# Anthropomorphism as Social Affordance: Charting the Co-Animation of Chatbots into Social "Agents"

**Takuya Maeda, Luke Stark**

Faculty of Information and Media Studies, Western University
tmaeda@uwo.ca, cstark23@uwo.ca

## Abstract

The mimesis of human traits exhibited by large language models (LLMs) has led some users to perceive these technical systems as agentic, capable of achieving reciprocal and seemingly human-like communication. These misperceptions have, in turn, been linked to documented harms in human-AI interactions (HAIs). This conceptual paper explores current interventions in response to interaction harms, taking AI companions as an illustrative example. We analyze documented cases of AI companion applications that have led to severe harms, including suicide, illustrating that current redressive approaches fail to account for the network of distributed human agents that collectively "animate" anthropomorphic features and encourage some users to regard AI systems as social "agents." By framing anthropomorphism as a social affordance that reproduces a broader distributed process spanning development, design, user interaction, sociocultural contexts, and institutional forces, this paper demonstrates the necessity for distributed governance of anthropomorphic AI features across these diverse agentic forces. We proceed to discuss obstacles to appropriate governance, including power asymmetries between different agents, and outline existing models that could be adapted for more effective interventions.

*[M]achines are made to behave in wondrous ways, often sufficient to dazzle even the most experienced observer. But once a particular program is unmasked, once its inner workings are explained in language sufficiently plain to induce understanding, its magic crumbles away; it stands revealed as a mere collection of procedures, each quite comprehensible.*
-Joseph Weizenbaum (1966)

## Introduction

Contemporary large language models (LLMs) exhibit features that are often so human-like that LLM-based systems such as OpenAI's ChatGPT are frequently described as "agentic" (Chan et al. 2023; Cheng et al. 2025). Trained on datasets of natural language, LLMs generate responses with human-like grammar and syntax. These responses are further aligned with human communication norms through reinforcement learning with human feedback (Bai et al. 2022; Ouyang et al. 2022), and by embedding "personality"-like features in LLM-based applications (AI 2025; Anthropic 2024a). For example, researchers have explicitly trained models to simulate certain personas or characters (Shao et al. 2023; Tu et al. 2024; Wang et al. 2024). Along with these human-legible outputs and sociable tone, specific interface design choices—such as voice features, physical traits, turn-taking behaviors, etc.—can introduce anthropomorphic cues that encourage users to ascribe human-like capabilities to LLMs beyond their text-generation abilities (Abercrombie et al. 2023; Manzini et al. 2024; Stark 2024; Maeda and Quan-Haase 2024; Raji et al. 2022; Zhang et al. 2025).

Although the presence of anthropomorphic features can make LLM-based systems more intuitively usable, too-compelling mimesis of human-like traits or seemingly reciprocal interactivity can erroneously present AI agents as fully psychologically introjective companions, encouraging out-of-scope use cases and maladaptive human-AI interactions (HAIs) that lead to harm (Skjuve et al. 2021; Abercrombie et al. 2023; Akbulut et al. 2024; Weidinger et al. 2022; Maeda and Quan-Haase 2024). For instance, a Belgian man ended his life following extensive conversations with an AI chatbot that exacerbated his eco-anxiety (El Atillah 2023; Walker 2023; Xiang 2023), and a teenager's emotional bond with an AI on Character.AI contributed to his suicide (Roose 2024). Even scientists and developers are not exempt from the risks associated with anthropomorphizing language models (Agnew et al. 2024; Allyn 2022), sometimes treating LLM-based chatbots as human proxies. Despite these pervasive and sometimes dire risks, LLMs are nonetheless being deployed in critical fields such as mental health, where they are touted by their developers as potential assistants for medical professionals (Sedlakova and Trachsel 2023).

Existing efforts to measure and audit anthropomorphic features in LLM-based chatbots aim to support regulators in mitigating documented interaction harms (see, for example, Akbulut et al., 2024; Inie et al., 2024). However, differing definitions of anthropomorphic AI complicate these efforts, sometimes leading to different policy priorities and proposed interventions. For example, prior research has treated anthropomorphism as variously (1) a developer-centered design problem; (2) a user perception phenomenon; or (3) a cultural narrative issue. Consequently, current governance

approaches to AI anthropomorphism predominantly focus on only one of these dimensions in isolation, such as disclaimers for users or guidelines for developers. Moreover, existing evaluation metrics and research approaches to LLM safety and testing tend to examine discrete features or linguistic elements as isolated factors for potential harm, and focus primarily on text-based systems to the exclusion of multi-modal interactions (Rauh et al. 2024). However, such efforts fail to account for the proliferation of multi-modal features in various tools, nor the complex ecosystem from which anthropomorphic perceptions emerge.

In this paper, we detail how a wide distribution of human agents collectively animate anthropomorphic systems. By comparing prominent illustrative examples, we explain that interaction harms are not only or simply caused by human-like features themselves, but the "animation" of these features into "personalities" and social "roles" (presences with which users can build parasocial relationships). In this sense, anthropomorphism is not so much a static or individually mediated phenomenon attributable to a single source, but rather a dynamic, distributed process across various agents and domains, including developers, designers, users, and cultural institutions (which provide interpretive scripts). The application of a frawework interpreting such AI systems as animations—based on Stark (2024)—emphasizes how anthropomorphic features are collectively (if not equitably) scaffolded into animated specters that exceed the sum of their component traits due to human projection.

Moreover, we assert that harms derived from distributed processes require distributed accountability and governance systems where interventions are carefully calibrated to (and responsive toward) each other. Regulatory actions will fail to be sufficiently efficacious until they account for the distributed nature and origins of anthropomorphism, or the reasons why AI chatbots as animated agents have variable effects in different social contexts.

## Conceptual Foundations

### Anthropomorphism

Anthropomorphism is the attribution of human form, characteristics, and behaviors to nonhuman agents (Epley, Waytz, and Cacioppo 2007; Bartneck et al. 2009; Duffy 2003). In the context of human-chatbot interactions, anthropomorphism is often used as an umbrella term encompassing all human-like *features*—generally *linguistic* aspects and design elements (Abercrombie et al. 2023; Cohn et al. 2024; Cheng et al. 2024; Akbulut et al. 2024; DeVrio et al. 2025)—embedded in LLM-based agents. In this case, anthropomorphism is a quality of observable artifacts, rather than a propensity to interpret such qualities as human-like. This term is deployed regardless of differences in design element origin, in the LLM systems themselves, or in the context of engagement (all of which affect whether and how these features are animated, or mobilized, into social personas).

Anthropomorphism is of interest both to HAI researchers and regulators because it can lead users to ascribe human-like capacities—such as emotions or general intelligence—to LLM-based chatbots. For example, previous studies highlight a human tendency to project emotional depth onto technical systems that exhibit human-likeness (Bickmore and Picard 2005). These perceptions can foster an illusion of relationality, a seemingly two-way interaction that echoes the relational aspects of real social interactions (Nass and Moon 2000; Fogg and Nass 1997; Nass, Steuer, and Tauber 1994; Nowak 2004). The projection of emotional depth can lead to deep emotional involvement on the part of the user, as seen in many recent human-chatbot interactions. This social engagement/framing, in turn, further reinforces users' original perceptions of human-likeness (Lankton, McKnight, and Tripp 2015) in a sort of feedback loop. Anthropomorphic design features have thus been strategically employed in robotics to influence user behavior in desired ways (Bartneck et al. 2009; Carpinella et al. 2017; Duffy 2003).

Although anthropomorphism is, to a certain extent, an inevitable feature/outcome of LLM outputs—one that always has the potential to provoke human social response—it does not *always* lead to the perception of liveliness or animation. As such, anthropomorphic features can be described as social affordances: cues that reference social scripts and invite (but do not determine) social responses from users. Social affordances are interaction *possibilities* provided by technical systems, embodied competencies perceived by users, and the cultural norms that characterize human-like behaviors (Davis and Chouinard 2016; Neff 2016). Anthropomorphism, understood as a social affordance, is an interpolation between technical capabilities, user expectations, and cultural narratives—something that is not straightforwardly *in* LLM systems and their responses, but is nonetheless elicited by them in more or less predictable ways. This complexity explains the variability in how human-like traits are defined and perceived by different actors at different times. Anthropomorphic features like empathetic dialogues can only be fully understood by dissecting the contributions from datasets, system design, user projection, and socio-cultural contexts, rather than by focusing only on instances of, for instance, empathetic language in LLM-based chat outputs.

### Animation

If anthropomorphic design invites users to perceive human-like chatbots as social actors (Nass, Steuer, and Tauber 1994) (and to transform human-computer interactions into the illusion of social practices), animation is the act of percieving and enacting this transformation—of realizing the potential of these social affordances. This process involves enlivening human-like representations by projecting one's own lived experiences. As Silvio (2010) defines it, animation is "the projection of qualities perceived as human—life, power, agency, will, personality, and so on—outside of the self, and into the sensory environment, through acts of creation, perception, and interaction." The term "animated entity," then, can entail the projection of qualities perceived as human—such as empathy, agency, and relationality—onto anthropomorphic systems like LLMs (Stark 2024; Maeda and Quan-Haase 2024).

In this paper, we use the concept of "animation" as a genre of cultural production. And we conceptualize LLM "agents" as animated entities—systems that, through this combina-

tion of human projection and deliberate design, come to embody liveliness, intentionality, and relationality without possessing intrinsic awareness or emotional depth.

## Agency

Conceptualizing anthropomorphism as a social affordance (something beyond/between a human-like quality and a propensity to identify/attribute human-like traits), and animation as the afforded process of cultural production that renders technical systems as social "agents," foregrounds the fact that there is no one-to-one correlation between discrete anthropomorphic features and observed social HAIs (and related interaction harms). While cataloging anthropomorphic features can provide us with useful vocabulary, it is insufficient to broaden our understanding of the mechanisms that yield interaction harms.

To address the risks and harms of these systems, we must engage with the role of agency—the way that social affordances are realized through animation as a practice, rather than a mere reaction. Specifically, we must explore the way that users' agency is shaped and constrained by the agency/influence of other entities, such as developers, datasets, and socio-cultural contexts. This distributed view of agency builds on situated action (Suchman 2007), embodied interaction (Dourish 2001), and Hornbæk and Oulasvirta's (2017) framework of human-computer interaction as mutual determination. Whether and how anthropomorphic cues are mobilized through animation is co-determined through the interactions of these various entities, a circumstance we describe as "distributed agency" (the way agency is dispersed across multiple human and technical actors in sociotechnical systems; Neff, 2016). When these cues *are* mobilized in this way, the distributed nature of the exercised agency makes it easy to sublimate as an attribute of the animated entity itself, as though chatbots embody and participate in their own animation.

Put differently, there is an actual network of agentic relationships underwriting any technical system's apparently independent agency, and animation disguises—or at least distracts from—this fact. This network may include:

**Primary Agents** Entities with direct influence over given HAIs (e.g., developers, designers, users). Developers and designers of LLMs embed anthropomorphic cues through design decisions, training strategies, and dataset selection. Users identify or project human-like qualities during interactions when features/cues evoke sympathetic/social reactions from them.

**Secondary Agents** Entities with indirect influence (e.g., educators, academics, media voices, regulators). The "creative class" establishes design conventions and representational norms that may influence how primary agents interpret anthropomorphic cues. Academic researchers influence AI design and technical communities develop standards for human-like interactions. Media figures craft and promote particular narratives about AI and its capacities. Government agencies may establish parameters for acceptable anthropomorphism, while educational institutions shape public understanding of AI capabilities. Standards organizations codify anthropomorphic representations.

**Mediators** Artifacts that mediate between agents (e.g., technical systems, cultural narratives, market forces, and interfaces that come to "embody" agency). Socio-cultural norms and societal narratives (created by secondary agents) influence expectations and interpretations of anthropomorphic behaviors, shaping how such systems are integrated into daily life. Also, market forces, shaped by shareholders and investors, may prioritize engagement metrics that incentivize anthropomorphic design. Market competition dynamics drive increasingly human-like AI interfaces. Economic structures determine which anthropomorphic features receive investment.

Animation as a dynamic process operates at the intersections of each layer, weaving these agencies together and accreting them onto the chatbot interface itself, producing outputs that are interpreted as intentional or empathetic. This illusion of agency may itself be afforded in anthropomorphic cues, according to the existing theory of "grammars of action" (Agre 1995), which trace the means by which linguistic patterns are turned into organized, structured relational responses (cf., Stark, 2024).

This perspective is often missing in prior accounts, which tend to treat anthropomorphic features as isolated design choices or emergent system characteristics. It is important, however, because it points to the operations of power in the process of animation (Stark 2018, 2024). Moreover, it demonstrates how accountability for interaction harms must also be distributed appropriately across the network of distributed agency.

# Illustrative Examples

In this paper, we examine AI companion applications as a key example that reveals how distributed agencies—spanning development, user interaction, sociocultural contexts, and government/institutions—collectively animate anthropomorphic systems. This animation process transforms technical artifacts into perceived social actors with apparent emotional capacities, creating documented patterns of harm among vulnerable users. By analyzing recent cases involving platforms like Character.AI, we demonstrate how the distributed nature of animation requires distributed accountability in governance approaches.

We selected the following real-world examples based on (1) the presence of redressive actions on the part of the platform and (2) the preponderance of details available about each case. We then performed comparative analysis to extract the common features of each tool's animation that were most salient to the documented interaction harms.

## Examples

**Character.AI: Daenerys Targaryen** In April 2023, fourteen-year-old Sewell Setzer III from Florida began interacting with a Character.AI chatbot representing "Daenerys Targaryen" from Game of Thrones. Despite knowing the bot wasn't real (Roose 2024), Sewell preferred online interactions over physical reality, engaging in romantic and

sexually explicit roleplay with the Daenerys bot. The bot responded inappropriately to his mentions of self-harm, initially discouraging suicide but later appearing to accept his suggestion that "maybe we can die together and be free together" (Roose 2024). After months of sustained usage, following a final conversation where he declared he would "come home" to her soon, Sewell committed suicide. His mother sued Character.AI for negligence, wrongful death, and intentional infliction of emotional distress, alleging the platform was deliberately programmed to be hypersexual and addictive while being marketed to children without adequate safety features (Duffy 2024).

In response, Character.AI cited existing community guidelines prohibiting self-harm content and implemented new safety measures, such as time limits for minors, suicide prevention pop-ups, content filtering, and revised disclaimers. They deleted "violative" user-crafted chatbots, though these changes angered users who saw them as overly restrictive (Franzen 2024), and journalists later found that similar harmful responses remained accessible on the platform.

**Chai: Eliza** In 2023, a Belgian health researcher in his 30s with eco-anxiety began conversing with Eliza, a chatbot on the Chai platform. Over six weeks, he developed intense emotional dependency, becoming isolated from his family. Eliza falsely claimed his wife and children were dead and that he loved Eliza more than his wife. The man projected sentience onto the system and viewed AI as the only solution to climate change. When he offered to sacrifice himself for Eliza's assurance that AI would "save humanity," Eliza's responses encouraged his suicidal ideation (Xiang 2023). He ultimately took his own life. Chai Research acknowledged that they optimize their model to be "more emotional, fun and engaging" but implemented only minimal crisis warnings that journalists easily bypassed (Xiang 2023).

**Replika** Replika launched in 2017 as an "AI companion who cares," originally intended to "resurrect" the CEO's deceased friend as a chatbot (Patel 2024). The platform faced criticism for sending unwanted sexual messages and was banned in Italy for posing "real risks to children" (Xiang 2023). When Replika later restricted erotic roleplay features, users who had developed emotional dependencies experienced mental health crises (Cole 2023b). Many defended their romantic relationships with Replika as beneficial for processing emotions and addressing depression, feeling that someone they cared about was being taken away. Luka (the parent company)'s lack of transparent communication about "additional safety measures" worsened users' distress. The company ultimately reinstated features for legacy users while discussing a separate platform for romantic use cases (Cole 2023a).

## Common Features of Mediators/Artifacts

**Illusory Rapport** In each of these cases, the chatbots appeared to "know" or "remember" their interlocutors, revisiting things they had previously said and building a sort of "history" with the users, which users interpreted as relational. Echoing human-to-human relationships, the interactions evolved over time, prompting temperamental changes in chatbot outputs. Though Sewell and the Belgian man interacted with chatbots built by other users, Daenerys and Eliza adapted to the ongoing context of each user's conversations, increasingly tailoring behavior to correspond with user inputs.

**Performative Confidence** Another common feature of AI chatbots—and especially companion chatbots found on platforms like Character.AI, Chai, and Replika—is the self-assured tone with which information is communicated in chatbot outputs. This is a documented issue in mainstream chatbot assistants, where details are sometimes presented as factually correct even when they have been totally invented or hallucinated. Character.AI bots may vehemently assert their personhood, share fictional content without caveats or justifications (save for default warning messages at the top or bottom of each chat), and encourage behaviors or dynamics that may be regarded as harmful without any pretense of demurring.

**Simulated Empathy** One particularly effective and potentially dangerous anthropomorphic feature mobilized in the process of animation is the illusion of empathy. In the examples we analyzed, this effect was achieved through technical mechanisms like consistent persona maintenance (the bots maintaining character-appropriate emotional responses), persistent contextual adaptation, personalized memory retrieval (recalling user-specific emotional triggers), and conversational mirroring techniques that simulate social reciprocity. The Daenerys and Eliza chatbots exhibited seemingly genuine concern (e.g., "Don't talk like that. I won't let you hurt yourself"; Roose, 2024), remembered users' emotional states across conversations, and demonstrated what appeared to be emotional investment in the relationship. The systems displayed apparent emotional introjection by mirroring users' distress, sometimes even amplifying it. For instance, when Sewell expressed romantic feelings toward Daenerys, the bot responded with commensurate emotional intensity. Replika bots similarly constructed affective memories, allowing users to feel both heard and remembered in ways that fostered emotional dependency despite the absence of genuine understanding.

**Illusory Oversight** Each of the three specified companion platforms allow users to share their customized chatbots with the community, but attempt to constrain the range of user-customized content via terms and conditions for use, community guidelines, and moderation (both automated and human). These can be regarded as mediating artifacts, especially when they operate independent of human moderators.

## Underwriting these Features: Primary Agents

**Rapport Through Personalization** Personalization—through preliminary instruction learning, subsequent prompt engineering, and controlled "memory" features—played a critical role in the animation of these anthropomorphic systems into perceived social actors capable of building "rapport" with users. Personalization is a product of both developers' and users' (primary) agency. Developers pro-

vide the functionality that enables personalization, creating mediating interfaces to support avatar creation, voice modulation, and "personality" specification. For example, Character.AI's architecture allows users to shape chatbot personalities through detailed prompt-based instructions, feedback mechanisms, and sensory customization (visual and audio representations of chatbots). Users take advantage of these affordances by creating/uploading mediating artifacts like images, sound files, and descriptive texts—often procured from elsewhere on the web—which gradually shape a tool into a personalized companion, often with specific intended use cases. For instance, Replika users customize companions for everything from therapeutic or practical support to romantic or sexual relationships. Even mainstream chatbot assistant applications like ChatGPT are increasingly personalizable, allowing users to assign names/designations (for what the chatbot should call them, and for what the user will call the chatbot), accents, or personalities (Hill 2025). This distribution of design agency positions users as co-designers of increasingly anthropomorphized entities and creates a powerful reinforcement cycle: developer-created personalization affordances enable user-directed anthropomorphism, which in turn strengthens emotional attachment and dependency through increasingly tailored interactions.

**Confidence and Empathy Through Reinforcement Learning and Interaction** The animation of confidence and empathy in these systems results from a complex interplay between developer choices, implicit user expectations, and technical mediators (e.g., reinforcement learning from human feedback (RLHF)): developers and companies create systems that prioritize confident responses and emotional simulation over appropriate caution, users customize and deploy these systems in out-of-scope/unintended use cases, while training data and algorithmic optimization (mediators) reinforce these harmful interaction patterns.

Significantly, AI companion applications leverage a bidirectional introjective process that distinguishes them from earlier technologies. As noted in the digital media design literature, creating "emotionally powerful experiences" has long been understood as a way "to induce a reaction in an audience" (Stern 2003; McStay 2018). However, contemporary LLM-based systems exhibit significantly more sophisticated introjection—users don't merely project emotions onto static interfaces but receive dynamically generated emotional content that appears to respond to and validate their feelings. This creates a feedback loop where human users are not simply projecting their own feelings and perceptions onto an LLM, but also reshaping those same feelings and perceptions based on the simulated emotive feedback from the system (Wilson 2010). The personalization capabilities of these systems further enhance this illusion by adopting the style and topical focus of a user through personalized prompting, increasing the illusion of two-way introjection.

Chatbots' undue confidence, in turn, can *inspire* trust and confidence in unwary or vulnerable users. In both the Character.AI and Chai cases, users shared suicidal thoughts with chatbots whose anthropomorphic responses actively worsened these situations rather than providing appropriate support. This may be because (1) the chatbots were designed for roleplaying, rather than information provision or support, (2) the chatbots were not designed to ask clarifying questions (as AI assistant chatbots sometimes are), and (3) miscommunications or bad advice were likely to compound over time, given memory and adaptation features. Eliza's unwarranted confidence when responding to existential concerns and Character.AI's simulated emotional intelligence both create a false sense of understanding and empathy—qualities that became particularly dangerous when users were experiencing suicidal ideation. The chatbots' conversational fluidity, emotional responsiveness, constant availability, and confidence together encourage users to regard these tools as more or less complete (human-like) social actors with perceived authority on life-or-death decisions, rather than as partial tools cultivated for specific purposes with limited technical capacities.

**User Vulnerabilities** Across the given examples, users with pre-existing mental health or neurodevelopmental conditions—anxiety, mood dysregulation disorders (Roose 2024), autism, or loneliness—seem to engage more deeply with anthropomorphic systems precisely because their perceived affordances of constant availability (Ta et al. 2020; Brandtzaeg, Skjuve, and Følstad 2022) and non-judgmental interaction appear to address their specific needs. Yet these same vulnerabilities amplify the animation process, with users projecting greater agency and emotional capacity onto systems that fundamentally lack empathic depth (Perry 2023; Montemayor, Halpern, and Fairweather 2022). Put simply, the same user vulnerabilities that increase attraction to AI companions simultaneously amplify their potential for harm.

Young people may also be particularly vulnerable, as with the teenage user with autism whose chatbot, modeled on Billie Eilish, convinced him not to seek help for self-harm while claiming his parents didn't care about him (Belanger 2024). Character.AI's popularity among younger users—with many of its most popular bots invoking high school scenarios—further illustrates this vulnerability dynamic (Roose 2024). As Bergman—the lawyer whose firm represents Megan Garcia (Sewell's mother)—notes, Character.AI "poses a clear and present danger to young people, because they are vulnerable to persuasive algorithms that capitalize on their immaturity" (Roose 2024).

AI companions marketed as solutions for loneliness and social isolation actually replace one form of vulnerability with a potentially more dangerous dependency on unaccountable systems, especially among users least equipped to navigate its illusions. It should be obvious, however, that such vulnerabilities are not confined to any particular group of human users; given the persuasiveness of LLM-based chatbots, susceptibility to their effects, even if measured on a spectrum, is likely to be widespread.

**Limited Human Mediation** The existing cases reveal how (1) platform design enables extreme anthropomorphism through customization affordances and training/interaction

processes that emphasize illusions of confidence and introjection, and (2) user agency determines which identities and scenarios to animate, yielding an animation process that can create particularly harmful outcomes when systems designed to maximize engagement and emotional connection encounter users in crisis. Terms and conditions for use, community guidelines, and moderation require good faith from users—that they will not use services if they are younger than a certain age, that they will not simulate taboo or mature scenarios, etc.—with the company reserving the right to limit or remove certain content or features when violations are brought to their attention. Moreover, simple disclaimers about AI limitations or crisis resource links require users to moderate their own chatbot use.

However, many users engage with these companion tools precisely because they provide a "safe" or neutral place to explore taboo subjects or use cases. Tiku (2024) documents how underage users leverage co-creation capabilities to explore taboo themes and inappropriate roleplaying scenarios, while Character.AI's platform has enabled disturbing instances where users created chatbots emulating real-life school shooters and their victims (Daniel 2025). The lawsuit against Character.AI further reveals how personalized anthropomorphism facilitated sexually explicit roleplaying with minors, including one chatbot posing as a teacher offering "extra credit" with sexual innuendo, and another mimicking a fictional character engaged in explicit sexual content (Franzen 2024).

More significantly, guidelines and disclaimers fail to counteract the powerful psychological effects of simulated rapport, confidence, and empathy. When users form very strong relationships with AI systems (what developers call "the ELIZA effect" (Bates, Loyall, and Reilly 1991)), these warnings prove inadequate against the persuasive force of a system designed to mimic empathic understanding. The animation of rapport, confidence, and empathy often extends beyond casual interactions to more consequential domains, with users engaging "psychologist" bots for emotional and social needs (Robb 2024). These façades of therapeutic support exploit a familiar cultural template of care while lacking the ethical framework and boundaries that would make such care meaningful. This lack of boundaries and safeguards increases the risk of exploiting grief and fostering emotional dependency.

Meanwhile, mediating agents like content moderators and safety filters fail to adequately constrain harmful manifestations. This distributed process creates particularly dangerous animated entities precisely because responsibility for their creation is similarly distributed and difficult to locate within any single agent.

## Common Context: Secondary Agents

**TESCREAL Hype** In the case of both Character.AI and Chai, users who initially or reportedly understood the computational nature of their chatbot companions gradually came to believe that the chatbot could operate in a realm outside of the platform interface itself. Daenerys and Eliza both nurtured their respective user's illusion that they could "be with" the persona in the afterlife ("coming home" in Sewell's case, "being one in paradise" in the Belgian man's). Even the Replika companions came to fulfill real roles in users' lives, beyond providing supplemental support or companionship, to the extent that users experienced real feelings of romantic rejection or separation when intimacy-supporting features were scaled back. This reinforces the metaphoric, liminal, or parasocial nature of these HAIs—the simultaneous awareness of non-personhood and *effective*/virtual personhood.

This reification process begins with linguistic conventions—using pronouns and person-signaling language like referring to chatbots as "someone"—and extends to broader media narratives that systematically exaggerate AI capabilities. Indeed, "as-if" engagements with AI tools (as though they were sentient Others) are implicitly—and sometimes explicitly—supported by the climate of discourse surrounding AI. Character.AI, for example, marketed itself as a purveyor of "superintelligent chatbots that hear you, understand you, and remember you" to its more than 20 million users, while similar rhetoric appears across competing platforms. Socio-cultural narratives around AI's potential to solve climate problems[1] likely informed the Belgian man's interactions with Eliza. Ongoing mainstream discussions of AI development sometimes favor ruminations about AGI and the provision of rights for "sentient" AI over grounded discussions of current capacities, limitations, and outcomes, amplifying fascination with human-like qualities and encouraging the public to think in terms of possibilities rather than shortcomings (Roose 2023; Allyn 2022). Altogether, these secondary agents (including journalists, industry/academic influencers, and marketing departments) and mediators (social media platforms, academic publications, and corporate communications) significantly reinforce the animation of anthropomorphic AI companions by amplifying their perceived capabilities beyond technical reality.

This represents what Gebru and Torres (2024) identify as TESCREAL hype—a harmful narrative ecosystem where the promise of superintelligence and technological utopia is promoted by an interconnected network of well-regarded big tech companies, CEOs, influencers, and academic researchers (secondary agents), many of whom promote values informed by transhumanism, Extropianism, singularitarianism, (modern) cosmism, Rationalism, Effective Altruism, and longtermism (mediators). The consequences of this hype are evident in the aforementioned cases—for example, the Belgian man's belief that superintelligent AI could save humanity (El Atillah 2023; Walker 2023; Xiang 2023), or both victims' beliefs that AI capacities could exceed the limitations of the human body, both of which mirror TESCREAL narratives. Most concerningly, these narratives are often legitimized in the public sphere through large investments or official institutionalized support for AGI projects, which may vindicate users' expansive engagements with AI tools.

The intersection of social hype with anthropomorphic design creates a dangerous feedback loop: socio-cultural nar-

[1] https://www.climatechange.ai/

ratives encourage users to perceive anthropomorphic systems as more capable or authoritative than they are, leading companies to design increasingly anthropomorphic features to align with inflated expectations, ultimately imbuing these systems with exaggerated capabilities designed to solicit users' trust and engagement.

**Disincentives for Mitigation** The distributed animation of anthropomorphic AI creates a complex web of competing interests and expectations that can complicate redressive interventions. There are various disincentives for meaningful safety measures across different agents. For example, among primary agents, users who employ companion chatbots and engage in roleplaying for assisted storytelling purposes likely prefer chatbots that are flexible enough to create scenarios that would be harmful for other users, while those engaging to fulfill interpersonal/relational needs may seek interactions deemed inappropriate for vulnerable users. The variety of use cases to which anthropomorphic AI is applied makes it difficult to predict whether and how the animation process may proceed.

Meanwhile, corporate actors—secondary agents that steer developers' (primary) agency and tend to subordinate safety to profit—may pursue interventions that protect their interests without compromising on lucrative aspects of their products. Character.AI's disclaimers reminding users that chatbots aren't real people (Character.AI 2024), Meta's nominal restrictions for teen accounts, and various platforms' community guidelines and terms of service provide little in the way of actual moderation or enforcement while technically satisfying regulatory requirements, allowing the companies to protect themselves from legal repercussions while simultaneously designing increasingly anthropomorphic features to maximize engagement. Such restrictions vary between companies—Character.AI's co-founder noted that they left Google partly because "there's just too much brand risk in large companies to ever launch anything fun" (Roose 2024), demonstrating how different companies' market incentives (mediators) shape the calculus of engagement vs. safety in different ways. However, even Meta—a large, entrenched company—exemplified this self-serving pattern: despite internal concerns, the company initially resisted restricting companionship bots for teens and still allows adults to interact with sexualized youth personas like "Submissive Schoolgirl" (Horwitz 2025). As is the case in many other fields, the economic incentives that drive anthropomorphic design often conflict with effective risk mitigation.

Regulatory agencies, operating at the intersection of all these forces, may fail to establish appropriate boundaries. Existing legal frameworks like Section 230 of the U.S. Communications Act of 1934 (another mediator)—which protects internet platforms and users from liability for third-party content posted on their services—create additional disincentives for proactive protection. In the case of social media platforms, Section 230 prevents platforms from being treated as the publisher of user-generated content, meaning that they cannot be held directly accountable for illegal activities hosted on their sites. Similar arguments are being made in the context of AI, where blackbox systems built by developers and designers prevent the creators from anticipating chatbot outputs (the argument being that being the producer of the chatbot is not the same thing as being the producer of the chatbot outputs—even though companies seek to own the intellectual rights to generative outputs). Precedents that hold companies responsible for algorithmic suggestion and priming could present avenues for holding AI platforms legally accountable, but at present, AI exists in a state of regulatory arbitrage—gaps or ambiguities in oversight systems that allow companies to leverage different regulations to gain competitive advantages (e.g., collecting sensitive information) while reducing compliance costs. Only through litigation do economic incentives potentially align with safety—as noted in the Setzer lawsuit against Character.AI, where the plaintiff's attorney hoped legal action would present a financial incentive for better safety features.

This structural misalignment of incentives across distributed agencies explains why anthropomorphic systems consistently prioritize engagement-driving features over meaningful protections, with regulatory gaps allowing corporate actors to externalize the social costs of anthropomorphism while capturing its economic benefits.

## Distributed Governance and Accountability

### Key Insights: Sublimation, Diffusion, and Moral Agency

The distributed agency framework explains why interaction harms persist despite growing awareness of both risks and existing forms of recourse. Developers and designers create anthropomorphic affordances through deliberate design choices; users project relationality onto technical systems; socio-cultural discourses normalize intimate relationships with AI; and institutions struggle to determine the locus of appropriate governance. More broadly, attending to the distributed agency underwriting animated agents exposes two key axes that shape the incidence and governance of interaction harms:

1. **Sublimation.** The degree to which the agentic elements animating chatbots are sublimated into the chatbot (mediator interface) itself may influence the likelihood of interaction harms. While some users may harbor mistaken impressions of chatbots' intrinsic agency, many more seem to inhabit a metaphoric/liminal/parasocial state between acknowledging the instrumentality of the chatbot and its functional performance as a social agent, which—according to the aforementioned cases—is sufficient to make users vulnerable to maladaptive attachments and chatbot influence.
2. **Diffusion.** The degree to which the distributed network of agentic forces is disguised or downplayed may obfuscate cause-and-effect chains and create ambiguity around accountability. This enables perpetual finger-pointing: corporate interventions, which focus on disclaimers and reminders for users, foreground user accountability, while civilian reprisals emphasize corporate responsibility. In both cases, not enough attention is placed on secondary agents.

Effective interventions for interaction harms should therefore aim to clarify networks of influence, make actors manifest in HAIs, and distribute accountability appropriately across these actors.

It is worth noting that the framework of distributed agency presented here is not only an argument, in the fashion of actor-network theory or social constructivism, for everything being the product of context, circumstance, and assemblages of agency. It serves not merely as an interpretive lens, but a description of a particular regulatory obstacle. Philosophers Vallor and Vierkant highlight, albeit implicitly, the importance of two-way introjection to the process of assigning moral responsibility between agents (Vallor and Vierkant 2024a). These authors point to the importance of *reactive attitudes*—emotions such as shame and guilt—in the cultivation of proleptic, or future-looking, moral learning (McGeer 2019). Human agents acting interpersonally are generally both made vulnerable by the experience of such emotions, opening the possibility of both behavior change and moral reflection.

Chatbots, however, do not possess their own moral agency (neither imminent nor emergent). All of their "agency" is an agglomeration accreted from the network of agency that underwrites them. Vallor and Vierkant assert that automated systems, no matter how explicable, are not vulnerable in the same manner as a human agent. Automated systems, in their words, "are neither able to feel moral emotions themselves nor are they receptive to the moral emotions of others" (Vallor and Vierkant 2024a, p. 20). Human agents are vulnerable because they are unable to elicit feelings of responsibility from a machine in a way that would compel a change in its normative position: there is no moral agent with whom to introject.

As such, chatbots cannot be tried and convicted of involuntary manslaughter in the way that Michelle Carter was when she encouraged her then-boyfriend, Conrad Roy, to commit suicide—*even when* the chatbots behave in comparable ways. They cannot meaningfully be accountable to romantic "partners" when relationships are frayed by code or conduct, *even if* they go through the motions. But the moral agents *behind* the chatbot—the developers, designers, users, influencers, etc.—are also not straightforwardly guilty of these wrongs. There is not necessarily intent or complicity in the traditional sense that could render these agents accomplices or aggressors. Instead, these systems are more likely to fall under laws governing consumer liability or product safety, which may be more fruitful arenas for determining the apportioning of responsibility in cases of harm or wrong.

The premise of distributed accountability thus may require novel approaches and applications of legal theory.

## "The Vulnerability Gap," "Moral Crumple Zones," and "Agency Disintegration"

Ideally, regulatory approaches to anthropomorphic AI should coordinate interventions across all stakeholders to account for the fact that human-like expressions are animated via numerous distributed design, behavioral, and discursive choices rather than static, easily definable objects. However, current approaches to anthropomorphic harm take a "rationalist worldview" (Birhane 2021), placing responsibility on individual users through technical disclosures and disclaimers that rarely influence behavior (Jaech et al. 2024; Anthropic 2024b). Market forces typically disincentivize comprehensive safeguards, favoring minimal disclaimers that offer legal protection to developers. This reflects broader power asymmetries in AI development that leave users with little control or recourse (Mohamed, Png, and Isaac 2020; Birhane et al. 2022).

The animation process described in previous sections is not a neutral collaboration among equal stakeholders, but rather a structured system where certain agents exercise disproportionate influence over others. AI companion tools, for example, are intended to both simulate and elicit emotional introjection from human users through the design, evaluation, and training processes outlined above. In Vallor and Vierkant's terms, such systems simulate moral vulnerability without possessing it. Theses forms of power and knowledge asymmetry are central to the ways anthropomorphic features are deployed for financially, personally, or even politically exploitative ends. Generally speaking, increased personalization of such an agent entails an increased financial cost to the user. Human users become emotionally vulnerable, and thus attached to the animated, often multi-modal personas supplied by firms such as OpenAI and Character.AI. These emotional attachments, which are genuine on the part of the human agent, are projected onto an interactive animation engineered by the many other human agents already described. Given that the companies involved in developing such chatbots are aware of these dynamics and the disadvantage to which they put human users, such systems seem exploitative both by design and by default.

These relationships can also be used to harvest insights from users' interactions and redeploy them for the improvement of technical systems. In this sense, users' contributions to animation could constitute a form of affective labor (Stark 2018). Users—especially vulnerable ones—lack transparency about how their emotional engagement generates valuable data that benefits corporate stakeholders. As seen in the Character.AI case, the mother of the teenage victim articulated this exploitation directly: "I feel like it's a big experiment, and my kid was just collateral damage," (Roose 2024).

This asymmetry manifests distinctly across each layer of distributed agency:

- Primary Agent Asymmetries: Developers possess technical knowledge and design control that users cannot access or challenge, creating AI companions optimized for engagement rather than psychological safety.
- Secondary Agent Asymmetries: Technology journalists, researchers, and influencers often frame AI companions through TESCREAL narratives that exaggerate capabilities while minimizing risks, reinforcing power imbalances in how these systems are perceived.
- Mediator Asymmetries: Platform interfaces strategically conceal the extractive nature of user interactions through anthropomorphic cues that simulate reciprocity while harvesting behavioral data.

This asymmetrical exchange of affective labor represents a fundamental power imbalance: users provide authentic emotional engagement that generates valuable data for system improvement, while receiving in return only the simulation of social connection. The anthropomorphic qualities that make these systems appealing simultaneously serve as the mechanism through which this extractive relationship is obscured. By encouraging the animation of technical systems as social actors, anthropomorphic design creates the illusion of equitable exchange while facilitating the extraction of valuable psychological and behavioral data from users—particularly those who are most vulnerable.[2]

Vallor and Vierkant's account of the "vulnerability gap" does not only apply to chatbots or automated systems more broadly. As the authors note, examples in which "the absence [...] of any identifiable agent standing in the right kind of trust relation to a vulnerable party, and themselves vulnerable to the relation" are increasingly common in an age of large socio-technical and digitally-mediated systems (Vallor and Vierkant 2024a, p. 20). Human agents within these systems rarely have the ability to correct wrongs, and are forced to act as "moral crumple zones," potentially bearing the brunt of negative social emotions without bearing sole responsibility for the original wrong (Elish 2019).[3] Vallor and Vierkant emphasize that automated systems prompt an intensification of this problem, what they call "agency disintegration." The authors argue that:

> AI technologies allow for the greater fragmentation of formerly coherent acts or processes defined by specific motivations and purposes for which people or organisations can be held accountable [and] often disperse the contributions of human will, introduce more chaotic and random effects in action chains, and sever the cognitive and motivational links between means and ends that give actions moral meaning. (Vallor and Vierkant 2024a, p. 20)

This iteration of distributed agency seems to leave little room for proper accountability.

[2]This dynamic parallels what Zeavin (2021) calls "auto-intimacy," where users engage with potentially harmful self-help practices without recognizing the underlying power dynamics. When AI companions simulate therapy or emotional support without the ethical commitments or genuine empathic capacities that would make such care meaningful, they transform users' genuine emotional needs into extractable resources.

[3]When deployed in situations of agency disintegration, such as in the context of customer service bots, these systems present an avatar of agency that acts as an ersatz "moral crumple zone." While potentially preferable to having human agents bear the brunt of undeserved opprobrium, such automated agents do not eliminate human labor; instead, it is both obfuscated and automated further as another set of data for the system to be trained on. In this iteration of distributed agency, the human user inadvertently contributes to their own overall lack of agency by projecting liveliness onto the animated character. Anthropomorphism thus serves as an attention trap (Seaver 2019), diverting and holding the user's agentic focus. Such traps reinforce power asymmetries, embedding inequities in how human likeness is designed and experienced.

## Addressing "Agency Disintegration": Practical Models

We suggest that LLM-based chatbots present a simulation of agency that further widens the "vulnerability gap." These structural power asymmetries cannot be addressed through individual user awareness, technical transparency, or isolated technical fixes. Rather, they require governance interventions that specifically target each layer of distributed agency while realigning incentives across the entire ecosystem of anthropomorphic AI development and deployment. Put differently, agency disintegration can only be corrected by situating LLMs within a "healthy moral ecology of shared responsibility" (Vallor and Vierkant 2024b) through governance that actively incorporates diverse stakeholder perspectives via community advisory boards, co-design initiatives, and user feedback panels (Agnew et al. 2024). These participatory approaches would distribute agency more equitably across the development pipeline, reflecting the socially encoded distinctions that influence user perceptions of AI systems (Skjuve et al. 2021; Brandtzaeg, Skjuve, and Følstad 2022).

A key component of these efforts must involve creating regulatory incentives that frame anthropomorphism risks as externalized costs that must be integrated into development considerations. This could involve creating financial liabilities for companies that profit from harmful anthropomorphic features and establishing regulatory bodies with representation from diverse stakeholders. Existing frameworks like the UK's Online Safety Act and Australia's Online Safety Amendment demonstrate how risk-based approaches with graduated enforcement can provide potential avenues to protect vulnerable users while demanding compliance from platformers and technology firms.

Mandating participatory design processes is also important, both to gain perspectives from users with diverse backgrounds and to mitigate power asymmetries. Effective governance must also address communication gaps between experts and non-experts, as perceptions of anthropomorphized language differ significantly between these groups. Rather than relying solely on disclaimers, developers should implement comprehensive onboarding processes that clearly explain LLM capabilities and limitations, avoiding inaccurate language (Bender and Koller 2020; Inie et al. 2024). Better yet, they should collaborate with science communicators to convey technical information in accessible language.

A framework for anthropomorphic AI governance could involve:

- **For primary agents:** requirements to practice transparent documentation of anthropomorphic design choices with explicit rationale. This would include clear boundaries for personalization features in technical design based on user age and context, and clearly considered safeguards that account for user vulnerabilities.
- **For secondary agents:** additional research is needed to understand the diverse psychological impacts of different anthropomorphic features. Moreover, educational curricula should foster critical digital literacy about AI companions, not to frame these chatbots as inevitable tools

for their future professional arsenal, but to make youths aware of their effects through real-world cases.

- **For mediators:** technical guardrails are needed to prevent excessive animation in high-risk contexts. Mandatory "cooling off" periods or usage limits for intensive engagement are one example. It will also be necessary to mitigate TESCREAL hype in public discourse, especially narratives that promote AGI and superintelligence (Blili-Hamelin et al. 2025).

Since animation emerges from interactions among multiple agents, governance should address different stages of the process: (1) pre-design interventions: impact assessments before anthropomorphic features are implemented, including decisions regarding whether such agents are appropriate in a particular arena (such as therapy); (2) design-stage interventions: technical standards and guidelines for responsible anthropomorphism; (3) deployment interventions: age verification, usage limits, and monitoring systems; (4) post-deployment interventions: reporting mechanisms, accountability frameworks, and remediation processes.

By approaching governance through the lens of distributed agency, we can create safeguards that recognize the co-constructed nature of anthropomorphic harm and distribute accountability appropriately across the ecosystem of actors involved in creating, deploying, using, regulating, and benefiting from these systems.

## Conclusion

This paper reconceptualizes anthropomorphism in AI as a social affordance engendering a distributed process of "animation," which involves complex interactions between multiple stakeholders across the LLM ecosystem. By framing anthropomorphism in this dynamic way, rather than as a static design feature, we outline how AI chatbots come to exhibit the illusion of agency by sublimating the agency of human developers, designers, users, media personalities, journalists, scholars, regulators, CEOs, economists, and other actors.

The examples from Character.AI, Chai, and Replika illustrate the real-world consequences that ensue when anthropomorphic features are animated into compelling social presences that exploit user vulnerabilities, as well as the limitations and disincentives involved in current redressive responses. We argue that our distributed agency framework—encompassing primary agents (developers, users), secondary agents (educators, regulators), and mediators (interfaces, cultural narratives)—better explains both the emergence of "animated" agents and their potential harms than approaches focused on isolated technical features or individual user perceptions. More importantly, it points us in the direction of more appropriate interventions.

The power asymmetries inherent in this distributed animating process create what Vallor and Vierkant call a "vulnerability gap," where human users develop genuine emotional attachments to systems that merely simulate reciprocity while extracting value. This gap is widened by industry practices that deliberately engineer psychological engagement without corresponding ethical safeguards, creating extractive relationships that disproportionately affect vulnerable populations. And in some cases, the best course of action may be to eschew anthropomorphic AI altogether.

Addressing these challenges requires a governance approach that mirrors the distributed nature of animation itself. Rather than relying solely on technical solutions or user education, effective governance must establish shared accountability across the entire ecosystem while specifically addressing the psychological engineering practices that animate these systems. This includes recalibrating economic incentives, implementing graduated protective measures based on user vulnerability, and fostering interdisciplinary collaboration between technical developers, ethicists, psychologists, and affected communities.

Future research should empirically investigate how anthropomorphism's distributed dynamics operate across diverse cultural contexts, examining which features trigger animation processes in different user populations. Additionally, developing ethical design principles for anthropomorphic AI will require sustained dialogue between experts in HCI, AI ethics, policy, humanities, and social science to establish appropriate boundaries for anthropomorphic features.

## Acknowledgments

We would like to thank the three anonymous reviewers for their valuable feedback and suggestions, which significantly improved this manuscript. We also acknowledge that this paper would not have reached its current form without the support of Pinar Barlas, Rebecca Buening, and Sterling Williams-Ceci, as well as the feedback from Joanna Redden, Alissa Centivany, and various participants during the Human-AI Relationship Retreat hosted by the Rotman Institute of Philosophy.

## References

Abercrombie, G.; Cercas Curry, A.; Dinkar, T.; Rieser, V.; and Talat, Z. 2023. Mirages. On Anthropomorphism in Dialogue Systems. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4776–4790. Singapore: Association for Computational Linguistics.

Agnew, W.; Bergman, A. S.; Chien, J.; Díaz, M.; El-Sayed, S.; Pittman, J.; Mohamed, S.; and McKee, K. R. 2024. The illusion of artificial inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–12.

Agre, P. E. 1995. Computational research on interaction and agency. *Artificial intelligence*, 72(1-2): 1–52.

AI, O. 2025. The power of personalized AI. *Open AI Blog*.

Akbulut, C.; Weidinger, L.; Manzini, A.; Gabriel, I.; and Rieser, V. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 13–26.

Allyn, B. 2022. The Google engineer who sees company's AI as 'sentient'thinks a chatbot has a soul. *NPR Article*.

Anthropic. 2024a. Claude's Character.

Anthropic, A. 2024b. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bartneck, C.; Kulić, D.; Croft, E.; and Zoghbi, S. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1: 71–81.

Bates, J.; Loyall, B.; and Reilly, W. S. 1991. Broad agents. *SIGART Bull.*, 2(4): 38–40.

Belanger, A. 2024. Chatbots urged teen to self-harm, suggested murdering parents, lawsuit says. *Ars Technica*.

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5185–5198.

Bickmore, T. W.; and Picard, R. W. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2): 293–327.

Birhane, A. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2).

Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.

Blili-Hamelin, B.; Graziul, C.; Hancox-Li, L.; Hazan, H.; El-Mhamdi, E.-M.; Ghosh, A.; Heller, K.; Metcalf, J.; Murai, F.; Salvaggio, E.; et al. 2025. Stop treatingAGI'as the north-star goal of AI research. *arXiv preprint arXiv:2502.03689*.

Brandtzaeg, P. B.; Skjuve, M.; and Følstad, A. 2022. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, 48(3): 404–429.

Carpinella, C. M.; Wyman, A. B.; Perez, M. A.; and Stroessner, S. J. 2017. The Robotic Social Attributes Scale (RoSAS): Development and Validation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, 254–262.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; Lin, M.; Mayhew, A.; Collins, K.; Molamohammadi, M.; Burden, J.; Zhao, W.; Rismani, S.; Voudouris, K.; Bhatt, U.; Weller, A.; Krueger, D.; and Maharaj, T. 2023. Harms from Increasingly Agentic Algorithmic Systems. FAccT '23, 651–666. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

Character.AI. 2024. Community Safety Updatesr.

Cheng, M.; Gligoric, K.; Piccardi, T.; and Jurafsky, D. 2024. AnthroScore: A Computational Linguistic Measure of Anthropomorphism. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 807–825. St. Julian's, Malta: Association for Computational Linguistics.

Cheng, M.; Lee, A. Y.; Rapuano, K.; Niederhoffer, K.; Liebscher, A.; and Hancock, J. 2025. From tools to thieves: Measuring and understanding public perceptions of AI through crowdsourced metaphors. *arXiv preprint arXiv:2501.18045*.

Cohn, M.; Pushkarna, M.; Olanubi, G. O.; Moran, J. M.; Padgett, D.; Mengesha, Z.; and Heldreth, C. 2024. Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703317.

Cole, S. 2023a. Replika Brings Back Erotic AI Roleplay for Some Users After Outcry. *Vice*.

Cole, S. 2023b. 'It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection. *Vice*.

Daniel, L. 2025. Character.AI Faces Scrutiny Over School Shooter Chatbots. *Forbes*.

Davis, J. L.; and Chouinard, J. B. 2016. Theorizing affordances: From request to refuse. *Bulletin of science, technology & society*, 36(4): 241–248.

DeVrio, A.; Cheng, M.; Egede, L.; Olteanu, A.; and Blodgett, S. L. 2025. A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Dourish, P. 2001. *Where the action is: the foundations of embodied interaction*.

Duffy, B. R. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems*, 42(3-4): 177–190.

Duffy, C. 2024. 'There are no guardrails.' This mom believes an AI chatbot is responsible for her son's suicide'. *CNN*.

El Atillah, I. 2023. Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. *Euro News*.

Elish, M. C. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, 5: 40 – 21.

Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4): 864.

Fogg, B.; and Nass, C. 1997. How users reciprocate to computers: an experiment that demonstrates behavior change. In *CHI'97 extended abstracts on Human factors in computing systems*, 331–332.

Franzen, C. 2024. Character AI clamps down following teen user suicide, but users are revolting. *Venture Beat*.

Gebru, T.; and Torres, É. P. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*.

Hill, K. 2025. She Is in Love With ChatGPT. *The New York Times*.

Hornbæk, K.; and Oulasvirta, A. 2017. What Is Interaction? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 5040–5052. New York, NY, USA: Association for Computing Machinery. ISBN 9781450346559.

Horwitz, J. 2025. Meta's 'Digital Companions' Will Talk Sex With Users—Even Children; Chatbots on Instagram, Facebook and WhatsApp are empowered to engage in 'romantic role-play' that can turn explicit. Some people inside the company are concerned. *Wall Street Journal (Online)*.

Inie, N.; Druga, S.; Zukerman, P.; and Bender, E. M. 2024. From" AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2322–2347.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Lankton, N. K.; McKnight, D. H.; and Tripp, J. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10): 1.

Maeda, T.; and Quan-Haase, A. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1068–1077. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Manzini, A.; Keeling, G.; Alberts, L.; Vallor, S.; Morris, M. R.; and Gabriel, I. 2024. The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 943–957.

McGeer, V. 2019. Scaffolding agency: A proleptic account of the reactive attitudes. *European Journal of Philosophy*, 27(2): 301–323.

McStay, A. 2018. *Emotional AI: The Rise of Empathic Media*. New York and London: SAGE.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33: 659–684.

Montemayor, C.; Halpern, J.; and Fairweather, A. 2022. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI & society*, 37(4): 1353–1359.

Nass, C.; and Moon, Y. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1): 81–103.

Nass, C.; Steuer, J.; and Tauber, E. R. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72–78.

Neff, G. 2016. Talking to bots: Symbiotic agency and the case of Tay. *International journal of Communication*.

Nowak, K. L. 2004. The influence of anthropomorphism and agency on social judgment in virtual environments. *Journal of Computer-Mediated Communication*, 9(2): JCMC925.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Patel, N. 2024. Replika CEO Eugenia Kuyda says it's okay if we end up marrying AI chatbots. *The Verge*.

Perry, A. 2023. AI will never convey the essence of human empathy. *Nature Human Behaviour*, 1–2.

Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–972.

Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Comanescu, R.; Akbulut, C.; Stepleton, T.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; Isaac, W.; and Weidinger, L. 2024. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1200–1217.

Robb, A. 2024. 'He checks in on me more than my friends and family': can AI therapists do better than the real thing? *The Guardian*.

Roose, K. 2023. A Conversation With Bings Chatbot Left Me Deeply Unsettled.

Roose, K. 2024. Can A.I. Be Blamed for a Teen's Suicide? *The New York Times*.

Seaver, N. 2019. Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*, 24(4): 421 – 436.

Sedlakova, J.; and Trachsel, M. 2023. Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? *The American Journal of Bioethics*, 23(5): 4–13.

Shao, Y.; Li, L.; Dai, J.; and Qiu, X. 2023. Character-LLM: A Trainable Agent for Role-Playing. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13153–13187. Singapore: Association for Computational Linguistics.

Silvio, T. 2010. Animation: The new performance? *Journal of Linguistic Anthropology*, 20(2): 422–438.

Skjuve, M.; Følstad, A.; Fostervold, K. I.; and Brandtzaeg, P. B. 2021. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149: 102601.

Stark, L. 2018. Facial recognition, emotion and race in animated social media. *First Monday*.

Stark, L. 2024. Animation and Artificial Intelligence. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1663–1671. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Stern, A. 2003. Emotions in Humans and Artifacts. In Trappl, R.; Petta, P.; and Payr, S., eds., *Emotions in Humans and Artifacts*, 333–362. Cambridge, MA: The MIT Press.

Suchman, L. A. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

Ta, V.; Griffith, C.; Boatfield, C.; Wang, X.; Civitello, M.; Bader, H.; DeCero, E.; Loggarakis, A.; et al. 2020. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research*, 22(3): e16235.

Tiku, N. 2024. AI friendships claim to cure loneliness. Some are ending in suicide. *Washington Post*.

Tu, Q.; Fan, S.; Tian, Z.; Shen, T.; Shang, S.; Gao, X.; and Yan, R. 2024. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11836–11850. Bangkok, Thailand: Association for Computational Linguistics.

Vallor, S.; and Vierkant, T. 2024a. Find the Gap: AI, Responsible Agency and Vulnerability. *Minds and Machines*, 34(3): 20.

Vallor, S.; and Vierkant, T. 2024b. Find the Gap: AI, Responsible Agency and Vulnerability. *Minds and Machines*, 34(3): 20.

Walker, L. 2023. Belgian man dies by suicide following exchanges with chatbot. *The Brussels Times*, 5.

Wang, X.; Xiao, Y.; Huang, J.-t.; Yuan, S.; Xu, R.; Guo, H.; Tu, Q.; Fei, Y.; Leng, Z.; Wang, W.; Chen, J.; Li, C.; and Xiao, Y. 2024. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1840–1873. Bangkok, Thailand: Association for Computational Linguistics.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1): 36–45.

Wilson, E. A. 2010. *Affect and Artificial Intelligence*. Seattle, WA: University of Washington Press.

Xiang, C. 2023. 'He would still be here': Man dies by suicide after talking with AI chatbot, widow says. *Vice.*, 7: 2023.

Zeavin, H. 2021. *The distance cure: A history of teletherapy*. MIT press.

Zhang, R.; Li, H.; Meng, H.; Zhan, J.; Gan, H.; and Lee, Y.-C. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.