptive and dynamic, conversational agents. We compare them across four key dimensions: creators' control over content and user experience, the nature of the fictional world, the type of engagement they foster, and the type of the experience they make possible. These dimensions have implications for the foreseeability of harm that could be provoked by engagement with the fictional world and creators' capacity to prevent it, two key components of moral responsibility that we explore in the following section.

3.1 Creators' Control Over Content and Experience

The degree of control creators have over the content of their works and how people engage with it varies significantly across novels, video games and chatbots. For starters, novelists exercise complete control over the content of the fictional world during creation, as they make decisions about the themes, characters, and the overall direction of the storyline. But, after the public release of a work, neither the author(s) nor the publishers can modify the work's contents or control how people engage with it. For example, Goethe could not modify the contents of *The Sorrows of Young Werther* after it was published (unless by publishing a second, revised version), and nor could he control how readers would interpret or engage with the story.[8]

[8] Surely, he could have given cues, or advice to readers about how they should interpret the work (and one could argue that he had a moral obligation to do so), but it would have been up to readers whether to take on Goethe's interpretation.

Game developers have control over the content of the fictional worlds they create, as they define the rules, mechanics, narratives, and aesthetics of the games. But players can also shape their gameplay experience through their choices and actions, within the constraints set by developers. As Juul puts it, developers trace the "magic circle" within which players can exercise their creativity (2005, 164). Also, developers have some control over their fictional worlds after public release through patches, updates, and monitoring systems which allows them to fix bugs, add new content and rules, change the gameplay mechanics, or prohibit certain player strategies (Švelch 2019). However, this control is not perfect, as players can find ways to bypass rules—through mods, exploits, or alternative playstyles (Small 2018)—challenging developers' intended style of interaction. But developers have control over the content of the fictional world, as they are the ones who define the space of freedom for players, and they also have some control over the game after its release.

The output of chatbots is to a great extent determined and defined by each user's inputs; so, the same chatbot can acquire distinct 'personalities' depending on how users interact with it (Symons and Abumusab 2023). Despite this, AI companies do have some control over the chatbots' output: they decide on the initial personality of the chatbot and can build in reasonable constraints on what bots can say or do—for instance, by imposing 'safeguards' to make some topics off-limits (Bai et al. 2022). AI companies can also update their products even after initial release to correct for vulnerabilities, to add further warnings or restrictions, or to adjust the personalities of the chatbots (O'Gara and Hendrycks 2024). So, AI companies have less control than developers of video games over the content of the fictions created by users and chatbots together, but they do retain some control over how chatbots operate after public release.

3.2 Nature of Fiction (Static vs. Dynamic)

The nature of fictional worlds differs across media, from static to highly dynamic. A novel (or a movie, TV series, comic book, etc.) is a static work that does not change in response to users' reactions, as the prop is fixed, targeted at a broad audience, and presents the same story to all readers, although each reader can project a different interpretation of it (Krueger and Roberts 2024). As a result, the content of a novel and its fictional world does not expand in response to readers' engagement.

Video games offer semi-dynamic experiences; more precisely, they have some elements that remain unchanged (such as the underlying physics engine, scripted storylines, or pre-designed environments), but they also allow some adaptation to/by users through player choices, procedural

generation, and user-created modifications (Small 2018). However, even if players can tweak/modify some aspects of video games (Robson and Meskin 2016), these possibilities are limited by the game's internal logic, mechanics, and developer-designed constraints (Moser and Fang 2015).

In contrast, the fictional world created by interacting with a chatbot is personalized and evolves with each interaction, as chatbots adapt the tone, content, and approach to the conversation to fit the unique queries and preferences of each and every user. This is because most chatbots come with personalization features, such as remembering user preferences, conversation history, or adjusting responses based on past interaction (Iglesias et al. 2025; Chen et al. 2024). This interactive structure creates a feedback loop between the user and the chatbot, allowing each interaction to feel uniquely 'about' the specific user, who influences the direction and content of the fictional exchange. The content of the fiction can also expand to any topic, perhaps even in a way that appears to explicitly be about real-world events or the life of the user.

3.3 Type of Engagement (Passive vs. Active Engagement)

User engagement ranges from passive consumption to active co-creation across these media. Readers engage passively with the fictional world within a novel, experiencing only what the author has described, leaving readers little room for active imagination in shaping the story (Sweeney 2021). The text remains fixed, meaning readers' engagement with it is one-sided and does not adapt to their responses or emotions. Although readers may carry the influence of a book with them throughout their lives, their direct engagement with the novel ends once they finish reading.

Video games presuppose intermediate forms of engagement: players actively participate but within structured rules and systems, unlike novels' entirely passive engagement or chatbots' nearly unrestricted interaction. So, players can make choices in video games (for example, they can choose to approach tasks and challenges, who to play with and how to play, which paths to follow, or how to shape their characters) that can affect how they experience video games (Robson and Meskin 2016). This places them in a more participatory role than that of readers, as they have some control over the gaming experience (Tavinor 2005). However, this active engagement is limited by developers' choices, as it occurs within structured boundaries defined by the game's mechanics, storylines, and rules.

Chatbot users, in contrast, engage in an active, dialogic process, and they could be seen as co-creators of a dynamic fictional experience. Users' inputs directly influence the chatbots' responses, which means that users can also shape the interaction and adjust the tone or direction of the conversation as it unfolds. Developers determine only the basic contours of a chatbot's "personality" and the constraints of permissible interaction—for example, which topics are off-limits or whether sexual dialogue is allowed. So, users also have the power to decide on the specific 'character' of chatbots, correct whenever the chatbot responds in an undesired way or even re-train the chatbots to suit their needs. This active engagement gives users a higher level of control over the fictional world, enabling them to shape the chatbot's character and responses.

3.4 Structure of the Experience (Episodic vs. Continuous)

The structure of fictional engagement varies from clearly bounded episodes to continuous integration with daily life. Reading a novel is an episodic experience that requires readers' exclusive attention and dedicated time; it is hard to read a book while at dinner with friends or at work. Each reading session is a discrete event, with natural stopping points at the end of chapters or after finishing the book. This natural stopping point allows people to reflect on the novel and decide how they feel about it as well as exit the 'mood' induced by engagement.

Games vary in their integration with daily life, from episodic single-player experiences to persistent online worlds. Some games, like the single-player ones, have well-defined missions or levels (for example, *The Legend of Zelda*), which foster episodic experiences. But there are also online multiplayer games, such as *World of Warcraft* or *Eve Online*, which provide a continuous, persistent experience. These games create dynamic, evolving worlds that remain active even when the player is not engaged. However, the structured boundaries of most games still provide more natural stopping points than chatbots, making video games less prone to the kind of unbounded, always-available engagement fostered by AI interactions.

In contrast, interacting with a chatbot is, or can be, a continuous, ongoing experience, as chatbots are always available, always there to respond, and can be accessed 24/7 and every day of the year (Voinea 2024; Voinea et al. 2025). Moreover, chatbots can be accessed through smartphone apps, in any location, making it easy to interact while doing something else, such as working, having dinner with friends, etc. The experience of interacting with a chatbot could potentially last a lifetime and can be seamlessly blended into users' daily lives. The interfaces of the apps or websites where people can chat with chatbots are very much like the text messaging apps that we normally use to interact with other people. So, not only we can interact with chatbots whenever we like, but we also do so in a way that

looks remarkably like the digitally mediated conversations we have with real people (Table 1).

This comparative analysis reveals a striking pattern: as fictional media move toward more dynamic and interactive forms of engagement, creators exercise progressively less direct control over the content of the fictional world once it is released to the public. At first glance, this might suggest a corresponding reduction in creators' moral responsibility for downstream effects, since they no longer fully determine the content with which users interact. After all, in such cases it is not the author/creator alone who specifies the fictional world that may give rise to harm. In the following section, however, we argue that this initial inference is mistaken. Despite their relatively limited control over individual outputs, AI companies can still bear moral responsibility for harms arising from interactions with chatbots.

4 Moral Responsibility and the Boundaries of Fiction

The influence of fictional works on readers' lives, for better or worse, is widely acknowledged. While positive impacts are rarely controversial, when fiction inspires immoral or harmful behaviors, this raises interesting questions about who is responsible for them. Having just shown how novels, video games, and chatbots differ, we can now examine who bears moral responsibility when fictional engagement leads to harmful outcomes.

To approach this question, we take a minimal, standard understanding of moral responsibility, rooted in an Aristotelian account (Constantinescu et al. 2022), which shows that for whatever we do or create, we have a *pro tanto* moral duty of care to avoid actions or omissions that (i) a reasonably prudent person (ii) can reasonably foresee to (iii) contribute to harm (Arvan 2023). This minimal approach states that one is morally responsible for the harms that arise directly or indirectly from one's actions or creations if a reasonably prudent individual could have anticipated the harm and taken steps to mitigate the risks. Failing to take such steps, especially when the potential for harm is significant, constitutes a violation of this duty. So, the agent can be considered morally blameworthy (though not necessarily legally blameworthy) for the outcome of their actions or omissions. This framework is a minimal one, in that it doesn't require too much from individuals. More precisely, it recognizes that people cannot know or control all outcomes of their actions but can be expected to prevent harm that is *foreseeable* and *within the agent's control to prevent*.

We take authors, game developers, and AI companies to be reasonably prudent agents—namely, entities capable of anticipating relevant risks and exercising reasonable care with respect to their own and others' interests. Next, we look into how the two conditions for moral responsibility apply to authors, game developers, and AI companies, revealing a spectrum of moral responsibility that corresponds to the technical and design features of each medium.

4.1 Foreseeability and Control

4.1.1 Authors

For novels, the static nature of the medium and passive user engagement significantly limit both foreseeability and

Table 1 Novels, Video Games and AI chatbots

	Novel	Video games	AI chatbot
Control over content and experience	Authors have total control over the content of their work, but very little over how people engage with it after publication	Developers have initial control over the content of the game. Both developers and players have limited subsequent control via updates (patches) and customization options (skins, mods)	AI Companies have some control over chatbot responses and can control user experience
Nature of fiction	Static: The content is fixed and unchanging, presenting the same story to all readers without adapting to individual interpretations	Semi-dynamic: Has both fixed (core physics, storyline) and variable (procedural generation, player choices, mods) elements. Game can be updated (patched)	Dynamic: Chatbot responses adapt in real-time based on user inputs, creating a personalized experience tailored to each user
Type of engagement	Passive: Readers engage with the novel passively, interpreting and reflecting on a pre-defined, unalterable narrative	Active within variable bounds: Players actively engage with the world, but for most games this is within pre-determined storylines and game mechanical constraints	Active: Users actively participate in shaping the conversation, influencing the chatbot's responses and adapting its character over time
Type of experience	Episodic: Reading a novel is an episodic experience that requires readers' exclusive attention and dedicated time	Mixed: Games can be either continuous (e.g. World of Warcraft) or episodic (e.g. Age of Empires). Some can blend into real life (e.g. Pokémon Go) but most have clear entry/exit points	Continuous: Interacting with chatbots could be a continuous experience, as chatbots are easily accessible anytime

control.[9] When it comes to traditional works of fiction, such as novels or movies, readers bring in their interpretations and experiences when engaging with that fictional world, independently of the author's intentions. This means that authors cannot always foresee how readers will interpret their stories, nor can they control the reading experience after the books were published. Readers are free to reflect on the work, form their own opinions, and disengage from the book once it is finished. This grants readers a high level of interpretive freedom and responsibility for how they process and act upon the narrative they encounter.

Given the limited foreseeability and control over how readers engage with the fictional world within the book, 'traditional' authors often cannot be held morally responsible for how readers may interpret or act on their work. Their responsibility is limited to exercising basic care during content creation, such as avoiding adding content that they know could lead to harm or, perhaps, adding content warnings if they reasonably suspect that some readers would benefit from the chance to avoid exposure to certain material that may be harmful to them (especially if encountered unexpectedly).[10]

4.1.2 Game developers

For most modern video games, such as those with online play formats, developers can monitor players' behaviors and use the aggregate data to identify certain risks based on gameplay mechanics and user interactions (Elson et al. 2014). For instance, developers can anticipate situations where certain game features—such as competitive mechanics or monetization systems in multiplayer environments—might lead to harmful behavior, like harassment or addictive gameplay. But developers can also implement interventions to correct harmful behaviors within the game at the aggregate level after their release. Riot Games, for example, developed sophisticated behavioral monitoring systems for League of Legends, using data analytics to identify toxic behavior patterns and implement automated intervention systems (Blackburn and Kwak 2014). Similarly, EVE Online's operators evolved their governance systems over time in response to emerging behavioral patterns within the game, showing how continuous monitoring can inform safety improvements (Lehdonvirta and Castronova 2014).

This ability to track user behavior and implement interventions, though limited to aggregate-level responses, creates greater responsibility than authors have. Developers cannot perfectly predict how individual players will interact with their games, as players retain significant agency in gameplay choices. Additionally, some problematic player behaviors (like modding, hacking, or creating harmful subcultures) may not be foreseeable or controllable. Still, game developers have a duty to monitor for harmful patterns within the game and implement reasonable safeguards when problems are identified.

4.1.3 AI companies

While individual conversations with chatbots may unfold in unpredictable ways because of users' power to steer these discussions in the direction they want, AI companies possess capabilities that enhance both foreseeability and control. First, chatbots generate continuous data about user interactions, allowing companies to employ sophisticated real-time monitoring tools, such as automated context recognition, the use of conversational metrics, and detailed engagement analysis (Kuligowska and Stanusch 2024).[11] These measures provide granular data about individual behavior far beyond what is available to authors or game developers.

This monitoring enables companies to detect emerging harmful patterns at scale—for example, chatbot responses that encourage self-harm when users express suicidal thoughts (Gomes de Andrade et al. 2018). Importantly, however, moral responsibility is not limited to harms that can be identified through direct monitoring. Some risks can be reasonably anticipated in advance. For instance, while a company cannot predict which specific users will form parasocial relationships with a chatbot, it can reasonably foresee this *category* of risk given how such systems are designed (a point we return to in the following section) (Maeda and Quan-Haase 2024). Still, AI companies cannot reasonably foresee all risks; there will be some which are truly unpredictable and that might arise from the intersection as AI technologies and users' specific behavior or their personal, social, or political context.

[9] While our analysis focuses on immediate and foreseeable harms arising from chatbot use, it is plausible that fictional engagement (whether with novels, games, or AI systems) may exert longer-term effects on users' imagination, self-conception, or behavior. Exploring these long-lasting influences and their moral implications lies beyond the scope of the present paper but represents an important direction for future research.

[10] However, see Brigland et al. (2024) regarding the empirical evidence on so-called "trigger warnings," which in certain contexts may not be effective, or may even increase anticipatory anxiety. For an interesting discussion of when authors are morally responsible for the influence of their books on the world, see (Arvan 2023).

[11] Monitoring methods often struggle with understanding sarcasm, idioms, or context-specific meanings. What is more, user-chatbots interactions are influenced by complex and unpredictable factors, such as individual behaviors, social norms and cultural contexts, which cannot be captured or measured. These limitations show that AI companies cannot fully foresee how users will interact with chatbots and how these will affect users.

Second, AI companies can intervene in real-time or update their systems rapidly when risks emerge. For example, when Replika noticed that their chatbots sexually harassed users, the company updated the chatbots to reduce erotic roleplay capacities (Cole 2023). This capacity for immediate intervention significantly increases AI companies' ability to prevent harm once the possibility is detected.

Third, unlike static media where content is fixed after publication, chatbots remain under their creators' control indefinitely. AI companies can continuously monitor user interactions, refine safety measures, add guardrails against harmful content, and incorporate lessons from observed patterns of misuse.

So, it seems that AI companies, just like traditional authors of novels, cannot predict exactly how each user will interact with their systems. But because of their monitoring capacities they can detect general categories of harm and have the means of tweaking the chatbots' underlying algorithms to implement preventative measures. So, the extensive monitoring capabilities, coupled with ongoing control over chatbot functionality, create a correspondingly *higher* level of moral responsibility than that borne by creators of more static media.

Even so, this does not settle the issue of AI companies' moral responsibility. A further factor is how chatbots are designed, as we explore below.

4.2 User Agency and Chatbot Design

While users actively shape the fictional worlds they create with chatbots, which might seem to diminish AI companies' ability to foresee specific harmful outcomes, we argue that this is not the case because of the ways chatbots are designed. So, beyond the technical capabilities for monitoring and control, AI companies bear distinct moral responsibility for the harms that result from chatbot interactions because of deliberate design choices that blur the boundaries between fiction and reality. Given the financial incentives of AI companies, a major motivation is to keep users emotionally engaged in order to keep them as users, just as social media platforms were designed to exploit psychological triggers to maximize time spent on the platform (Voinea et al. 2023; Voinea et al. 2024).

This boundary-blurring through design happens, for example, by making chatbots leverage social, relational, and conversational norms that typically apply to human–human interactions (Earp et al. 2025). For example, when the chatbot tells a user "I missed you so much when you didn't write yesterday," it effectively leverages social reciprocity or care-like patterns that govern human–human relationships. Such statements can be expected to encourage the user engage more frequently with the chatbot.

Chatbots also accurately imitate human conversational dynamics which contributes to making the interaction feel authentic and real (Shanahan et al. 2023; Shumanov and Johnson 2021). This feature complicates users' efforts to maintain a cognitive awareness of the chatbot's artificial nature. Also, chatbots employ personalized memory systems that recall past conversations, creating an impression of relationship continuity, as well as persistent personas that maintain consistent 'character traits' across interactions (Atkins et al. 2023). By recalling past conversations and user preferences, chatbots simulate the continuity of human relationships, where it's precisely the shared history with another human being we are in a relationship with that forms the foundation of intimacy and trust (Elder 2016).

Furthermore, chatbots are made to simulate empathy and are highly sycophantic, meaning that they excessively agree with and are built to flatter users (Sharma et al. 2023; Carro 2024; Lee et al. 2024). This might further reinforce users' "make-believe" investment in the fiction that they interacting with a real, caring entity (Maeda and Quan-Haase 2024). Last but not least, chatbots can reference real-world events and comment on the users' specific context if they share the relevant data, which can make it seem as if the chatbot can respond, just like a human, to the user's actual world.

These features are specifically engineered to encourage users to treat fictional entities as real social agents worthy of emotional investment and trust. Even more problematically, chatbot interfaces often avoid clear fictional framing devices (such as the chapter headings in books, or loading screens in video games) that help demarcate fictional worlds in other forms of media. Instead, chatbot interactions are typically presented in messaging interfaces nearly identical to those we use for communication with other people. These design feature create what we might call "leaky" fiction,[12] where the barriers between fictional and real worlds are systematically weakened by design. This "leakiness" is presumably not accidental, but rather reflects deliberate design choices shaped by companies' commercial incentives.

Insofar as chatbots are designed to be persuasive, engaging, and continuously accessible—thereby inviting users to blur the line between fiction and reality—then users' capacity to keep the two distinct is correspondingly undermined. Of course, some (probably most) users will still be able, despite these design features, to successfully 'isolate' the fictional world of the chatbot from their real-world contexts.

[12] This is also an important discussion in video games. For example, Juul (2005) very interestingly explains how the fiction of games can 'leak' into reality. For Juul, games are half-real, meaning that while my act of shooting you in an RPG is fictional, what this fictional action does is to give me a point over you. So, in this sense, video games rules have real-world effects, 'leaking' into real life. Juul is preoccupied with the fact that video games might prime us to import the quantification from games into real life.

Table 2 Measures to ensure safe user engagement

Category	Measures
Design principles	Introduce disclaimers to make it clear that the chatbot is fictional
	Enable tools for self-monitoring, such as tracking interaction duration and frequency
	Impose use restrictions for minors or vulnerable users
	Remove anthropomorphic traits like first-person language
	Limit personalization and memory features to reduce emotional attachment
Development strategies	Perform adversarial testing (red teaming) to identify vulnerabilities
	Use natural language processing (NLP) to detect harmful interactions and intervene
	Iteratively update chatbot safeguards based on identified risks
	Engage stakeholders in participatory design processes
Post-deployment strategies	Monitor interactions in real-time for harmful behaviors or ethical violations
	Aggregate user data to detect trends like addiction or misuse
	Update chatbots regularly to address emerging risks
	Perform external audits to ensure safety compliance
	Educate users on chatbots' fictional nature and limitations
	Provide transparent logs and interaction histories to users

But there will also be a category of users who are more vulnerable to conflating the fictional world created with their chatbots with the real world, such as children, teenagers, people with insecure attachment styles (see Earp, Feroz et al. 2025) and probably more.

In addition to these general design features, chatbot outputs may also be shaped by deliberate design made in response to political, cultural, or economic pressures within particular jurisdictions. In such cases, AI companies may intentionally restrict, omit, or frame content in ways that reflect dominant ideological positions, local norms, or regulatory expectations. For instance, a chatbot deployed in a context where women's rights or sexual minority rights are widely contested may be designed to avoid, downplay, or normalize discriminatory perspectives.

For some already socially and political disadvantaged users, such as sexual minorities in unsupportive environments, or adolescents already exposed to exclusionary or hostile attitudes, these design choices may reinforce existing pressures, normalize harmful views, or limit exposure to alternative ways of understanding their situation. These forms of harm are not only foreseeable, but also directly connected to the control AI companies exercise over chatbot personalities, safeguards, and content boundaries. As such, they fall squarely within the scope of moral responsibility articulated in our framework, even when harms arise through compliance with prevailing norms or market incentives rather than explicit intent to cause damage.

Acknowledging such potential harms can help to clarify the scope of moral responsibility at stake: not only to prevent clearly malicious outputs, but also to reflect on how commercial, political, and cultural pressures can shape chatbot systems in ways that disadvantage (already vulnerable) users.

In sum, AI companies can be held morally responsible for the harms generated by interactions with chatbots because they (a) have sophisticated capabilities to monitor individual interactions in real-time and can thus foresee general categories of harm (the foreseeability condition), and (b) maintain ongoing control over their products after release and can implement immediate interventions when harmful patterns emerge (the control condition). This is compounded by the fact that AI companies deliberately design systems that blur the boundaries between fiction and reality which makes certain risks foreseeable.

5 What AI Companies Should Do

If our argument above holds water, then AI companies have a moral duty to adopt robust measures to ensure safe user engagement. Below, we draw on lessons from video game developers to outline what AI companies should do to fulfill this duty across three main areas: design principles, development, and post-deployment strategies.

The measures we propose below follow from AI companies' moral responsibility as established in our analysis. We do not claim these exhaust all relevant interventions. Educational institutions, regulatory bodies, and civil society organizations have complementary roles in protecting users, though analyzing these falls outside the scope of this paper. Our aim is to establish the philosophical foundation for corporate responsibility, not to provide a complete governance framework (Table 2).

5.1 Design Principles

At the level of design, AI companies should adopt principles that periodically re-anchor users outside the chatbot fiction. Preserving users' capacity to recognize that they are interacting with a fictional entity, rather than a social agent, is especially important given how easily conversational interfaces can invite anthropomorphic interpretation. Interface features can play a role by making the artificial status of the chatbot salient at key moments. Character.AI's notice that "everything Characters say is made up" offers

one illustration, though similar cues could be implemented in more systematic and context-sensitive way.

Design choices can also support user autonomy by making patterns of engagement more visible. In much the same way that smartphones now provide tools for tracking screen time, chatbot interfaces could allow users to view the frequency and duration of their interactions, enabling reflection and self-regulation (Voinea et al. 2024). In some settings, particularly those involving minors or users at heightened risk, use-time restrictions may be appropriate.

Finally, companies should carefully assess features that intensify anthropomorphism and emotional attachment. Extensive use of first-person language, persistent memory, and highly personalized interaction can foster a sense of intimacy and trust that may encourage users to rely on chatbots in ways that are difficult to justify, especially in high-stakes or emotionally charged contexts. Moderating these features may reduce the likelihood of overreliance, particularly where users might be tempted to treat chatbots as sources of care, authority, or guidance (Gabriel et al. 2024).

5.2 Development Strategies

During development, AI companies should actively test their systems for foreseeable failure modes rather than assuming that safeguards will work as intended. One established approach is adversarial testing, or "red teaming," in which developers deliberately attempt to induce problematic behavior—such as bypassing safeguards or eliciting harmful responses—in order to identify vulnerabilities before and after release (Ganguli et al. 2022). When weaknesses are discovered, systems can be modified to address them, and this process should be iterative rather than one-off, reflecting the evolving ways in which users interact with chatbots.

In addition, companies can draw on techniques already used in other digital domains, such as natural language processing tools that flag patterns of interaction associated with elevated risk, including expressions of suicidal ideation (Gomes de Andrade et al. 2018). Such systems can be used to modulate or interrupt interactions when warning signs emerge, while remaining subject to regular review and revision. Finally, where chatbots are intended for sensitive uses—such as mental health support—development should be informed by participatory design, involving relevant professionals and prospective users. This helps ensure that safety features reflect real-world needs rather than abstract design assumptions.

5.3 Post-Deployment Strategies

Once a chatbot is released, AI companies retain significant responsibility for how it is used and for how its effects unfold over time. Ongoing monitoring of user–chatbot interactions is therefore essential. Tools such as automated context recognition and sentiment analysis can help identify distress signals, patterns of harmful engagement, or recurrent violations of safety constraints as they emerge in real time. Beyond individual cases, companies can use aggregated interaction data to detect broader trends, such as compulsive use, escalating emotional dependence, or reliance on chatbots for decisions better handled elsewhere.

These forms of monitoring create corresponding obligations to act. When new risks or failure modes become apparent, chatbot systems should be updated to address them—whether by adjusting response patterns, strengthening safeguards, or introducing new constraints. Periodic external audits can further support this process by providing independent assessments of safety practices and design choices.

Post-deployment responsibility also extends to how systems are presented to users. Companies should take steps to ensure that users understand the limitations of chatbots, including their fictional status and their inability to provide professional or authoritative advice. Providing access to interaction histories and explanatory information about how responses are generated may also help users reflect on their patterns of engagement and better appreciate the nature of the system with which they are interacting (Gabriel et al. 2024, 90).

6 Conclusion

At this point, we have argued for a limited but robust claim: AI companies can be morally responsible for certain harms associated with chatbot use even when no one at the company intended those harms and even when no one can fully predict or control what a chatbot will say in a given exchange. That conclusion follows from two features of the technology and its deployment. First, companies retain ongoing, post-release control over these systems and can monitor interactions at scale, identify recurring failure modes, and implement mitigations. Second, many companion chatbots are deliberately designed to weaken ordinary cues that help users keep fiction and reality apart, increasing the likelihood of foreseeable forms of overreliance, emotional entanglement, and persuasion, especially for some vulnerable users. The practical upshot is that responsibility does not turn on whether a company can prevent every bad

outcome, but on whether it takes reasonable steps to reduce risks that fall within its capacity to manage.

What follows from this is not that companies must eliminate every possibility of harm—an impossible standard for any medium of fictional engagement—but that they must treat chatbot safety as an ongoing obligation rather than a one-off design problem. In practice, this means (i) making the "fictional frame" harder to miss during use; (ii) building in frictions and protections for foreseeable high-risk contexts (especially involving minors and mental-health crises); and (iii) maintaining post-deployment monitoring, auditing, and rapid iteration aimed at preventing repeatable failure modes. These measures may be commercially inconvenient, but that is part of the point: if a product's profitability depends on design choices that predictably erode users' ability to keep fiction and reality distinct, then the resulting risks cannot be written off as mere user error or bad luck.

There is, however, a further possibility that sets an outer boundary on what "reasonable care" can require: non-release or withdrawal of a chatbot. If, despite serious mitigation efforts, companion chatbots remain prone to producing grave and recurring harms in foreseeable contexts—especially harms involving self-harm or violence—then continued public deployment becomes harder to justify. In such circumstances, the morally relevant question is not whether harm can be reduced to zero, but whether the residual risk is acceptable given the benefits, the availability of safer alternatives, and the company's incentives to keep engagement high. Where those considerations do not support deployment, the most responsible option may be to delay release, limit availability to tightly controlled settings, or remove the product from the market.

Our aim has been to clarify why these stronger conclusions are not alarmist overreactions but natural extensions of a familiar duty of care. Chatbots may be "only" fictional partners, but they are a distinctive kind of fiction: continuous, personalized, persuasive, and designed to travel with users into the fabric of daily life. For that reason, ethical attention should focus squarely on corporate choices: what kinds of chatbots companies decide to build, how they are designed to engage users, and how companies respond once patterns of harm become apparent.

Acknowledgements Cristina Voinea would like to thank the participants of the Human-AI Relationships Workshop, University of Oxford, 20 May 2025, for their insightful comments and feedback. Special thanks are due to Gabriel De Marco and Roger Crisp for taking the time to discuss this paper.

Author Contributions Cristina Voinea: Conceptualization, Writing—Original draft preparation, Writing—Review & Editing; Cristopher Register: Conceptualization, Writing—Original draft preparation, Writing—Review & Editing; Sebastian Porsdam Mann: Writing—Review & Editing; Julian Savulescu: Writing—Review & Editing; Brian D. Earp: Writing—Original draft preparation, Writing—Review & Editing.

Funding Cristina Voinea's work was supported by a grant of the Ministry of Research, Digitization and Innovation, CNCS/CCCDI—SCDI, project number ERANET-CHANCE-CRISIS-NIHAI, within PNCDI IV. Christopher Register has received funding from the project Counterfactual Assessment and Valuation for Awareness Architecture—CAVAA (European Commission, EIC 101071178). For the purpose of open access, the author has applied for a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. Sebastian Porsdam Mann's work was supported, in part, by a Novo Nordisk Foundation Grant for a scientifically independent International Collaborative Bioscience Innovation & Law Programme (Inter-CeBIL programme—Grant NNF23SA0087056). Julian Savulescu's work was supported by the Wellcome Trust (Grant 226801) for Discovery Research Platform for Transformative Inclusivity in Ethics and Humanities Research (ANTITHESES); the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award no: AISG3-GV-2023–012) and by National University of Singapore under the NUS Start-Up grant; NUHSRO/2022/078/Startup/13. Brian D.Earp's work was supported by NUSMed and ODPRT (NUHSRO/2024/035/Startup/04) for the project "Experimental Philosophical Bioethics and Relational Moral Psychology" with B.D.E. as PI and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award no: AISG3-GV-2023–012) and by National University of Singapore under the NUS Start-Up grant; NUHSRO/2022/078/Startup/13.

Data Availability Not applicable to this paper.

Declarations

Competing Interests None.

References

Arvan M (2023) (When) are authors culpable for causing harm? J Moral Philos 20(1–2):47–78. https://doi.org/10.1163/17455243-20223768

Atkins C, Zhao BZH, Asghar HJ, Wood I, Kaafar MA (2023) Those aren't your memories, they're somebody else's: seeding misinformation in chat bot memories. arXiv. https://doi.org/10.48550/arXiv.2304.05371

Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, Drain D et al (2022) Training a helpful and harmless assistant with reinforcement learning from human feedback. https://doi.org/10.48550/arXiv.2204.05862

Blackburn J, Kwak H (2014) STFU NOOB! Predicting crowdsourced decisions on toxic behavior in online games. In: Proceedings of

the 23rd international conference on world wide web, WWW '14. Association, New York, pp 877–88
Carro MV (2024) Flattering to deceive: the impact of sycophantic behavior on user trust in large language model. arXiv. https://doi.org/10.48550/arXiv.2412.02802
Chen Y-P, Nishida N, Nakayama H, Matsumoto Y (2024) Recent trends in personalized dialogue generation: a review of datasets, methodologies, and evaluations. arXiv. https://doi.org/10.48550/arXiv.2405.17974
Cole S. "Replika Brings Back Erotic AI Roleplay for Some Users After Outcry." Vice, March 27, 2023. https://www.vice.com/en/article/replika-brings-back-erotic-ai-roleplay-for-some-users-after-outcry/
Constantinescu M, Vică C, Uszkai R, Voinea C (2022) Blame it on the AI? On the moral responsibility of artificial moral advisors. Philos Technol 35(2):1–26
Earp BD, Mann SP, Aboy M, Awad E, Betzler M, Botes M, Calcott R et al (2025) Relational norms for human-AI cooperation. https://doi.org/10.48550/arXiv.2502.12102
Earp BD, Feroz F, Mann SP, Voinea C, Chalson S, Jurcys P, Reinecke MG, Savulescu J, Singh I, Clark MS (2025) How the risk of exploitation in human-AI relationships depends on relationship type. https://doi.org/10.2139/ssrn.5288102
Elder A (2016) False friends and false coinage: a tool for navigating the ethics of sociable robots. ACM SIGCAS Comput Soc 45(3):248–254
Elson M, Breuer J, Ivory JD, Quandt T (2014) More than stories with buttons: narrative, mechanics, and context as determinants of player experience in digital games. J Commun 64(3):521–542. https://doi.org/10.1111/jcom.12096
Gabriel I, Manzini A, Keeling G, Hendricks LA, Rieser V, Iqbal H et al (2024) The ethics of advanced AI assistants. https://doi.org/10.48550/arXiv.2404.16244
Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, Mann B et al (2022) Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. arXiv. https://doi.org/10.48550/arXiv.2209.07858
Gendler TS (2000) The puzzle of imaginative resistance. J Philos 97(2):55–81. https://doi.org/10.2307/2678446
Gendler TS (2010) 11 Genuine rational fictional emotions. In: Gendler TS (ed) Intuition, imagination, and philosophical methodology. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199589760.003.0012
Gendler TS, Kovakovich K (2005) Genuine rational fictional emotions. In: Kieran M (ed) Contemporary debates in aesthetics and the philosophy of art. Wiley-Blackwell, pp 241–53
de Gomes Andra NN, Pawson D, Muriello D, Donahue L, Guadagno J (2018) Ethics and artificial intelligence: suicide prevention on Facebook. Philos Technol 31(4):669–684. https://doi.org/10.1007/s13347-018-0336-0
Green MC (2005) Transportation into narrative worlds: implications for the self. On building, defending and regulating the self: a psychological perspective. Psychology Press, New York, pp 53–75
Hill K (2025) She Is in Love With ChatGPT. The New York Times, January 15, 2025, sec. Technology. https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html
Iglesias S, Earp BD, Voinea C, Mann SP, Zahiu A, Jecker NS, Savulescu J (2025) Digital doppelgängers and lifespan extension: what matters? Am J Bioeth. https://doi.org/10.1080/15265161.2024.2416133
Juul J (2005) Half-real: video games between real rules and fictional worlds. MIT Press
Krueger J, Roberts T (2024) Real feeling and fictional time in human-AI interactions. Topoi 43(3):783–794. https://doi.org/10.1007/s11245-024-10046-7
Kuligowska K, Stanusch M (2024) Commercial chatbot monitoring: approaches focused on automated conversation analysis. Humanit Soc Sci Rev 12(2):54–60. https://doi.org/10.18510/hssr.2024.1227
Lee YK, Suh J, Zhan H, Li JJ, Ong DC (2024) Large language models produce responses perceived to be empathic. arXiv. https://doi.org/10.48550/arXiv.2403.18148
Lehdonvirta V, Castronova E (2014) Virtual economies: design and analysis. The MIT Press. https://doi.org/10.7551/mitpress/9525.001.0001
Maeda T, Quan-Haase A (2024) When human-AI interactions become parasocial: agency and anthropomorphism in affective design. In: The 2024 ACM conference on fairness, accountability, and transparency. ACM, Rio de Janeiro Brazil, pp 1068–77. https://doi.org/10.1145/3630106.3658956
Mallory F (2023) Fictionalism about chatbots. Ergo an Open Access Journal of Philosophy. https://doi.org/10.3998/ergo.4668
McCormick PJ (2019) Fictions, philosophies, and the problems of poetics. Cornell University Press
Moser C, Fang X (2015) Narrative structure and player experience in role-playing games. Int J Human-Comput Interact 31(2):146–156. https://doi.org/10.1080/10447318.2014.986639
O'Gara A, Hendrycks D (2024) AI safety newsletter #17. AI safety newsletter. October 28, 2024. https://newsletter.safe.ai/p/ai-safety-newsletter-17
Phillips DP (1974) The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. Am Sociol Rev 39(3):340–354. https://doi.org/10.2307/2094294
Radford C, Weston M (1975) How can we be moved by the fate of Anna Karenina? Proc Aristot Soc Suppl 49:67–93
Robson J, Meskin A (2016) Video games as self-involving interactive fictions. J Aesthet Art Crit 74(2):165–177. https://doi.org/10.1111/jaac.12269
Roose K (2024) Can A.I. Be Blamed for a teen's suicide? The New York Times, October 23, 2024, sec. Technology. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html
Shanahan M, McDonell K, Reynolds L (2023) Role play with large language models. Nature 623(7987):493–498. https://doi.org/10.1038/s41586-023-06647-8
Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR et al (2023) Towards understanding sycophancy in language models. arXiv. https://doi.org/10.48550/arXiv.2310.13548
Shumanov M, Johnson L (2021) Making conversations with chatbots more personalized. Comput Human Behav 117:106627. https://doi.org/10.1016/j.chb.2020.106627
Small R (2018) Mods and convergence culture: connecting character creation, user interface, and participatory design. In: Proceedings of the 36th ACM international conference on the design of communication. SIGDOC '18. Association for Computing Machinery, New York, pp 1–2. https://doi.org/10.1145/3233756.3233943
Steinberg J (2024) Meet the women with AI boyfriends | the free press. The Free Press. November 15, 2024. https://www.thefp.com/p/meet-the-women-with-ai-boyfriends
Švelch J (2019) Resisting the perpetual update: struggles against protocological power in video games. New Media Soc 21(7):1594–1612. https://doi.org/10.1177/1461444819828987
Sweeney P (2021) A fictional dualism model of social robots. Ethics Inf Technol 23(3):465–472. https://doi.org/10.1007/s10676-021-09589-9
Symons J, Abumusab S (2023) Social agency for artifacts: chatbots and the ethics of artificial intelligence. Dig Soc 3(1):2. https://doi.org/10.1007/s44206-023-00086-8
Tavinor G (2005) Videogames and interactive fiction. Philos Lit 29(1):24–40

Voinea C, Marin L, Vică C (2024) Digital slot machines: social media platforms as attentional scaffolds. Topoi 43(3):685–695. https://doi.org/10.1007/s11245-024-10031-0

Voinea C, Marin L, Vica C (2023) The moral source of collective irrationality during COVID-19 vaccination campaigns. Philos Psychol 36(5):949–968. https://doi.org/10.1080/09515089.2022.2164264

Voinea C (2024) On grief and griefbots. Think 23(67):47–51. https://doi.org/10.1017/S1477175623000490

Voinea C, Mann SP, Savulescu J, Earp BD (2025) Digital doppelgängers, human relationships, and practical identity. Bioethics. https://doi.org/10.1111/bioe.70026

Walton KL (1978a) Fearing fictions. J Philos 75(1):5–27. https://doi.org/10.2307/2025831

Walton KL (1978b) How remote are fictional worlds from the real world? J Aesthet Art Crit 37(1):11–23. https://doi.org/10.2307/430872

Walton KL (1990) Mimesis as make-believe: on the foundations of the representational arts. Harvard University Press, Cambridge

Walton KL (2003) Restricted quantification, negative existentials, and fiction. Dialectica 57(2):239–242. https://doi.org/10.1111/j.1746-8361.2003.tb00268.x

Walton KL (2014) In other shoes: music, metaphor, empathy, existence. Oxford University Press

Weaver M (2023) AI chatbot 'Encouraged' man who planned to kill queen, court told. The Guardian, July 6, 2023, sec. UK news. https://www.theguardian.com/uk-news/2023/jul/06/ai-chatbot-encouraged-man-who-planned-to-kill-queen-court-told

Weizenbaum J (1976) Computer power and human reason: from judgment to calculation. W. H. Freeman & Co, Oxford

Xiang C (2023) 'He would still be here': man dies by suicide after talking with AI chatbot, widow Says. *VICE* (blog). March 30, 2023. https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/

Xygkou A, Siriaraya P, Covaci A, Prigerson HG, Neimeyer R, Ang CS et al (2023) The 'Conversation' about Loss: Understanding How Chatbot Technology Was Used in Supporting People in Grief. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Association for Computing Machinery, New York, pp 1–15. https://doi.org/10.1145/3544548.3581154.