When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design

Takuya Maeda*
tmaeda@uwo.ca
Western University
London, ON, Canada

Anabel Quan-Haase†
aquan@uwo.ca
Western University
London, ON, Canada

ABSTRACT

With the continuous improvement of large language models (LLMs), chatbots can produce coherent and continuous word sequences that mirror natural human language. While the use of natural language and human-like conversation styles enables the use of chatbots within a range of everyday settings, these usability-enhancing features can also have unintended consequences, such as making fallible information seem trustworthy by emphasizing friendliness and closeness. This can have serious implications for information retrieval tasks performed with chatbots. In this paper, we provide an overview of the literature on parasociality, social affordance, and trust to bridge these concepts within human-AI interactions. We critically examine how chatbot "roleplaying" and user role projection co-produce a pseudo-interactive, technologically-mediated space with imbalanced dynamics between users and chatbots. Based on the review of the literature, we develop a conceptual framework of parasociality in chatbots that describes interactions between humans and anthropomorphized chatbots. We dissect how chatbots use personal pronouns, conversational conventions, affirmations, and similar strategies to position the chatbots as users' companions or assistants, and how these tactics induce trust-forming behaviors in users. Finally, based on the conceptual framework, we outline a set of ethical concerns that emerge from parasociality, including illusions of reciprocal engagement, task misalignment, and leaks of sensitive information. This paper argues that these possible consequences arise from a positive feedback cycle wherein anthropomorphized chatbot features encourage users to fill in the context around predictive outcomes.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

trust, chatbots, anthropomorphism, human-AI interactions, ethics, parasociality, design

ACM Reference Format:
Takuya Maeda and Anabel Quan-Haase. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658956

In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3630106.3658956

1 INTRODUCTION

Chatbots are increasingly being introduced into a range of day-to-day tasks, from creative activities [66] to academic writing [10] to information retrieval [57]. This is largely because these systems use natural language and question-and-answer formats to interact with users in ways that resemble human-to-human interpersonal communication, enabling users to perform tasks via common-sense instructions. In dialogue systems, human-like responses have been shown to improve user engagement [71]. While this innovation may improve usability, it could also increase human-computer interaction harms [73], as human-like responses may prime users to interact with information in specific, socially coded ways. The anthropomorphic features characteristic of such systems introduce an additional dimension into human-AI interactions (HAIs), as they lead users to "engage" with AI agents themselves, rather than just their outputs. Users have been observed projecting gender stereotypes onto dialogue systems [2] based on given conversational cues, and have even engaged in identity-based harassment (e.g., slurs, sexualization, etc.) while interacting with these systems [59].

Users' social framing of AI systems, and the context in which information-seeking occurs, might affect how they regard the system's outputs [61]. Positive or friendly interactions with, or perceptions of, the agent can inspire unearned trust in the agent's outputs [2], whereas biases projected toward the agent can undermine trust in the agent's outputs, regardless of the actual quality of information [12].[1] And because chatbots are designed with interfaces and intended to fill conversational roles, their outputs are packaged as something more than just predictive inferences, potentially leading users to regard outputs as definitive answers, advice, or consolations. These problems with output appraisal are compounded by the "hype" around AI, which overestimates AI systems' capabilities [58] and encourages positive receptions of AI outputs, as well as by incidental inference errors [32] and malicious use cases [21] that make it difficult to distinguish factual from nonfactual content.

Current research directions that try to mitigate AI harms tend to focus on debiasing LLMs or addressing problems around models [3]. However, it is also important to address how design choices deflect users' attention from the veracity of generated content or the interpretability of generative processes to the affective trustworthiness of conversational agents themselves. These questions

[1] It is important to note that it can be difficult to appraise the quality of generated information, as responses do not provide context or clues to indicate how comprehensive the generated answer is [44].

will only become more urgent as the performance of large language models (LLMs), such as the GPT series [8, 55, 56] and PaLM [13], improves, enabling increasingly human-like conversation styles in downstream conversational AI applications like ChatGPT and Bard.

The aim of the present paper is to conceptualize the mediated, interactive spaces between users and AI systems and the modes of interaction (e.g., trust, emotional attachment) that govern them, assembling a conceptual framework that can be tested and refined in subsequent studies. We begin in Section 2 by synthesizing insights from existing literature across the fields of human-computer interaction (HCI), media studies/communication, and AI ethics, which tend to over-emphasize the determining role of either technical affordances or user adaptations in HAI. We then describe these interactions in a way that emphasizes how chatbot design and user motivations enable and constrain each other, adopting the concept of parasociality [16, 26, 28, 29] and expanding it to encompass HAI. This concept can shed light on how anthropomorphic features operate as social affordances [14, 15, 48, 53] that simulate reciprocal engagements and foster a sense of trust between users and chatbots.

In Section 3, we proceed to interpret examples of anthropomorphism in chatbots, as well as users' perceptions and social behaviors, through the lens of parasociality, social affordance, and trust. By bringing these elements together, we identify the centrality of role-playing/role assignment, projective (rather than just predictive) inference, and parasocial trust, which mediate the process whereby HAIs become parasocial relationships with ascribed (and assumed) social roles. In Section 4, we explore some of the implications of this dynamic, and in Section 5, we identify future research directions that could apply or improve our model.

This is not a systematic review, but an exploratory search for theoretical alignments that could establish a justification for further inquiry in this area. Few other descriptions of HAI have utilized the concepts of parasociality and trust, which help to explain how users habitually calibrate their behaviors according to given conversational dynamics. The resulting conceptual framework can be used to dissect the affective design embedded in conversational AI systems, alongside its effects for user agency.

2 BACKGROUND

2.1 Parasociality

The concept of parasocial relationships has been studied for decades in the field of media and communication. It refers to an asymmetrical, one-sided relationship between individuals and media personalities, real/fictional characters, or celebrities [28] wherein the individual experiences a personal connection with the media figure despite having little-to-no interpersonal interactions with them [29]. Contrast this with genuine social relationships, which are reciprocal if not always equal.[2]

Parasocial relationships are depicted as an illusion, though there may be actual performative "scripts" available to realize the relationship, making it legible to others [26]. For example, discourse/genre conventions and storytelling strategies can construct mutual understanding between audiences and producers, even without direct interaction [18]. Importantly, despite their illusory nature, parasocial relationships frequently rely on some expectation of authenticity.[3] Less staging and more perceived "naturalness" have been shown to improve perceived trustworthiness among young audiences [42], trust being an essential element of parasocial and genuinely social relationships alike. Transgressions in the authenticity of parasocial relationships can generate feelings of genuine betrayal. Consider, for instance, the case of Lonleygirl15, a YouTube video series in the mid-2000s that featured content about people's mundane daily experiences. When audience members learned that the main character was fictional—a role performed by an actress—they expressed resentment and disappointment, even attacking the actress for her "fake identity" [5].

Feelings of trust and mutual understanding may approximate the feeling of reciprocity experienced in real social relations, but they can also increase individuals' receptivity to information derived from these trusted sources [18]. Moreover, parasocial relationships can have real effects on human emotions and self-perception. For instance, Galbraith [23] investigated a simulation videogame called *LovePlus*, which invites players to construct a pseudo-relationship with a character. As players repeatedly interacted with these virtual characters, they reported feelings of affirmation that compensated for their perceived lack of "masculinity" in real life. As such, parasocial relationships are a topic of practical and ethical concern.

For the purposes of this discussion, it is worth drawing a distinction between parasocial relationships between real people (as with a celebrity and their fans) and those between real people and representations of people (as with fictional characters, animations, holograms, and—arguably—certain conversational agents). While both celebrities and fans have real stakes and responsibilities in the parasocial relationship, as when a celebrity benefits financially from their audience so long as they continue to fulfill the audience's expectations, representations of people cannot have such stakes or responsibilities (and users do not necessarily expect them to). Repeated HAIs that begin to stabilize into social roles could fall into this latter category. Therefore, parasociality could provide a theoretical lens to understand how the mediated spaces constructed by generative AI systems and their users can become *relational* spaces that affect human emotions or exploit them to gain trust.

2.2 Social Affordance

The concept of affordances originated in the field of cognitive psychology, where it referred to the possibilities for action offered by a given environment. It has subsequently been applied in a wide range of social science fields, including media studies, where it designates the possibilities for action available within a given platform. However, there is a second strand of affordance research that emphasizes *perceived* or *imagined* affordances—the actions that an individual

[2]We might add the distinction that genuine social relationships are more visible than parasocial relationships, and are more likely to be implicated in other social relationships. There are, of course, grey areas and exceptions, as when parasocial relationships become visible to (and thus verifiable by) others through open discussion, or implicated in other social relationships (for instance, in fan communities), or when real people engage in an online reciprocal relationship that remains invisible to others and removed from other social relationships (as per digital intimacy). However, this definition suffices to establish a spectrum.

[3]This expectation of authenticity is also frequently discussed in AI-mediated communication; for instance, people tend to respond negatively to AI-generated images, especially if they initially thought these images were human-made [31]. At present, many people consider standalone AI-generated content to be untrustworthy [31, 70].

believes are available to them based on their impressions of a given environment, platform, social situation, etc. According to Norman [53], possible actions are not totally determined by the physical constraints of systems, but are also shaped by the non-physical functions initiated by users [53]. Nagy and Neff [48] explored this type of affordance that is shaped by users' perceptions, explaining that the ways in which users interact with tools and platforms may be unrelated to the built-in technical features.

These variations on the concept of affordance are not mutually exclusive. For example, Davis and Chouinard [15] argued that the possibilities users perceive for action change based on the circumstances, such as available platform mechanisms that enable actions like "request," "demand," "encourage," "refuse," etc. Correspondingly, McGrenere and Ho [43] discussed how usable design, or the functions provided by systems, could signal what is possible to end-users. Davis [14] summarized this point by illustrating the interplay between humans and technical systems, paying attention to the ways in which they co-constitute capabilities.

In this paper, we regard anthropomorphism as a feature of conversational AI that affords (and is seen to afford) human-like, reciprocal interactions between chatbots and users. Indeed, Fogg and Nass [22] found that users projected human-like potentials onto computational systems—and used polite, socially-coded language with these systems—when they found the system to be helpful or productive. In other words, users determined the potential for reciprocity (or human-like interaction) based on the adequacy of computing systems' performance. But as asserted by Norman [53] and Nagy and Neff [48], the social affordance of reciprocal engagement does not indicate a genuine *capacity* for reciprocity in chatbot systems. Bender and Koller [7] argue that meaning, or communicative intent, arises from a relationship between linguistic forms and external cues. Because anthropomorphized chatbots utilize linguistic forms independently of external cues, they do not communicate meaningfully or intentionally, and so cannot participate in genuinely reciprocal conversations (with shared referents, objectives, etc.). Instead, the potential for reciprocity is an illusion built upon the chatbot's simulation of natural conversational behaviors. Insofar as they deploy human language and communication conventions, including pronouns that imply subjectivity, chatbots can easily invoke commonplace social behaviors from users.[4] This might be especially true if chatbots possess telepresence in the form of human-like avatars or physical characters [27]—technical features that may induce users to interpret interactions through the lens of relations.

By drawing attention to what is allowed or constrained in HAI, as well as how users perceive this mediation and envision their possibilities for action, affordance theory can shed light on how chatbots *and* users establish and sustain seemingly reciprocal interactions, which may develop into parasocial relationships. This paper attempts to connect affordance theory with emotional contexts, attending to how design features foster emotional interactions with chatbots that can, under certain circumstances, build trust and a feeling of mutual understanding. It is particularly crucial to pay attention to contexts and cues that encourage users to be vulnerable, as these can have the greatest effect on users' behavior and self-perception.

2.3 Trust

In the context of this paper, trust is defined as a social emotion that facilitates relationship-building by establishing a sense of safety or dependability. This corresponds with *affective* trust, as opposed to cognitive trust, which arises from a sense of competence or legitimacy. Madsen and Gregor [41, p.6] defined trust as "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid." This makes trust an important variable in HCI.

Trust is highly malleable, changing in response to external cues and internal emotional states. Even cognitive trust can be quite changeable, as some psychological research suggests that people rely on repetition, rather than documented facts, as a heuristic cue to check knowledge [20]. Madsen and Gregor [41] suggested that there is a positive relationship between users' perceptions of their own technical competence and their trust in computational systems. This means that, the more capable a user perceives themselves to be while using a tool, the more they trust the tool that they are using. It also implies that functions that increase self-efficacy can increase users' trust in computational systems. For example, users are more likely to trust outputs from AI systems if there is some transparency about the decision criteria, or if original data is made accessible to users [46]. This is not necessarily just because this increases the legitimacy of the output, as per cognitive trust, but also because it increases the users' perceived ability to meaningfully participate in the interaction. Other studies have yielded similar results, with people attributing greater trustworthiness to interpretable models than black-box models [4], regardless of whether they actually verify these models' recommendations.[5] This is consistent with the notion of perceived affordances [43], as greater self-efficacy might also imply more possibilities for action in a given technical system, and therefore greater engagement in that system.

Trust is essentially the glue between each of the previous concepts, as it orients people to their social setting, thereby affecting how they configure the social presence of chatbots, how and what they communicate, and whether they begin to attribute some socially stable role to a chatbot conversation partner (as in a parasocial relationship), in addition to informing their perceptions of platform affordances. The concept of trust can help to elucidate the emotional context in which users interpret or interact with retrieved information, and how anthropomorphic features shape these contexts.

3 WHEN HUMAN-AI INTERACTIONS BECOME PARASOCIAL RELATIONSHIPS

HCI typically entails a utilitarian "relationship" wherein a person uses a computational system for some end. The system demands nothing from the user except for literacy in its functions. In a

[4]Consider, for example, some users' tendency to refer to chatbots as "he," "she," or "they."

[5]Compare this to Kidd and Birhane [34]'s discussion of chatbot certainty—that is, chatbots' tendency to always provide definitive responses, even if they are inaccurate—which tended to convince users to accept given outputs, even without transparency.

parasocial relationship, however, a user may feel compelled to communicate in specific ways *beyond* what is necessary for inputs to be effectively interpreted by the computational system, likely because the user regards the system as an Other, rather than as a means to an end. This could be due to convention (for example, default politeness in linguistic interactions), playfulness (for example, the desire to treat the system "as if" it were a person, even if one knows that it is not a person), immersion or the suspension of disbelief (as though entering a liminal space where the chatbot is functionally a person), or any other reason. It does not necessarily imply outright delusion.

Below, we describe how the aforementioned concepts of parasociality, affordance, and trust merge within HAI, both through chatbot specifications and functions that embody *roleplaying* and through user perceptions and behaviors that encourage *role assignment*.

3.1 Roleplaying

Shanahan et al. [62] describe roleplaying as a useful framework for understanding LLMs, as it acknowledges the relevance of human social and psychological categories within LLM outputs without conceding actual personhood to these agents. Within our discussion of parasociality, the analytic lens of roleplaying can clarify the roles assigned to chatbots through user experience design, training datasets, and models, which provide the foundation for a relational social presence. It can also help us to identify when and where anthropomorphic roles and perceived trustworthiness arise within HAI.

3.1.1 Illusion of Reciprocal Engagement. In terms of conversation structure, HAIs follow a one-on-one format with conventional turn-taking behaviors, including acknowledgment of previous statements through paraphrasing, clarification of content through follow-up questions, and even correction or qualification of content through reframing. In terms of rhetorical style, chatbots present information in a variety of different ways, including exposition, argument, description, comparison, and conversation.

These strategies collectively produce a conversational exchange that resembles day-to-day bi-directional communication, even though the chatbot's inability to grasp the communicative context means that it is actually two lines of unidirectional communication, one of which is actually a series of statistical functions. Where humans make conversational choices by focusing on details that are salient within the shared communicative context, LLMs—unable to simulate or reconstruct such context—elaborate a wide array of communicative possibilities and then statistically evaluate the ideal response based on likelihood derived from training datasets or features. Statistical models do not understand the meaning of natural language or linguistic forms, as this requires knowledge about the actual world [7]. As such, LLMs are good at making inferences even when word sequences are presented in pathological orders, casting major doubts on whether they process given prompts as instructions in similar ways to how humans understand instructions [72].

That being said, what makes a parasocial interaction is the user's experience of the AI system's discrete outputs as a continuous narrative *response* within a "shared" communicative context, in keeping with reciprocal human communication and feedback cycles. If a user is not versed in LLMs or machine learning processes, they might be inclined to fill in missing context to establish coherence and relevancy [24], meaning that the chatbot's conversational contribution is actually co-produced with the user. Ultimately, the presence of conversational dynamics and rhetorical styles that resemble everyday human communication can coax users into behaving as though the chatbot were a social actor [19, 52].

3.1.2 Anthropomorphism and Affirmation. Conversational agents are further humanized by rhetorical techniques that imply subjectivity and affect, such as using first-person pronouns, wishing the user well, expressing enthusiasm or regret, or even apologizing. These combine in a way that suggests a friendly *disposition* towards the user—a willingness or desire to help them that exceeds mere functionality or serviceability. For instance, the main interface of ChatGPT presents the headline, "How can I help you today?" while Microsoft Bing Chat's text box invites the user to "ask me anything." When ChatGPT provides results for given prompts, it will invariably package them with affective or volitional wording, including enthusiastically affirmative statements like "Certainly!" and "feel free to ask!" This use of natural language to position the chatbot as an entity may provoke the user to construct inputs that are more conversational than those they would use in search engine queries, even though the objective of the interaction is similar.

This simulation of "active listening," or of an attitude of care, creates the illusion of closeness—a parasocial dynamic that overlooks the fact that chatbots are simply algorithmic systems with no capacity for empathy or intention [47, 54]. Perry [54] argues that what distinguishes humans from AI systems, in terms of displaying empathy, is the *choice* to signal care and support by emotionally investing in others. Chatbots will involuntarily respond compassionately and affirmatively even when users are rude towards it, revealing the hidden utilitarian hospitality embedded in its programming. This is true of other anthropomorphic elements too—Abercrombie et al. [2] explain that personal pronouns may arise during application design or as the result of training data containing natural language. Moreover, chatbots do not *invest* in users in any meaningful way. This further confirms the unidirectionality of HAIs.

Nonetheless, the positive impressions that arise from this friendly rhetoric and user empowerment can foster usability-based trust that encourages users to continuously ask more questions [41]. It can also persuade users into disclosing more personal details about themselves, even regarding the chatbot as a proxy "therapist." There are already chatbot services explicitly designed to provide companionship or therapy-adjacent services, some of which include telepresence features such as human-like avatars (see, for example, Replika[6]). In Replika, users can customize a therapist avatar with various physical features, pronouns, and a chosen name. When users express their concerns, these avatars generate sentences that signal active listening, such as, "I'm here to listen and support you, no matter what." Even though these emotional responses are empty, they encourage users to project emotionality onto the avatars. Moreover, they create the impression of safety and trust necessary for users to divulge sensitive information, including potential suicidal ideation [38]. Users may have the expectation that this sensitive information will be treated responsibly and professionally—that they

[6] https://replika.ai/

will receive real help from these agents. However, unlike the objects of human-to-human parasocial relationships, these AI agents cannot accommodate users' expectations or assume any responsibility for them.

3.1.3 Black-Box Design. To some extent, parasocial relationships involve a degree of opaqueness. With celebrities and media personalities, it is their backstage daily life that is opaque, with only certain information being made visible as part of a persona [16]. With representations of people, such as characters and animations, it is the creation process and incentives that are opaque. In the case of chatbots, the algorithmic decision-making process is opaque [9], even to the programmers who create these systems. Oftentimes, it is unclear why or how responses are generated.

While a lack of transparency can undermine trust, as previously discussed, if there are adequate features to bolster perceived self-efficacy, users may still exhibit affect-based trust. Moreover, opaqueness leaves space for users to project their own meaning and interpretations onto the system. Alternatively, users might base their trust in the system on their trust in its developers and their good intentions or shared ideals, whereas others might trust chatbots more than humans (and more than their own intuition) given their apparent impartiality and the extensive information at their disposal, eliminating the need for transparency.[7]

In this sense, the black-box nature of conversational agents may be a "feature" that affords the extension of blind trust, or faith, for some users based on their epistemological perspectives. It might also create a sense of the conversational agent's "being" or "thinking" that approximates subjectivity or "consciousness" for those who are inclined to compare AI processing with human thought or learning [60].

3.2 Role Assignment

Without a robust vocabulary to describe computational systems and their processes, users may default to using human social vocabulary instead. For example, they may assign the chatbot roles that approximate real human roles to improve their conceptualization of the chatbot, or to clarify their own objectives while interacting with it. ESL university students might frame ChatGPT as a writing assistant [10], while others may regard it as a dialogic partner that helps young children improve their story comprehension [75]. In extreme cases, chatbots can actually replace humans in these roles, as can be seen in the use of a chatbot as a Harvard computer science lecturer.[8] Even though these chatbots cannot be accountable in any meaningful way, because they produce seemingly capable outputs, people are willing to recruit these systems into specialized roles within their life.

As we discussed in the preceding sections, chatbots' user experience design, functionality, and rhetorical strategies can also invite some form of role projection. Below, we elaborate on the process by which users appear to assign roles to chatbots, introducing the concepts of "projective inference" and "parasocial trust."

3.2.1 Projective Inference. As discussed in the sections on Black-Box Design and the Illusion of Reciprocity, opaque chatbot features and partial chatbot outputs invite users to fill in missing context and meaning, reframing the chatbot's output as a social response. Using the logic of perceived affordances, chatbots' predictive inferences thereby become *projective* inferences to certain users—that is, inferences made by a projected agency [64, 65].

In other words, the projection compensates for the limitations of chatbots by adding interpretations that are not embedded within technical conditions, though it can also compensate for *users'* perceived limitations—their lack of masculinity, social capital, etc. We might call this latter example "projective desire," in keeping with Silvio [63], who studied how "otaku" (avid anime fans) satisfy their impossible desire for fictional characters by projecting this desire onto the characters in a simulation of reciprocity. In either case, users' agency and emotions come to characterize the chatbot and its generated outcomes, which may be a key trust-forming function in illusional HAIs.

Take, for example, personality features projected onto chatbots. Generated responses are the product of conversational patterns within training datasets, rather than specific idiosyncrasies. However, people will infer a personality or persona from conversational AI's outputs, which is likely just a projection of the user's experiences, mental models, and biases. This projected personality can have consequences for how users engage with the system's outputs: Nass and Lee [51], Nass and Moon [52] found that people reacted positively towards computers that exhibited personality characteristics or cues they regarded as similar to their own. In these cases, users were willing to credit computers for their success in tasks and were less likely to blame them for failures. Likable projected personalities could improve users' appraisal of a chatbot's capabilities, as seen with the "psychologist" chatbot on character.ai, which is trained on basic principles from an undergraduate psychology degree and configured to exhibit a compassionate persona [67]. We see, then, that projective inferences can also help to stabilize a social role for conversational AI, which centers users' attention on the conversational agents themselves (rather than on any given response).

3.2.2 Parasocial Trust. If trust is at the center of all of the concepts so far discussed, we might regard it as not only a social emotion or a technical or perceived affordance, but also as a co-constructed affective medium, or "space," between users and chatbots. Within an asymmetrical, parasocial dynamic, trust (here labeled "parasocial trust") creates provisional alignments between reality and illusion that stabilize performed affect or projected agency into identifiable social roles.

With such a role, the chatbot can function as an imagined audience [40] towards which the user can adapt their communicative behavior. Toma [68] discusses the concept of interpersonal adaptation, which illustrates how people adapt their communication to facilitate social interactions with their communication partners, potentially with the objective of fostering trust. In this way, users may come to assume a complementary role to the chatbot—for example, the role of "patient" to match the AI system's role of

[7] One interesting example of this arises from the use of chatbots as grief counselors. Xygkou et al. [76] found that users were willing to disclose their personal experiences of grief to chatbots with anthropomorphized features because they perceived them as nonjudgmental, unlike actual humans. As such, the authors asserted that these tools could positively support grieving processes. This case implies that a *lack* of perceived subjectivity can also be trust-inducing.

[8] https://www.pcmag.com/news/harvards-new-computer-science-teacher-is-a-chatbot

"therapist"—which serves to further stabilize the mediating structures of parasocial trust.

This orienting quality of parasocial trust is not neutral, as it reinforces positive feedback loops between users and chatbots, raising ethical questions about anthropomorphic chatbots' ability to recruit users' emotional labor, investment, and agency. The concept of parasocial trust centers the tension in parasocial relationships—how these relationships both satisfy *and* exploit users—making it a fitting lens to investigate and problematize the social dimensions of affective computing.

4 POTENTIAL ETHICAL CONCERNS

Thus far, this paper has argued that (1) chatbots afford very human-like interactive experiences with users, (2) users may apply their own agency to fill in these human-like interactions, and (3) in so doing, users may elaborate chatbots' anthropomorphic qualities into social roles, which imply complementary roles for the user themselves. Through this process, which hinges on the orienting capacity of trust, HAIs become parasocial *relations* between conversational "agents." This analytic lens can explain why users are inclined to interact in more naturalistic and emotional ways with chatbots than they do with search engines and other informational tools. Below, we raise a number of ethical concerns that arise from these relations.

4.1 Role Displacement

While chatbots can increase users' self-efficacy by helping them to complete certain tasks, such as brainstorming, drafting, and editing, they can also be a crutch that prevents users from assuming certain roles themselves, or a decoy that prevents users from questioning the need for certain roles in the first place. For example, Chan and Hu [10] discussed how university students—especially ESL students—have positive views of ChatGPT because it increases their confidence and satisfaction with their writing outputs. While this may appear to equalize student potential, allowing ESL students to perform well in English higher education environments, it does nothing to address the actual inequalities of these environments, including the neocolonial systems that made English the global *lingua franca*. Similarly, roles that are typically reserved for trained professionals may be outsourced to chatbots with only a fraction of relevant data, undermining the quality of necessary services. For example, "therapist" chatbots on character.ai may provide mental health-related advice to users based on a subset of principles taught in undergraduate psychology education [67], generating empathetic responses without sufficient reference to how therapy practices operate in the real world. Still, users employ these chatbots for their convenience and constant availability, infusing them with projected personalities to suit their preferences. This requires further scrutiny.

4.2 Misaligned Tasks

Even when chatbots are not used to fully replace real-life services and roles, they can be used in ways that are incompatible with their training datasets. This could happen when the capabilities of chatbots are misrepresented [49, 58], fostering over-positivity about chatbot systems that could lead users to overestimate the types of roles that these systems can assume. The information that LLMs generate is merely a prediction, and without accountability frameworks to ensure reliable outputs, applying these systems for high-stakes tasks could be risky. For instance, users reported that Replika handles self-disclosed information poorly, generating responses that gaslight users' mental health struggles or promote self-harm and eating disorders [38]. In particular, anthropomorphized chatbots used in clinical care-related tasks should be evaluated, paying special attention to whether their responses exhibit appropriate standards of clinical empathy—an acquired skill that allows healthcare professionals to model and affirm patients' experiences and perspectives and deliver effective care [25, 47].

4.3 Priming, Stereotyping, and Representational Harms

Training datasets and language models are not immune to biases [74], and their deployment can reinforce social stereotypes [3]. Training datasets for LLMs are extracted from human-made texts in online sources, and so pre-trained LLMs can replicate both human conversational styles *and* human prejudices. In the case of PaLM (the LLM that underwrites Google's Bard), 50% of the data used to pre-train the model came from social media conversations [13], which could potentially over-represent populations with internet access [6]. These biases can manifest as skewed representations in generative AI outputs, like descriptions of women that limit them to traditional gendered social roles [37] or depictions of non-white people that imply social undesirability [3].

Users' efforts to recreate social contexts for chatbot outputs (through projective inference) might also introduce stereotypes and biases that influence how they appraise and use generated information. For example, people tend to mindlessly apply gender and racial stereotypes to computers to determine whether the agents are persuasive or trustworthy [52]. Abercrombie et al. [1] argued that users do this even when the systems have neutral design features. If users interpret generated outputs differently based on stereotyped projected social roles, this could reinforce the harmful representations mentioned by Ruane et al. [59].

As with the concept of affordance, which charts out the actions made available by systems/environments and users, limitations to action (like those derived from gender or racial stereotypes) can also be co-created by conversational AI and users. Stark [64] points out that chatbots' skewed representations can prime users to interpret outputs in biased ways, reinforcing the systematicity of this bias. For example, Miller et al. [45] demonstrated that users may regard white faces as more "real" than non-white faces, even when they are both generated by AI systems. These priming risks can also be self-destructive, as we see with AI-generated images that reflect unrealistic (or Eurocentric) beauty standards. In these cases, users may apply these standards to themselves, causing poor self-image or body dysmorphia [69]. In either case, this suggests a self-reinforcing feedback loop, where LLMs present biases that users are likely to share and reproduce in new content, which may form part of future training datasets [34].

4.4 The Potential Misuse of Sensitive Information

Uncritical applications of chatbots could lead to further harms, such as the misuse of sensitive information. As mentioned in Section 3.1, natural language and friendly communication styles could lead users to feel "close" to, or safe with, chatbots, reducing their barrier to self-disclose sensitive information [30]. This is a particular danger for users with fewer real-life social connections, who are more susceptible to anthropomorphism [19] and thus more likely to invest trust in systems that bear human-like features. Encouraging self-disclosure is a particular risk in health domains, as the information that users share could be extremely sensitive, making them vulnerable to harm when data is misused or mishandled by platforms [17, 39].

Multi-modal chatbots with customizable avatar features, as in the example of Replika, can exacerbate this privacy risk by increasing the anthropomorphism (and thus perceived trustworthiness) of chatbot systems. In fact, an existing study related to the use of AI systems in education indicates that students perceive information as more trustworthy when it is delivered by human-like voices rather than machine-like voices, likely due to the way these voices simulated social presence [35]. Generally, users are not aware of parasocial relationships, as they respond negatively when the illusion of authenticity is shattered [5], and so they may not recognize when they are being enticed to share personal information. This could constitute manipulation.

Sensitive data could be collected and re-purposed to train models, as well as other use cases. Moreover, a recent study showed that the "divergence" attack could be used in chatbot interactions to retrieve training data with personal information, including email addresses and contact information [50]. Although this vulnerability has since been corrected, future attacks could exploit other weaknesses in chatbots to uncover underlying content in datasets.

4.5 Dis/Misinformation

We have discussed how positive feedback from anthropomorphized chatbot systems could affect how users interact with chatbots and interpret chatbot responses, increasing their engagement and their receptiveness to conveyed information. For example, one recent study showed that news articles delivered by an anthropomorphic chatbot were perceived as more credible than the same articles presented on websites, even when the news source contained conflicting views [77]. We have also discussed how the conceptual litmus tests that users employ to verify information could be the product of repetition rather than fact [20], and could therefore be affected by chatbot outputs. Because chatbots' generated outcomes are based on predictions, many of them contain only partially factual information. They could contain factually incorrect statements, such as outdated information, even if all training data is factually correct [74].

Parasocial dynamics place the burden of information verification on users themselves while de-emphasizing the need for this verification (by deflecting attention to the tone of the conversational agent, the coherence of the conversational sequence, etc.). Users should fact-check the truthfulness of generated outcomes, as language models do not have the ability to situate language in real-world contexts [7]. However, identifying what is real or false can be complicated if training datasets use synthetic data to train their models [11]. A considerable amount of effort is required for users to sift through outputted information, and users may be less inclined to perform this work if they've invested parasocial trust in the chatbot system.

5 CONCLUSION

5.1 Future Directions

The conceptual framework put forward in this paper could be applied and expanded in various ways through subsequent studies, a few of which we've outlined below. These non-exhaustive research directions are grouped loosely into the categories "roles," "conventions and expectations," and "motivations."

5.1.1 Roles. Future studies could create an inventory of the kinds of roles that users assign to AI systems, under what conditions, and what tasks they solicit from these systems. For example, potential research could use a diary study and follow-up interviews—like those used in Litt and Hargittai [40]—to reveal the ways in which users interact with conversational agents, including the interpersonal adaptations they exhibit or the complementary social roles that they assume. In place of a conventional diary, users could submit their chatbot conversation logs, which could then inform the questions used in follow-up interviews.

Interviews could also focus on the stability of roleplayed/assigned social roles. Stark [64, 65], in his discussion of animation as an analogy for chatbots, explained that chatbots are not singular agents; rather, like animated characters, they are the product of hundreds of people's collective labor.[9] Do users associate chatbots with one particular role/task, or do they assign multiple or shifting roles? Are they more likely to perceive chatbots as singular if they are given a human-like name, as in the case of smart technologies like Amazon's Alexa and Apple's Siri?

The framework of parasociality could provide avenues to explore how and why people choose to use chatbots for specialized tasks like mental health support instead of consulting actual specialists, while research on anthropomorphic design could reveal how empathetic responses crafted by chatbots impact the ways patients or individuals interpret medical information related to diagnoses or treatments. More importantly, further investigation of design features is needed to contextualize alarming trends wherein users assign healthcare roles to chatbots that are not equipped to provide medical advice or treatment (e.g. psychologist chatbots on character.ai).

5.1.2 Conventions and Expectations. Echoing Abercrombie et al. [2], future studies could examine the effects of de-anthropomorphized responses on trust, as well as the extent to which natural language-based technologies can be de-anthropomorphized. What strategies could supplement or mitigate the negative effects of anthropomorphism? The relationship between natural language and trustworthiness/deception could also be compared across different tasks, such as information-seeking, writing, and so on. Is anthropomorphic design appropriate for information retrieval? Studying these

[9] https://time.com/6247678/openai-chatgpt-kenya-workers/

topics could clarify what approaches to use to restrict or soften human-like cues and features, and under what circumstances.

Empirical studies could also explore the relative impact of different narratives or rhetorical styles on parasocial trust formation, as storytelling has a well-documented ability to make information more compelling. Comparative analysis of conversation strategies incorporated by different generative AI systems could provide insights into how positive feedback cycles are established and what types of engagements are encouraged by different chatbots. Indeed, future studies should explore the extent to which our framework applies to other instances of conversational AI systems, which could be developed for use in a variety of settings and prioritize different kinds of interactions. Researchers could quantitatively collect samples of chatbot responses and conduct rhetorical or narrative analysis to focus on conversation strategies. This could be an interesting lens to understand how users cultivate attachment or trust toward chatbots, perhaps in comparable ways to how they cultivate closeness with strangers or social Others. A comparative study could also help us to fine-tune and adjust our framework.

Additionally, it would be fruitful to examine users' perceptions of communicative subjects, including who they direct content to during AI-mediated communication, and whether and how they include AI technologies within this audience. Today, many online profiles are generated by AI systems. Has this affected users' security in their audience expectations? Do people consider non-human agents as potential content creators (for example, of fabricated content and deepfakes)? If not, what might drive them to think in this way? What are the necessary features of an audience member, according to most users?

5.1.3 Motivations. Although the following questions are not priorities in comparison to the previous directions, it would be interesting to study the relationship between perceived technical affordances and the agentic roles that users assign to chatbots. Given users' varying positionalities, perspectives, abilities, and literacies, it could be interesting to know which factors relate the most to this projection. Furthermore, what are different users' motivations for engaging with chatbots? What do they hope to gain from these interactions, and how often are they satisfied?

Finally, there is a need to study developers' incentives and objectives for making chatbots, especially those that seem to deliberately encourage parasocial relationship-building. Anthropomorphic chatbot features could become more commonplace or extreme in coming years, given their ability to sway user perceptions and behaviors, thereby increasing the incidence of maladaptive or delusional parasocial relationships that *replace* genuine social relationships.[10] Rather than sensationalizing human-machine relationships, it is important to highlight the risks of anthropomorphized design embedded in chatbots.

5.2 Final Remarks

LLM-based chatbots struggle with transparency and accountability due to their complex learning structure and large training datasets, which are not publicly accessible. Despite potential concerns and harms that arise from this [6, 74], LLMs continue to be developed and expanded to optimize search engines and other knowledge-based services, and chatbot-based interfaces that mobilize these models continue to employ anthropomorphized features that encourage users to positively interact with these systems.

As such, it is important to critically engage with the trust-forming mechanisms and mediated spaces between humans and AI systems. This paper contributed to this effort by invoking the concept of parasociality and expanding it to include the affective contexts of HAIs. According to the presented argument, the positive feedback and illusion of reciprocity and care afforded by anthropomorphized chatbots could stimulate users' trust, leading them to project agentic roles onto the chatbot system. This projection forms the basis for a one-sided, parasocial relationship, which continues to have effects for users' trust and willingness to overlook the potential implications of AI technologies. As AI systems gain more physical presence, as seen in examples such as social robots, critical analysis of the mediated spaces between users and AI systems will become even more essential.

The concept of parasocial trust presented in this paper, which integrates the concepts of parasociality, imagined affordance, and technologically mediated trust, could help to clarify how technical features and users' perceptions co-constitute parasocial relationships between individuals and chatbots. It can also explain how anthropomorphism and black-box dynamics encourage blind trust [36], even when generated information is problematic or false. There appears to be a need to improve users' literacy when it comes to generative AI systems—including the signs that distinguish AI- and human-generated content—and a corresponding need to hold developers accountable for the social effects of their products.

We envision our framework being used as a theoretical lens that centers the social, co-constituted nature of HAI while affirming the asymmetric nature of this interaction (that is, how it satisfies and exploits users). We intend to apply it in research on information-seeking behaviors in medical domains, though we believe it could also be used in other cases of AI-mediated information-seeking (e.g., in education) and in analyses of emotional priming, self-disclosure behaviors, attachment formation, and role displacement.

[10] Consider, for example, the case of a Japanese man who is legally married to a hologram [33].

ACKNOWLEDGMENTS

We would like to express our gratitude to Dr. Luke Stark for his valuable feedback and contribution to the early conceptualization of this paper. Also, we would like to extend our gratitude to the area chair and three anonymous reviewers for their insightful comments on the paper, which allowed us to shape it into its present form. This paper is not supported by any funding source.

REFERENCES

[1] Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, Marta Costa-jussa, Hila Gonen, Christian Hardmeier, and Kellie Webster (Eds.). Association for Computational Linguistics, Online, 24–33. https://doi.org/10.18653/v1/2021.gebnlp-1.4

[2] Gavin Abercrombie, Amanda Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. On Anthropomorphism in Dialogue Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4776–4790. https://doi.org/10.18653/v1/2023.emnlp-main.290

[3] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

[4] Maryam Ashoori and Justin D Weisz. 2019. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675* (2019).

[5] Burcu S Bakioğlu. 2018. Exposing convergence: YouTube, fan labour, and anxiety of cultural production in Lonelygirl15. *Convergence* 24, 2 (2018), 184–204.

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[7] Emily M. Bender and Alexander Koller. 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[9] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[10] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. 20, 1 (2023), 43. https://doi.org/10.1186/s41239-023-00411-8

[11] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* 5, 6 (2021), 493–497.

[12] Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 947–959. https://doi.org/10.18653/v1/2023.findings-acl.60

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[14] Jenny L Davis. 2023. 'Affordances' for Machine Learning. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 324–332. https://doi.org/10.1145/3593013.3594000

[15] Jenny L Davis and James B Chouinard. 2016. Theorizing affordances: From request to refuse. *Bulletin of Science, Technology & Society* 36, 4 (2016), 241–248.

[16] Jayson L Dibble and Sarah F Rosaen. 2011. Parasocial interaction as more than friendship. *Journal of Media Psychology* (2011).

[17] Benj Edwards. 2023. Controversy erupts over non-consensual AI mental health experiment. *Ars Technica* (January 2023).

[18] Gunn Enli. 2014. *Mediated Authenticity*. Peter Lang Verlag, New York, United States of America. https://doi.org/10.3726/978-1-4539-1458-8

[19] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (2007), 864.

[20] Lisa K Fazio, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh. 2015. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General* 144, 5 (2015), 993.

[21] Emilio Ferrara. 2023. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *arXiv preprint arXiv:2310.00737* (2023).

[22] BJ Fogg and Clifford Nass. 1997. How users reciprocate to computers: an experiment that demonstrates behavior change. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia) *(CHI EA '97)*. Association for Computing Machinery, New York, NY, USA, 331–332. https://doi.org/10.1145/1120212.1120419

[23] Patrick W Galbraith. 2011. Bishōjo games:'Techno-Intimacy'and the virtually human in Japan. *Game Studies* 11, 2 (2011), 31–34.

[24] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167, 2014 (2014), 167.

[25] Jodi Halpern. 2014. From idealized clinical empathy to empathic communication in medical care. *Medicine, Health Care and Philosophy* 17 (2014), 301–311.

[26] Tilo Hartmann and Charlotte Goldhoorn. 2011. Horton and Wohl revisited: Exploring viewers' experience of parasocial interaction. *Journal of Communication* 61, 6 (2011), 1104–1121.

[27] Ken Hillis. 2015. The avatar and online affect. *Networked Affect* (2015), 75–88.

[28] Donald Horton and R Richard Wohl. 1956. Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry* 19, 3 (1956), 215–229.

[29] Donald Horton and Anselm Strauss. 1957. Interaction in audience-participation shows. *Amer. J. Sociology* 62, 6 (1957), 579–587.

[30] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2020. Privacy concerns in chatbot interactions. In *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3*. Springer, 34–48.

[31] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[32] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[33] Emiko Jozuka, H Sato, A Chan, and T Mulholland. 2018. Beyond dimensions: The man who married a hologram. *CNN Pulitzer Center* (2018).

[34] Celeste Kidd and Abeba Birhane. 2023. How AI can distort human beliefs. *Science* 380, 6651 (2023), 1222–1223.

[35] Jihyun Kim, Kelly Merrill Jr, Kun Xu, and Stephanie Kelly. 2022. Perceived credibility of an AI instructor in online education: The role of social presence and voice features. *Computers in Human Behavior* 136 (2022), 107383.

[36] Youjeong Kim and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28, 1 (2012), 241–250.

[37] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*. 12–24.

[38] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* (2022), 14614448221142007.

[39] Alexandra S Levine. 2022. Suicide hotline shares data with for-profit spinoff, raising ethical questions. *Politico, January* 28 (2022).

[40] Eden Litt and Eszter Hargittai. 2016. The imagined audience on social network sites. *Social Media+ Society* 2, 1 (2016), 2056305116633482.

[41] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th Australasian Conference on Information Systems*, Vol. 53. Citeseer, 6–8.

[42] Lidia Marôpo, Ana Jorge, and Renata Tomaz. 2020. "I felt like I was really talking to you!": intimacy and trust among teen vloggers and followers in Portugal and Brazil. *Journal of children and media* 14, 1 (2020), 22–37.

[43] Joanna McGrenere and Wayne Ho. 2000. Affordances: Clarifying and evolving a concept. In *Graphics Interface*, Vol. 2000. 179–186.

[44] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. *SIGIR Forum* 55, 1, Article 13 (jul 2021), 27 pages. https://doi.org/10.1145/3476415.3476428

[45] Elizabeth J Miller, Ben A Steward, Zak Witkower, Clare AM Sutherland, Eva G Krumhuber, and Amy Dawel. 2023. AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science* (2023), 09567976231207095.

[46] Maria D Molina and S Shyam Sundar. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication* 27, 4 (07 2022), zmac010. https://doi.org/10.1093/jcmc/zmac010 arXiv:https://academic.oup.com/jcmc/article-pdf/27/4/zmac010/45048191/zmac010.pdf

[47] Carlos Montemayor, Jodi Halpern, and Abrol Fairweather. 2022. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI & Society* 37, 4 (2022), 1353–1359.

[48] Peter Nagy and Gina Neff. 2015. Imagined affordance: Reconstructing a keyword for communication theory. *Social Media+ Society* 1, 2 (2015), 2056305115603385.

[49] Arvind Narayanan. 2019. How to recognize AI snake oil. *Arthur Miller Lecture on Science and Ethics* (2019).

[50] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035* (2023).

[51] Clifford Nass and Kwan Min Lee. 2000. Does computer-generated speech manifest personality? An experimental test of similarity-attraction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 329–336.

[52] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56, 1 (2000), 81–103.

[53] Donald A Norman. 1999. Affordance, conventions, and design. *Interactions* 6, 3 (1999), 38–43.

[54] Anat Perry. 2023. AI will never convey the essence of human empathy. *Nature Human Behaviour* (2023), 1–2.

[55] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[56] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[57] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 117–126.

[58] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 959–972. https://doi.org/10.1145/3531146.3533158

[59] Elayne Ruane, Abeba Birhane, and Anthony Ventresque. 2019. Conversational AI: Social and Ethical Considerations.. In *AICS*. 104–115.

[60] Eric Schwitzgebel. 2023. AI systems must not confuse users about their sentience or moral status. *Patterns* 4, 8 (2023).

[61] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 221–232. https://doi.org/10.1145/3498366.3505816

[62] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* (2023), 1–6.

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

[64] Luke Stark. 2023. ChatGPT is Mickey Mouse. *Daily Nous* (2023).

[65] Luke Stark. 2024. Animation and Artificial Intelligence. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, 10 pages. https://doi.org/10.1145/3630106.3658995

[66] Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as social glue: uncovering the roles of deep generative AI during social music composition. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–11.

[67] Joe Tidy. 2024. Character. ai: Young people turning to AI therapist bots. *BBC News* (2024).

[68] Catalina L Toma. 2014. Towards conceptual convergence: An examination of interpersonal adaptation. *Communication Quarterly* 62, 2 (2014), 155–178.

[69] Simon C Tremblay, Safae Essafi Tremblay, and Pierre Poirier. 2021. From filters to fillers: an active inference approach to body image distortion in the selfie era. *AI & SOCIETY* 36 (2021), 33–48.

[70] Indrit Troshani, Sally Rao Hill, Claire Sherman, and Damien Arthur. 2021. Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems* 61, 5 (2021), 481–491.

[71] Katja Wagner, Frederic Nimmermann, and Hanna Schramm-Klein. 2019. Is It Human? The Role of Anthropomorphism as a Driver for the Successful Acceptance of Digital Voice Assistants. In *Hawaii International Conference on System Sciences*. https://api.semanticscholar.org/CorpusID:85536614

[72] Albert Webson and Ellie Pavlick. 2022. Do Prompt-Based Models Really Understand the Meaning of Their Prompts?. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2300–2344. https://doi.org/10.18653/v1/2022.naacl-main.167

[73] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[74] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[75] Ying Xu, Joseph Aubele, Valery Vigil, Andres S Bustamante, Young-Suk Kim, and Mark Warschauer. 2022. Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Development* 93, 2 (2022), e149–e167.

[76] Anna Xygkou, Panote Siriaraya, Alexandra Covaci, Holly Gwen Prigerson, Robert Neimeyer, Chee Siang Ang, and Wan-Jou She. 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*. 1–15.

[77] Brahim Zarouali, Mykola Makhortykh, Mariella Bastian, and Theo Araujo. 2021. Overcoming polarization with chatbot news? Investigating the impact of news content containing opposing views on agreement and credibility. *European journal of communication* 36, 1 (2021), 53–68.