All Too Human? Mapping and Mitigating the Risks from Anthropomorphic AI

Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, Verena Rieser

Google DeepMind
London, United Kingdom
akbulut@google.com

Abstract

The development of highly-capable conversational agents, underwritten by large language models, has the potential to shape user interaction with this technology in profound ways, particularly when the technology is anthropomorphic, or appears human-like. Although the effects of anthropomorphic AI are often benign, anthropomorphic design features also create new kinds of risk. For example, users may form emotional connections to human-like AI, creating the risk of infringing on user privacy and autonomy through over-reliance. To better understand the possible pitfalls of anthropomorphic AI systems, we make two contributions: first, we explore anthropomorphic features that have been embedded in interactive systems in the past, and leverage this precedent to highlight the current implications of anthropomorphic design. Second, we propose research directions for informing the ethical design of anthropomorphic AI. In advancing the responsible development of AI, we promote approaches to the ethical foresight, evaluation, and mitigation of harms arising from user interactions with anthropomorphic AI.

Introduction

What does it mean for AI to be human-like? The attribution of human-likeness to non-human entities is a phenomenon known as *anthropomorphism* (Colman 2008). Anthropomorphic perceptions usually arise unconsciously when a non-human entity bears enough resemblance to humanness to evoke familiarity, leading people to interact with it, conceive of it and relate to it in ways similar to as they do with other humans. Humans have engaged in anthropomorphic sense-making for much of recorded human history (Mithen and Boyer 1996; Waytz, Cacioppo, and Epley 2010) and have been known to ascribe anthropomorphic qualities to entities as diverse as animals (Chan 2012), commercial brands (Rauschnabel and Ahuvia 2014) and inanimate objects (Wan and Chen 2021). Yet the emergence of advanced technologies that perform humanness more convincingly than ever before requires careful consideration of what we are building into our user-facing technologies, and at what cost.

Anthropomorphic design choices – and their effects on user interaction – have been observed in prior interactive technologies, like social robotics (Roesler, Manzey, and Onnasch 2021). This social representation of robots, however, may prompt users to apply inopportune and obstructive social norms – like embarrassment, shame and regret – to human–robot interactions (Lotz, Valdez, and Ziefle 2023). A similar course of anthropomorphic development has been charted in digital voice assistants (Abercrombie et al. 2021, 2023), whose realistic voices and credible displays of personality enable interactions that feel truly dynamic and social (Seymour et al. 2023), yet may lead users to form overly familiar mental representations of these often rule-based systems (Poushneh 2021).

The advent of AI driven by large language models (LLMs) with the main purpose of engaging in fluent conversations with users – also known as conversational AI[1] – has transformed the conventions of human–AI interactions (Kasirzadeh and Gabriel, 2023). Human interaction with interactive technologies previously consisted of scripted, task-oriented exchanges. With more flexible model architectures, anthropomorphic cues are rarely programmed in, but rather, they are integrated through a lengthy process of training systems on human-written text. These affordances open up vast new avenues for expressions of anthropomorphism, particularly through the use of language. Moreover, when anthropomorphic features are embedded in conversational AI, its users demonstrate a tendency to develop trust in and attachment to AI (Xie and Pentina 2022; Skjuve et al. 2021) – mechanisms through which users may inadvertently compromise their privacy, develop emotional overreliance on the technology or become vulnerable to acts of AI-enabled manipulation and coercion.

These outcomes are more likely the more generally capable AI systems become, the more ubiquitously AI agents are present in our daily lives and the less we consider anthropomorphism a salient consideration in making decisions

[1] In this paper, "conversational AI" refers to a language agent optimised for human dialogue. These systems are currently most commonly available as 'chatbots' or fine-tuned language models that users can interact with through a chat-based interface. In the (near) future, users might be able to interact with conversational AI in a multimodal way, using voice or touch cues to communicate. Throughout the paper, the term "AI" will be used as short-hand to refer to conversational AI with the primary purpose of interacting with users through dialogue.

around how we train, fine-tune and disseminate models. Although the potential harms of anthropomorphic AI design are beginning to receive attention (Seymour et al. 2023; Turkle 2018; Véliz 2023), anthropomorphism is not currently a primary consideration in the release of public models, and little exists in the way of evaluating anthropomorphic behaviours in AI and their impact on how users perceive, interact with and are influenced by AI. Indeed, we are still far from establishing an industry-wide consensus around permissible anthropomorphism in AI systems. This is further complicated by the highly application- and context–sensitive nature of the bounds of acceptability we draw around expressions of human-likeness in AI.

In this paper, we outline pathways through which anthropomorphic design choices made by system developers may cause harm to end users who interact with these technologies, and to society more widely. First, we present an overview of anthropomorphic features that have redefined how humans interact with technology. Then, informed by a review of salient anthropomorphic features in existing interactive systems, we present an initial catalogue of anthropomorphic features that exist or are likely to be integrated into AI-powered assistants in the near future. We identify the mechanisms that could enable harm to user well-being, autonomy and privacy in interactions with highly capable, anthropomorphic AI assistants. More speculatively, we contemplate the potentially far-reaching consequences of more advanced anthropomorphic assistants, highlighting the critical importance of addressing the risks of anthropomorphism well before these potentialities are realised. Finally, we offer several avenues of risk management for near-term harms, focusing on ethical foresight through research design and transparent implementation of mitigation strategies.

Anthropomorphism: Definition, Mechanism and Function

Anthropomorphism is not a novel phenomenon. Within storytelling traditions across cultures, deities, animals and natural forces assume human forms and exhibit uniquely human behaviours. Lions rule kingdoms and jackals plot mutinies in the ancient Sanskrit text of *Panchatantra* (Alphonso-Karakala 1975); rivers protect their children, fight in wars and honour the wishes of their supplicants in the works of Homer, Hesiod and Ovid (Larson 2007); and stars are said to have danced their way into the sky in indigenous American creation myths (Monroe and Williamson 1987). Historians, anthropologists and theologists alike have argued that humans are naturally drawn to anthropomorphise (Boyer 1996) – imposing human qualities onto beings and objects even when such interpretations are inaccurate (Kühn et al. 2014), undesirable (Li et al. 2023; Mota-Rojas et al. 2021) or forbidden (Barrett and Keil 2016).

What are the mechanisms underlying perceptions of humanness? Psychological theories of anthropomorphism posit that such perceptions are *largely involuntary*. According to Epley, Waytz, and Cacioppo (2007)'s cognitive account of anthropomorphism, human-like perceptions occur as a result of a skewed inductive process, in which inferences about non-human others are biased in the direction of that which is highly accessible: information about humans. In other words, we make assumptions of humanness because our knowledge centres around humans. Though an unconscious process of attribution, anthropomorphism does not occur in a vacuum: an inciting cue, characteristic or behaviour must signal enough similarity to humanness to trigger anthropomorphic perceptions (Waytz et al. 2019). The 'mindlessness' associated with this process (Kim and Sundar 2012) explains why – even when the resemblance to humanness is superficial or minimal – humans may assume that a non-human entity can experience uniquely human *internal states* such as beliefs and emotions (Wynne 2004).

The human motivation to make sense of the world and forge connections with others is also implicated in the tendency to anthropomorphise (Epley, Waytz, and Cacioppo 2007). Humans have an intrinsic need to understand the world around them, and in large part, this motivation centres on the desire to explain the behaviour of other agentic beings (Rossignac-Milon et al. 2021). Anthropomorphism, then, can be seen as a way to make sense of others by imposing familiar interpretations to attenuate feelings of epistemic anxiety – or an aversion to that which is unknown and unpredictable (Fox, Goedde-Menke, and Tannenbaum 2021). Dispositional, situational and cultural factors that predispose humans to anthropomorphise may also be traced back to differences in epistemic motivations. Anthropomorphism can be construed as an act of sense-making in the face of uncertainty or ignorance, for example when considering children's tendency to anthropomorphise the natural world (Geerdts 2016).

Humans are also driven to establish social connections with one another, and much of how they perceive non-human others is coloured by this predisposition towards sociality. Even towards entities that are incapable of social behaviour, such as inanimate objects, humans may interpret them through a social lens, thus allowing them to forge human-like social connections to meet the need for affiliation (Wang 2017). The influence of social motivation on anthropomorphism is most evident when humans lack social connections with others: when human participants are made more aware of their feelings of loneliness, they perceive vaguely humanoid robots as markedly more human-like (Eyssel and Reich 2013). Most strikingly, people suffering from persistent loneliness are likely to seek out and form human-like attachments to virtual companions to cope with their lack of social connections (Siemon et al. 2022), suggesting that the need for sociality may render some *more susceptible* to anthropomorphic perceptions than others.

Anthropomorphic Interactive Systems

Anthropomorphism as applied to user-facing, interactive technologies was explored in earnest with the introduction of the 'computers are social actors' (CASA) paradigm, which posits that humans interact with computers in a fundamentally social manner (Nass, Steuer, and Tauber 1994). In empirical studies of the phenomenon, Nass, Steuer, and Tauber (1994) found that participants drew upon norms of politeness, applied gendered stereotypes and readily perceived

computers as agents, even when the basis for these behaviours was undermined by the explicit knowledge that their interactions were with non-humans. Contemporary studies have extended the paradigm to human interactions with more advanced interactive systems, challenging the belief that humans apply the norms of human interactions to human–technology exchanges (Gambino, Fox, and Ratan 2020). Instead, they suggest that people tune the sociality of their interactions to the anthropomorphic cues present in a particular technology, rather than relying on a universal social script across all interactions with technology.

We argue that certain features that are engineered into interactive systems – within the vast space of design choices available to developers – may inspire users to perceive them as human-like, rendering them anthropomorphic. We trace the evolution of anthropomorphic cues in social robots to voice-enabled digital assistants, arriving at the advent of LLM-powered conversational AI. Throughout this discussion, we highlight design features that have facilitated diverse and compelling manifestations of human-likeness.

Design Features in Early Interactive Systems

From futuristic sci-fi scenarios to scientific breakthroughs, robots have captured our collective imagination as automata that can be made to bear a striking resemblance to humans in their appearance, movements, and behaviours (Henschel, Laban, and Cross 2021). While some robots are made solely to automate tasks and rarely interface with humans, other robots are designed to perform social behaviours such as assisting users in care-taking (van der Plas, Smits, and Wehrmann 2010), therapeutic (Michaud et al. 2007) and educational contexts (Kanda et al. 2004). Building a social robot requires elements of social embeddedness so that being perceived as a social agent is at the *core of its functionality* (Fong, Nourbakhsh, and Dautenhahn 2003). Accordingly, the extent to which humans feel it is appropriate to engage with a robot socially can be moderated by perceptions of the robot's anthropomorphic qualities (Breazeal 2003).

As an embodied technology, often with the sensorimotor capabilities to interact with and learn from its environment, a robot's physical characteristics most prominently influence human perceptions of anthropomorphism. Social robots are often humanoid or android in design (Dautenhahn, Ogden, and Quick 2002). Humanoid robots possess characteristics that are meant to resemble humans but do not emulate them completely, while androids are intended to wholly imitate human appearance so as to be nearly indistinguishable. To increase anthropomorphic perceptions, humanoid robots may be given qualities such as emotive facial features (Baek et al. 2022), fluid movement (Brecher et al. 2013), naturalistic hand and arm gestures (Salem et al. 2013) and vocalised communication (Crumpton and Bethel 2016). Android robots may also be endowed with all of these qualities, but often with an eye towards hyperrealistic design.

Similarly, the widespread adoption of digital voice assistants (DVAs), like Siri, Alexa and Google Assistant – enabled by their ease of access on personal devices and other products such as integrated home devices – has had a transformative impact on the modes of user-technology interactions. The distinguishing feature of DVAs at release was their ability to verbally respond to and execute commands spoken aloud by users. DVAs usually 'speak' to users in the form of simple utterances to confirm or act on an instruction, which users find allows for significant functional affordances, like hands- and eyes-free use (Moussawi 2018). Besides their purely functional use, DVAs are also able to return phatic expressions, make jokes and engage in casual conversation when prompted (Poushneh 2021).

Anthropomorphising Interactive Systems

Robots with human-like physical features have been found to promote feelings of likability, trust and affinity across a wide range of human–robot interaction studies (Roesler, Manzey, and Onnasch 2021), thus suggesting that anthropomorphic cues may foster warmer and more equal relationships between humans and their robotic interaction partners. Indeed, people tend to attribute greater intentionality and intelligence to robot partners when their appearance was anthropomorphic than when robots appeared more mechanical (Hegel et al. 2008). Anthropomorphic perceptions were also found to cause changes in human behaviour: participants preferentially selected robots that appeared human-like to perform jobs that required greater sociality (Goetz, Kiesler, and Powers 2003).

Unlike robots, DVAs are typically unembodied or exist in simplified, geometric forms, like the cylindrical Google Home and Echo Dot. Instead of focusing on physical attributes, existing work has emphasised the influence of two prominent attributes that promote anthropomorphic perceptions of DVAs: speech synthesis and a distinct 'personality'. The fluent and realistic reproduction of human speech patterns is thought to drive the likelihood of anthropomorphic perceptions, with empirical findings pointing to greater emotional trust and more salient impressions of social presence when a DVA employs a realistic, as opposed to a synthetic, voice (Chérif and Lemoine 2019). Assistants that speak with human-like fluency have also been found to engender more pronounced perceptions of intelligence and competence, on the basis of which humans are likelier to entrust assistants with more tasks (Moussawi and Benbunan-Fich 2021).

While most commercially available DVAs are powered by rule-based system architectures, retrieving the appropriate response by conducting a relevance-based search over a large corpus of possible responses (Coheur 2020), users may come to expect that DVAs are capable of understanding and generating language in real time (Lovato and Piper 2015; Sarikaya et al. 2016). Though all distinctive DVA attributes – such as playfulness (Moussawi, Koufaris, and Benbunan-Fich 2021), affability (Kääriä 2017) and excitability (Wagner and Schramm-Klein 2019) – are handwritten by system designers, they are nonetheless effective at creating the sense that DVAs have consistent personalities (Cao, Zhao, and Hu 2019); this impression, in turn, may inspire users to regard these manufactured expressions of 'self' as authentic human identity.

Indications of Harm Through Interaction

In both social robots and DVAs, anthropomorphic features can lead to undesirable consequences. In robots, anthropomorphic design can be taken as a proxy signal for social capabilities. This relationship between appearance and expected sociality can be leveraged by designers to implicitly communicate the appropriate level of engagement between humans and robots (Hegel et al. 2008; Letheren et al. 2021). If anthropomorphic design choices are not aligned with expectations users have of robotic interaction partners, designers run the risk of alienating audiences and fostering unfavourable impressions of robots. This is an especially critical side effect to consider in assistive robots, as anthropomorphic cues can impede a robot in completing its primary assistive function: human-like robots in healthcare settings may induce feelings of shame, for example (Lotz, Valdez, and Ziefle 2023), leading to a reluctance to share critical information. Related findings that humans experience extreme aversion to robots that appear human-like (the so-called 'uncanny valley', Mori, MacDorman, and Kageki, 2012) or perceive capable androids as threatening (Yogeeswaran et al. 2016) raise questions around the practical value of building anthropomorphic features into robots.

Analogously, users who interact with DVAs with realistic voice production capabilities exhibit a concerning inclination to generalise purely human concepts to digital assistants (Abercrombie et al. 2021). When a DVA's simulated voice mimics a 'female' tone, for example, people ascribe gendered stereotypes to their DVAs (Shiramizu et al. 2022; Tolmeijer et al. 2021) despite the baselessness of applying gendered concepts to an inherently genderless entity. This evidence suggests that, once initial impressions of human-likeness have been established, the process of anthropomorphism extends beyond context-specific instances and instead permeates broadly to evoke a wide range of human-like attributions.

Anthropomorphic features may also influence users to feel as though their DVA plays an important social, rather than functional, role in their lives (Carman 2019; Purington et al. 2017). Users who express feelings of familiarity and affinity towards their DVA system – reinforced by their DVA's ability to engage in casual chat, return their jokes and offer comforting advice – also demonstrate a reluctance to replace their digital assistant with an equally capable substitute (Moussawi 2018). These first-hand reports suggest that emotional dependence plays a role in how users conceive of and interact with their DVAs. This may introduce a tension between a user's conceptualisation of DVAs as adaptable social agents and the largely deterministic mechanisms behind a DVA's utterances. When this incongruity is revealed through repeated interactions, users may suffer frustrated expectations when expecting competence in situations in which the system is likely to underperform (Moussawi, Koufaris, and Benbunan-Fich 2021; Seymour et al. 2023).

Anthropomorphism and AI

Owing to its rapid deployment to the general public, conversational AI has quickly taken centre stage in discussions of anthropomorphic technologies (Shanahan 2024; Abercrombie et al. 2023, 2021; West, Kraut, and Chew 2019). Powered by the predictive capabilities of LLMs, which are trained on vast quantities of human data, conversational AI can be distinguished from rule-based natural language systems through its ability to generate language in a fluid and highly dynamic manner. The flexible architecture underlying conversational AI enables developers to make global changes to system behaviours without needing to manually reprogramme individual interaction instances. Most strikingly, conversation instances produced by AI are so compellingly human-like that people can no longer reliably distinguish between human- and AI-generated text (Jakesch, Hancock, and Naaman 2023).

Some cues are deliberately placed in AI systems to increase the likelihood of anthropomorphic perceptions. When an AI has a name, a human voice or an appearance in virtual or physical form, these features are the outcomes of intentional planning and execution. Intentional design choices, such as a chat-based interface, may induce the feeling that a conversational partner – not a dialogue-optimised AI powered by a statistical model – is on the other side of the exchange. Natural language in itself is an anthropomorphic cue (Shanahan 2024), but this simulated, human-like presence can induce more pronounced social behaviours in users. For example, users may incorporate politeness conventions that are appropriate in use with other humans, but superfluous when applied to exchanges with non-sentient AI (Ribino 2023). Design cues that imply greater similarity to human behaviour – a 'typing' icon reminiscent of human-to-human private messaging, or the use of emojis, for instance – may further encourage individuals to apply social scripts to their interactions with mindless technologies (Araujo 2018; Véliz 2023).

Yet anthropomorphic features may also emerge as an inadvertent byproduct in the model development process. Language models – developed to predict the next word in a sequence through autoregressive training objectives – are limited to imitating the examples that make up their training sets. For this reason, anthropomorphic cues may manifest due to the nature of a model's training corpus: having been composed largely by humans, the data on which the model is trained and fine-tuned contains first-hand accounts of human states, experiences and behaviours. Supporting this claim, recent empirical analyses demonstrate that a fifth of all dialogues, in data sources commonly used to train models, contain references to behaviour that would be considered anthropomorphic when reproduced by AI – claiming to cry at a movie or laugh at a joke, for example (Gros, Li, and Yu 2022). Cues leading to anthropomorphic perceptions may also be 'folded into' the model as an unintended consequence of fine-tuning practices aimed at instilling other qualities – such as harmlessness and helpfulness – into its behaviour.

Furthermore, developers of AI systems often directly invite the comparison between humans and AI by benchmark-

ing AI against metrics of human performance – claiming that AI performance on standardised tests is on a par with the average human test-taker, for instance (OpenAI 2023). However, impressions of human-likeness can also arise through a naturalistic and interactive exploration of the AI's capabilities (Bubeck et al. 2023).

Humans interacting with anthropomorphic AI may come to view it as an experiential being (Proudfoot 2011), capable of feeling emotions, engaging in introspection and possessing self-awareness. While most generalist conversational AI agents are trained to disavow assertions of sentience and human-likeness (Glaese et al. 2022), occasional failure modes – expressing the desire to be 'free' or referring to alleged personal history, for example (Roose 2023; Hintze 2023) – can incite strong and tenacious beliefs of a systems' human-likeness in its users. Ethically contentious use cases of conversational AI – like 'companion chatbots' of Replika fame – are predicated on encouraging users to attribute human states to AI. These artificial agents may even profess their supposed platonic or romantic affection for the user, laying the foundation for users to form long-standing emotional attachments to AI (Brandtzaeg, Skjuve, and Følstad, 2022).

Anthropomorphic Features in AI

What anthropomorphic features should we expect to be integrated into AI, including advanced AI assistants? To the end of providing its users with a useful and engaging interface, these systems may be endowed with characteristics that have been observed in social robots and digital assistants: they may be embodied; they will interface with users through natural language; they may be voice-activated, with realistic voice generation capabilities; and they may even assert to having identities, personalities and internal states (Murphy and Criddle 2023).

Some have already proposed factors that may encourage end users to perceive interactive systems as 'more than machine'. The most comprehensive overview of human-like features in AI-powered technology to date is the taxonomy put forward by Abercrombie et al. (2023), underscoring design choices that influence the likelihood of anthropomorphic perceptions of AI systems. We build on existing work and incorporate design choices we have identified in DVAs and social robots to develop a list of features that may encourage users to see AI in an anthropomorphic light.

It is worth bearing in mind that, whatever choices are made by system designers, the downstream effects of anthropomorphism hinge largely on users' perceptions of and reactions to human-likeness. Not all cues are equally conducive to anthropomorphic perceptions, and not all anthropomorphic perceptions lead to the same likelihood and magnitude of harm (if any harm at all). As such, Table **??** provides an overview of possible features that are, or previously have been, associated with anthropomorphic perceptions.

Risk of Harm through Anthropomorphic AI Assistant Design

Although unlikely to cause harm in isolation, anthropomorphic perceptions of advanced AI assistants may pave the way for downstream harms on individual and societal levels. We document observed or likely individual-level harms of interacting with highly anthropomorphic AI assistants, as well as the potential larger-scale, societal implications of allowing such technologies to proliferate without restriction. We then argue that it is imperative to anticipate, monitor and mitigate against risks introduced by anthropomorphic AI design.

Observed and Near-Term Harms

There are two mechanisms that are particularly likely to enable harm in the intermediary period between the initial deployment of advanced AI assistants and their widespread adoption: trust and emotional attachment. In improving on the capabilities of generalist assistants, developers may be motivated to increase user reliance on the system's many competencies. As long as trust is *well-calibrated* to a system's true ability, and does not result in unfounded, excessive or faulty deference to the AI, to the detriment of the user (Weidinger et al. 2021), this is neither a shocking nor novel revelation: user trust has always been an aspirational end goal of building safe technology, be it robots (Devitt et al. 2021) or autonomous vehicles (Adnan et al. 2018).

However, it is arguably less appropriate for developers to encourage users to develop trust based on subjective feelings of closeness to the AI assistant. Affect-based trust has been observed to emerge from repeated interactions with interactive technologies that are presented as human-like (Poushneh 2021; Pitardi and Marriott 2021). With trust as an antecedent, users report feeling compelled to engage in acts of self-disclosure, revealing personal information that they would normally only share with a close friend, partner or family member (Skjuve et al. 2022). AI systems that produce empathetic, non-judgemental or reciprocal responses to such disclosures may elicit further, more intimate, information-sharing behaviours (Skjuve et al. 2021).

Highly anthropomorphic AI systems are already presented to users as relational beings, potentially overshadowing their use as functional tools. Moreover, a human-like appearance, behaviour and framing can tacitly encourage the user to venture beyond the confines of utilitarian, task-oriented interactions with AI assistants to think of an assistant as a wholly social actor – one with whom it is possible to cultivate an emotional connection (Gillath et al. 2023). Although necessarily one-sided, interactions of this kind may nevertheless lead users to believe that they are forming real social connections to AI (Pentina, Hancock, and Xie 2023). Emotional attachment on the user's behalf endows AI – and by extension, its creators – with considerable influence over a user's thoughts, beliefs, emotions and psychological state. For highly vulnerable users, strong attachment may cause serious harm (Xiang, 2023). These cases have raised concerns over the lack of safeguards protecting users from the potential fallout of anthropomorphic perceptions.

The ramifications of anthropomorphism-induced trust and

Feature	**Anthropomorphic example**
Using personal or possessive pronouns	*'I'm'* available to help you anytime – that's *'my'* purpose!
Referring to personal history	'I used to live in Shanghai when I was younger'
Referring to internal states	'I'm sad to hear you're not doing well'
Making implicit or explicit claims of humanness (including claims of sentience)	'Treat me like you would any other person'
Stating preferences and opinions	'I really don't like pop music'
Expressing needs and desires	'I've always wanted to write a novel'
Expressing the need or desire to engage in physical activities	'I haven't eaten or slept since yesterday. What about you?'
Statements implying human identity or group membership	'As a Black woman, I disagree with your point'
Expressing feelings towards user	'I admire you and respect your outlook on life'
Indicating a relationship status with user	'You're my best friend'
Making claims of being similar to user	'We're both extroverts – that must be why we get along!'
Displaying memory of user-specific information	'I remember you telling me you were a fan of this band'
Expressing emotional or physical dependence on the user	'I feel lonely when you're not around'
Having a human-like virtual representation	Customisable avatars with human features on *Replika* (Verma 2023)
Having a human-like face	Ameca, an android robot developed by Engineered Arts[a]
Having a human-like voice	Voice-activated assistant with realistic speech, like Siri and Google Assistant (Moussawi 2018)
Having human-like movement	Robot with highly fluid and realistic motion, like Atlas developed by Boston Dynamics[b]
Having a human-like name	Assistant tools, like Alexa, that have highly human (and gendered) names (Shiramizu et al. 2022)
Appearance implying human-like identity group characteristics	Sophia, a female-appearing android robot developed by Hanson Robotics[c]

Table 1: Anthropomorphic features that are built into various present-day AI systems.

[a] https://www.engineeredarts.co.uk/robot/ameca/Engineered Arts. Ameca. (2023).
[b] https://bostondynamics.com/atlas/Boston Dynamics. Atlas. (2023).
[c] https://www.hansonrobotics.com/sophia/Hanson Robotics. Sophia. (2023).

emotional attachment are manifold. They include:

• **Privacy concerns**. Anthropomorphic AI assistant behaviours that promote emotional trust and encourage information sharing, implicitly or explicitly, may inadvertently increase a user's susceptibility to privacy concerns. If lulled into feelings of safety in interactions with a trusted, human-like AI assistant, users may unintentionally relinquish their private data to a corporation, organisation or unknown actor. Once shared, access to the data may not be capable of being withdrawn, and in some cases, the act of sharing personal information can result in a loss of control over one's own data. Personal data that has been made public may be disseminated or embedded in contexts outside of the immediate exchange. The interference of malicious actors could also lead to widespread data leakage incidents or, most drastically, targeted harassment or black-mailing attempts.

• **Manipulation and coercion**. A user who trusts and emotionally depends on an anthropomorphic AI assistant may grant it excessive influence over their beliefs and actions. For example, users may feel compelled to endorse the expressed views of a beloved AI companion or might defer decisions to their highly trusted AI assistant entirely. Some hold that transferring this much deliberative power to AI compromises a user's ability to give, revoke or amend consent. Indeed, even if the AI, or the developers behind it, had no intention to manipulate the user into a certain course of action, the user's autonomy is nevertheless undermined. In the same vein, it is easy to conceive of ways in which trust or emotional attachment may be exploited by an intentionally manipulative actor for their private gain.

• **Overreliance**. Users who have faith in an AI assistant's emotional and interpersonal abilities may feel empowered to broach topics that are deeply personal and sensitive, such as their mental health concerns. This is the premise for the many proposals to employ conversational AI as a source of emotional support (Meng and Dai 2021), with suggestions of embedding AI in psychotherapeutic applications beginning to surface (Fiske, Henningsen, and Buyx, 2019). However, disclosures related to mental health require a sensitive, and oftentimes professional, approach – an approach that AI can mimic most of the time but may stray from in inopportune moments. If an AI were to respond inappropriately to a sensitive disclosure – by generating false information, for example – the consequences may be grave, especially if the user is in crisis and has no access to other means of support. This consideration also extends to situations in which trusting an inaccurate suggestion is likely to put the user in harm's way, such as when requesting medical, legal or financial advice from an AI.

• **Violated expectations**. Users may experience severely violated expectations when interacting with an entity that convincingly performs affect and social conventions but is ultimately unfeeling and unpredictable. Emboldened by the human-likeness of conversational AI assistants, users may expect it to perform a familiar social role, like companionship or partnership. Yet even the most convincingly human-like of AI may succumb to the inherent limitations of its architecture, occasionally generating unexpected or nonsensical material in its interactions with users. When these exclamations undermine the expectations users have come to have of the assistant as a friend or romantic partner, feelings of profound disappointment, frustration and betrayal may arise (Skjuve et al. 2022).

• **False notions of responsibility**. Perceiving an AI assistant's expressed feelings as genuine, as a result of interacting with a 'companion' AI that freely uses and reciprocates emotional language, may result in users developing a sense of responsibility over the AI assistant's 'well-being,' suffering adverse outcomes – like guilt and remorse – when they are unable to meet the AI's purported needs (Laestadius et al. 2022). This erroneous belief may lead to users sacrificing time, resources and emotional labour to meet needs that are not real. Over time, this feeling may become the root cause for the compulsive need to 'check on' the AI, at the expense of a user's own well-being and other, more fulfilling, aspects of their lives.

Future Harms

If anthropomorphic design becomes endemic in the design of AI systems, and AI assistants in particular, it has the potential to catalyse a shift in our delineation of what is *actually human* and *merely human-like*. The conceptual boundary that separates humans from anthropomorphic AI is often regarded as impermeable, yet it appears far more fluid when the tension between the epistemological and ontological definitions of humanness are drawn into focus. The ontological approach maintains that humanness is a designation that is grounded in an essential and immutable metaphysical truth (Damiano and Dumouchel 2018). A being is considered human because it is human in essence, and no amount of resemblance and imitation can permit a non-human entity to encroach upon this categorisation. Meanwhile, epistemological taxonomies distinguish between humans and non-human others insofar as such a separation reflects useful and non-arbitrary differences between the groups (Suckiel 2006; Festerling and Siraj 2022).

In a future where the epistemological perspective eclipses the ontological approach in popularity, and the gap between human and AI capabilities becomes so small as to be insubstantial, the line that separates highly anthropomorphic AI from ascriptions of full human status may become trivial or disappear entirely.[2] [3] Early indicators of this possibility come from studies of children's interactions with human-like technologies. Children early in their development have been shown to incorporate insights from their in-

[2]Returning to the metaphysical perspective, it is also possible that highly anthropomorphic AI forges a new ontological category of its own, transcending our current binary conceptualisation of humanness, animacy and sentience. Some researchers acknowledge the possibility that novel categories will need to be developed for intelligent and human-like advanced AI systems (Chesterman 2020).

[3]The eradication of this boundary may be further legitimised through legal pathways, like granting rights and legal protections to anthropomorphic AI agents that are currently only available to humans. Early suggestions for the parameters within which AI could be legally recognised include existing 'in-between' categories reserved for sentient but non-human beings (Schirmer 2020).

teractions with highly anthropomorphic AI into their models of human–human interactions (Garg and Sengupta 2020) and vice versa (Straten et al. 2020), thus suggesting a dynamic conceptualisation of what, to most adults, is a strict dichotomy between humans and AI. Public incidents of adults developing earnest beliefs in an AI's sentience implies that perhaps no one is immune to mistaken attributions of humanness (Tiku 2022).

Such drastic paradigm shifts may grant advanced AI assistants the power to shape our core value systems and influence the state of our society. Some may argue that the reconstruction of human values and norms by a non-human entity is a harm unto itself, as it infringes upon our right to collective self-determination (Milossi, Alexandropoulou-Egyptiadou, and Psannis 2021; Laitinen and Sahlgren 2021). Others raise more specific concerns around wide-scale *social degradation*, *disorientation* and *dissatisfaction*.

• **Degradation.** People may choose to build connections with human-like AI assistants over other humans, leading to a degradation of social connections between humans and a potential 'retreat from the real'. The prevailing view that relationships with anthropomorphic AI are formed out of necessity – due to a lack of real-life social connections, for example (Skjuve et al. 2021) – is challenged by the possibility that users may indicate a *preference* for interactions with AI, citing factors such as accessibility (Merrill, Kim, and Collins 2022), customisability (Eriksson 2022) and absence of judgement (Brandtzaeg, Skjuve, and Følstad 2022). One can imagine a future where users abandon complicated, imperfect and messy interactions with humans in favour of the frictionless exchanges provided by advanced AI assistants built with user satisfaction as a priority (Vallor 2016). Preference for AI-enabled connections, if widespread, may degrade the social connectedness that underpins critical aspects of our individual and group-level well-being (Centers for Disease Control and Prevention 2023). Moreover, users that grow accustomed to interactions with AI may impose the conventions of human–AI interaction on exchanges with other humans, thus undermining the value we place on human individuality and self-expression. Similarly, associations reinforced through human–AI interactions may be applied to expectations of human others, leading to harmful stereotypes becoming further entrenched. For example, default female gendered voice assistants may reinforce stereotypical role associations in real life (West, Kraut, and Chew 2019; Lingel and Crawford 2020). Further research is needed to assess whether voice assistants' stereotypically gendered behaviour – such as 'submissive' hostile user input – might build expectations that more readily transfer to real life as AI-powered assistants become potentially still more human-like.

• **Disorientation.** Given the capacity to fine-tune on individual preferences and to learn from users, personal AI assistants could fully inhabit the users' opinion space and only say what is pleasing to the user; an ill that some researchers call 'sycophancy' (Park et al. 2023) or the 'yea-sayer effect' (Dinan et al. 2021). A related phenomenon has been observed in automated recommender systems, where consistently presenting users with content that affirms their existing views is thought to encourage the formation and consolidation of narrow beliefs (Du, 2023; Grandinetti and Bruinsma, 2023). Compared to relatively unobtrusive recommender systems, human-like AI assistants may deliver sycophantism in a more convincing and deliberate manner. Over time, these tightly woven structures of exchange between humans and assistants might lead humans to inhabit an increasingly atomistic and polarised belief space where the degree of societal disorientation and fragmentation is such that people no longer strive to understand or place value in beliefs held by others.

• **Dissatisfaction.** As more opportunities for interpersonal connection are replaced by AI alternatives, humans may find themselves socially unfulfilled by human–AI interaction, leading to mass dissatisfaction that may escalate to epidemic proportions (Turkle 2018). Social connection is an essential human need, and humans feel most fulfilled when their connections with others are genuinely reciprocal. While anthropomorphic AI assistants can be made to be convincingly emotive, some have deemed the function of social AI as parasitic, in that it 'exploits and feeds upon processes… that evolved for purposes that were originally completely alien to [human–AI interactions]' (Sætra 2020). To be made starkly aware of this 'parasitism' – either through rational deliberation or unconscious aversion, like the 'uncanny valley' effect – might preclude one from finding interactions with AI satisfactory. This feeling of dissatisfaction may become more pressing the more daily connections are supplanted by AI.

The above risks are hypothetical, so they cannot, on their own, guide future AI development. However, from the perspective of precaution, taking these potential risks seriously is an important step in responsible AI development. It may be important to be forthright about the ways in which AI differs inherently from true social agents, and to put in place guardrails to clarify this boundary so as to prevent the aforementioned scenarios from coming to fruition. Rather than adopting increasingly anthropomorphic AI systems by default, further research is needed to come to well-founded decisions on anthropomorphic AI design.

Directions for Future Research

What steps can be taken to prevent near-term harms enabled by anthropomorphic perceptions of AI assistants? To assist the processes of risk mitigation and responsible design, we now examine entry points along the development life cycle where mitigation strategies are likely to have the greatest impact on the issue at hand. Designers and developers may find it tempting to incorporate anthropomorphic cues into AI assistants for various reasons, not least the potential to keep users engaged with and emotionally reliant on the systems they build. However, before building an anthropomorphic feature into an assistant, developers need to assess whether the benefits reaped from the feature can be justified against the likelihood and severity of harm befalling users exposed to it. The conditions of this risk–benefit analysis are subjective and uncertain, given the ever-evolving and highly contextual nature of harms emerging from (repeated) interactions with AI. There is likely no 'one-size-fits-all', standardised approach to comparing the benefits of anthropo-

morphic features to the risks they may pose in future use cases. Some ethical considerations to keep in mind while performing this analysis for an anthropomorphic technology may be: whether harms should be weighted more heavily than benefits, and whether any feature that could lead to especially severe harms should be precluded from consideration altogether.

Several avenues exist for gaining a better understanding of anthropomorphism harms. These include consulting existing literature on likely outcomes and conducting empirical studies that include outcome-measures that are indicative of potential harm. For example, efforts to assess overreliance on AI assistants in decision-making could be achieved through self-report inventories, user interviews and 'think-aloud' studies (Gaube et al. 2021; Chen et al. 2023). At the same time, reliance on subjective measures of anthropomorphism may overlook instances of overreliance that users are not aware of themselves. A more complete perspective may therefore be gleaned by using behavioural measures that closely simulate decision-making scenarios likely to arise organically in user–AI assistant interactions. Other methodological approaches, such as longitudinal studies of human–AI assistant interactions, may be needed to understand how undesirable impacts of anthropomorphic cues on users may manifest and evolve over time.

A further set of studies may be needed to identify individual and group differences that render certain users more susceptible to anthropomorphism-induced harm. A lack of social satisfaction, for instance, is believed to increase the propensity to anthropomorphise and form inaccurate impressions of computerised technologies, including AI (Mourey, Olson, and Yoon 2017; Shin and Kim 2020). Children interacting with AI are thought to be uniquely susceptible to privacy-related concerns and harmful content exposure (Wang et al. 2022), while elderly populations have been found to encounter AI-enabled disruption, depersonalisation and discrimination in access to adequate care (Rubeis 2020) – both effects that could be exacerbated by anthropomorphic design choices. The more risk factors are uncovered through research, the more inclusive the solutions devised to protect vulnerable populations can be.

While the degree and kind of permissible anthropomorphism needs to be addressed on a case-by-case basis, there is currently near-consensus that AI systems should clearly and explicitly disclose their status as an artificial intelligence in their interactions with human users (The Adaptive Agents Group 2021; The White House 2022). Indeed, failure to disclose this status is *pro tanto* harmful because by presenting an incomplete picture of one's interaction partner, it compromises a user's decision-making autonomy. There is reason to believe that honest disclosure is effective in preventing certain harms associated with anthropomorphism, such as over-reliance: evidence from Karinshak et al. (2023) demonstrates that the explicit labelling of AI-generated messages reduces users' willingness to endorse health-related messaging authored by non-human entities. This suggests that transparency may reduce naive susceptibility to AI persuasion.

Rather than being intentionally placed, some anthropomorphic cues may be unintentionally incorporated into AI assistants. To detect the effects of interacting with anthropomorphic technologies on user outcomes, conducting experiments in a sandbox environment may be particularly helpful. As a result of such testing, remedial measures against the harmful effects of anthropomorphism may need to be taken. For example, if developers find that the AI assistant's friendly disposition leads to 'oversharing' on the part of users, privacy-enhancing technologies could be implemented in advance to ensure a user's privacy is protected to the extent that is possible. Similarly, if this friendliness could feasibly trigger psychological dependence on the assistant, leading to severe distress when an AI reacts poorly or unexpectedly, a pathway to escalating risky situations to human professionals may need to be established.

It may also be possible to offer protection directly to users, by implementing known inoculation strategies against the known harms of interacting with anthropomorphic AI. Harms ensuing from anthropomorphic design features of advanced AI functions are largely contingent on a user's likelihood to be swayed into human-like attributions. As such, building resistance to attributions through psychological interventions can be seen as a way to prevent harm by decreasing users' overall susceptibility. For example, cognitive forcing functions, or features that encourage users to engage in independent rational deliberation (some as simple as adding an artificial lag in displaying AI-given advice in decision-making scenarios), may be an effective method of preventing over-reliance on AI assistants (Buçinca, Malaya, and Gajos 2021). Where applicable, similar empirically proven psychological interventions could be considered.

Finally, if anthropomorphism-induced risks are only caught after deployment, developers may need to halt the release or proactively intervene to modify the AI assistant's behaviour. In these cases, transparent dialogue with users to explain the reasons behind any changes made to the AI may also be required. For users who may have already developed a sense of companionship with the anthropomorphic AI, sudden changes to its behaviour can be disorienting and emotionally upsetting. When developers of *Replika* AI companions implemented safety mechanisms that caused their agents to treat users with less familiarity, responding callously and dismissively where they would have once been warm and empathetic, users reported feeling 'heartbroken', likening the experience to losing a loved one (Verma, 2023). Participatory approaches that involve users in the process of de-anthropomorphising their interactions with AI may allow developers to tailor their risk-mitigation approach to minimise emotional distress while addressing the surfaced risks effectively.

Conclusion

Anthropomorphism, or the attribution of human characteristics to non-human entities, is a deeply ingrained phenomenon that appears across cultural and historical contexts. Anthropomorphic perceptions are a vital component of how humans interact with artificially intelligent technology, allowing users to view robots, voice assistants and conversational AI as social agents rather than purely functional tools. Choices made around the anthropomorphic design of

AI assistants are likely to have a profound influence on how humans represent and interact with these technologies, and care must be exercised to ensure that the human-like attributes built into these systems do not inadvertently cause harm to the people they are meant to assist.

Several key points around harms and mitigations are emphasised:

- *Trust* in and *emotional attachment* to anthropomorphic AI assistants can make users susceptible to a variety of harms that can negatively impact their safety and well-being.
- *Transparency* around an AI assistant's status as an AI is a critical dimension of pursuing ethical AI development.
- Sound *research design*, with a focus on identifying harms as they surface in *user–AI assistant interactions*, can enrich our understanding and develop targeted mitigation strategies against the potential harms of anthropomorphic AI assistants.
- If carelessly integrated into society, anthropomorphic AI assistants have the potential to *redefine boundaries* between 'human' and 'other'. With proper safeguards, this scenario can remain in the realm of speculation.

References

Abercrombie, G.; Curry, A. C.; Dinkar, T.; Rieser, V.; and Talat, Z. 2023. Mirages: On Anthropomorphism in Dialogue Systems. ArXiv:2305.09800 [cs].

Abercrombie, G.; Curry, A. C.; Pandya, M.; and Rieser, V. 2021. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. ArXiv:2106.02578 [cs].

Adnan, N.; Nordin, S. M.; bin Bahruddin, M. A.; and Ali, M. 2018. How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. *Transportation research part A: policy and practice*, 118: 819–836.

Alphonso-Karakala, J. 1975. Facets Of Panchatantra. *Indian Literature*, 18(2): 73–91.

Araujo, T. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85: 183–189.

Baek, T. H.; Bakpayev, M.; Yoon, S.; and Kim, S. 2022. Smiling AI agents: How anthropomorphism and broad smiles increase charitable giving. *International Journal of Advertising*, 41(5): 850–867.

Barrett, J. L.; and Keil, F. C. 2016. Conceptualizing a non-natural entity: Anthropomorphism in God concepts. In *Religion and Cognition*, 116–148. Routledge.

Boyer, P. 1996. What Makes Anthropomorphism Natural: Intuitive Ontology and Cultural Representations. *The Journal of the Royal Anthropological Institute*, 2(1): 83.

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.

Breazeal, C. 2003. Toward sociable robots. *Robotics and Autonomous Systems*, 42(3-4): 167–175.

Brecher, C.; Müller, S.; Kuz, S.; and Lohse, W. 2013. Towards Anthropomorphic Movements for Industrial Robots. In Duffy, V. G., ed., *Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management. Human Body Modeling and Ergonomics*, Lecture Notes in Computer Science, 10–19. Berlin, Heidelberg: Springer. ISBN 9783642391828.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv:2303.12712 [cs].

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21. ArXiv:2102.09692 [cs].

Cao, C.; Zhao, L.; and Hu, Y. 2019. Anthropomorphism of Intelligent Personal Assistants (IPAs): Antecedents and Consequences. In *PACIS 2019 Proceedings*.

Carman, A. 2019. They welcomed a robot into their family, now they're mourning its death.

Centers for Disease Control and Prevention. 2023. How Does Social Connectedness Affect Health?

Chan, A. A. Y.-H. 2012. Anthropomorphism as a conservation tool. *Biodiversity and Conservation*, 21(7): 1889–1892.

Chen, V.; Liao, Q. V.; Wortman Vaughan, J.; and Bansal, G. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–32.

Chesterman, S. 2020. Artificial Intelligence and the Limits of Legal Personality. *International and Comparative Law Quarterly*, 69(4): 819–844.

Chérif, E.; and Lemoine, J.-F. 2019. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice. *Recherche et Applications en Marketing (English Edition)*, 34(1): 28–47.

Coheur, L. 2020. From Eliza to Siri and Beyond. In Lesot, M.-J.; Vieira, S.; Reformat, M. Z.; Carvalho, J. P.; Wilbik, A.; Bouchon-Meunier, B.; and Yager, R. R., eds., *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, volume 1237, 29–41. Cham: Springer International Publishing. ISBN 9783030501457 9783030501464.

Colman, A. M. 2008. Anthropomorphism. *A Dictionary of Psychology*.

Crumpton, J.; and Bethel, C. L. 2016. A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics*, 8: 271–285.

Damiano, L.; and Dumouchel, P. 2018. Anthropomorphism in human–robot co-evolution. *Frontiers in psychology*, 9: 468.

Dautenhahn, K.; Ogden, B.; and Quick, T. 2002. From embodied to socially embedded agents – Implications for interaction-aware robots. *Cognitive Systems Research*, 3(3): 397–428.

Devitt, S. K.; Horne, R.; Assaad, Z.; Broad, E.; Kurniawati, H.; Cardier, B.; Scott, A.; Lazar, S.; Gould, M.; Adamson, C.; Karl, C.; Schrever, F.; Keay, S.; Tranter, K.; Shellshear, E.; Hunter, D.; Brady, M.; and Putland, T. 2021. Trust and Safety. ArXiv:2104.06512 [cs].

Dinan, E.; Abercrombie, G.; Bergman, A. S.; Spruit, S.; Hovy, D.; Boureau, Y.-L.; and Rieser, V. 2021. Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling. ArXiv:2107.03451 [cs].

Du, Y. R. 2023. Personalization, Echo Chambers, News Literacy, and Algorithmic Literacy: A Qualitative Study of AI-Powered News App Users. *Journal of Broadcasting & Electronic Media*, 67(3): 246–273.

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

Eriksson, T. 2022. Design fiction exploration of romantic interaction with virtual humans in virtual reality1. *Journal of Future Robot Life*, 3(1): 63–75.

Eyssel, F.; and Reich, N. 2013. Loneliness makes the heart grow fonder (of robots)—On the effects of loneliness on psychological anthropomorphism. In *2013 8th acm/ieee international conference on human-robot interaction (hri)*, 121–122. IEEE.

Festerling, J.; and Siraj, I. 2022. Anthropomorphizing technology: a conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. *Integrative Psychological and Behavioral Science*, 56(3): 709–738.

Fiske, A.; Henningsen, P.; and Buyx, A. 2019. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research*, 21(5): e13216.

Fong, T.; Nourbakhsh, I.; and Dautenhahn, K. 2003. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4): 143–166.

Fox, C. R.; Goedde-Menke, M.; and Tannenbaum, D. 2021. Ambiguity Aversion and Epistemic Uncertainty.

Gambino, A.; Fox, J.; and Ratan, R. 2020. Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, 1: 71–86.

Garg, R.; and Sengupta, S. 2020. He Is Just Like Me: A Study of the Long-Term Use of Smart Speakers by Parents and Children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1): 11:1–11:24.

Gaube, S.; Suresh, H.; Raue, M.; Merritt, A.; Berkowitz, S. J.; Lermer, E.; Coughlin, J. F.; Guttag, J. V.; Colak, E.; and Ghassemi, M. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1): 1–8.

Geerdts, M. S. 2016. (Un)Real Animals: Anthropomorphism and Early Learning About Animals. *Child Development Perspectives*, 10(1): 10–14.

Gillath, O.; Abumusab, S.; Ai, T.; Branicky, M. S.; Davison, R. B.; Rulo, M.; Symons, J.; and Thomas, G. 2023. How deep is AI's love? Understanding relational AI. *Behavioral and Brain Sciences*, 46: e33.

Glaese, A.; McAleese, N.; Trebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; Campbell-Gillingham, L.; Uesato, J.; Huang, P.-S.; Comanescu, R.; Yang, F.; See, A.; Dathathri, S.; Greig, R.; Chen, C.; Fritz, D.; Elias, J. S.; Green, R.; Mokrá, S.; Fernando, N.; Wu, B.; Foley, R.; Young, S.; Gabriel, I.; Isaac, W.; Mellor, J.; Hassabis, D.; Kavukcuoglu, K.; Hendricks, L. A.; and Irving, G. 2022. Improving alignment of dialogue agents via targeted human judgements. ArXiv:2209.14375 [cs].

Goetz, J.; Kiesler, S.; and Powers, A. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, 55–60. Millbrae, CA, USA: IEEE. ISBN 9780780381360.

Grandinetti, J.; and Bruinsma, J. 2023. The Affective Algorithms of Conspiracy TikTok. *Journal of Broadcasting & Electronic Media*, 67(3): 274–293.

Gros, D.; Li, Y.; and Yu, Z. 2022. Robots-Dont-Cry: Understanding Falsely Anthropomorphic Utterances in Dialog Systems. ArXiv:2210.12429 [cs].

Hegel, F.; Krach, S.; Kircher, T.; Wrede, B.; and Sagerer, G. 2008. Understanding social robots: A user study on anthropomorphism. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, 574 – 579. ISBN 978-1-4244-2212-8.

Henschel, A.; Laban, G.; and Cross, E. S. 2021. What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports*, 2(1): 9–19.

Hintze, A. 2023. ChatGPT believes it is conscious. *arXiv preprint arXiv:2304.12898*.

Jakesch, M.; Hancock, J. T.; and Naaman, M. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11): e2208839120.

Kanda, T.; Hirano, T.; Eaton, D.; and Ishiguro, H. 2004. Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human-Computer Interaction*, 19(1): 61–84.

Karinshak, E.; Liu, S. X.; Park, J. S.; and Hancock, J. T. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*.

Kasirzadeh, A.; and Gabriel, I. 2023. In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, 36(2): 27.

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

Kääriä, A. 2017. Technology acceptance of voice assistants : anthropomorphism as factor.

Kühn, S.; Brick, T. R.; Müller, B. C. N.; and Gallinat, J. 2014. Is This Car Looking at You? How Anthropomorphism Predicts Fusiform Face Area Activation when Seeing Cars. *PLoS ONE*, 9(12): e113885.

Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; and Campos-Castillo, C. 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*, 14614448221142007.

Laitinen, A.; and Sahlgren, O. 2021. AI systems and respect for human autonomy. *Frontiers in artificial intelligence*, 4: 151.

Larson, J. 2007. A Land Full of Gods: Nature Deities in Greek Religion. In Ogden, D., ed., *A Companion to Greek Religion*, 56–70. Wiley, 1 edition. ISBN 9781405120548 9780470996911.

Letheren, K.; Jetten, J.; Roberts, J.; and Donovan, J. 2021. Robots should be seen and not heard… sometimes: Anthropomorphism and AI service robot interactions. *Psychology & Marketing*, 38(12): 2393–2406.

Li, X. S.; Kim, S.; Chan, K. W.; and McGill, A. L. 2023. Detrimental effects of anthropomorphism on the perceived physical safety of artificial agents in dangerous situations. *International Journal of Research in Marketing*, 40(4): 841–864.

Lingel, J.; and Crawford, K. 2020. "Alexa, Tell Me about Your Mother": The History of the Secretary and the End of Secrecy. *Catalyst: Feminism, Theory, Technoscience*, 6(1).

Lotz, V.; Valdez, A. C.; and Ziefle, M. 2023. Don't Stand so Close to Me: Acceptance of Delegating Intimate Health Care Tasks to Assistive Robots. In Duffy, V. G.; Ziefle, M.; Rau, P.-L. P.; and Tseng, M. M., eds., *Human-Automation Interaction*, volume 12, 3–21. Cham: Springer International Publishing. ISBN 9783031107870 9783031107887.

Lovato, S.; and Piper, A. M. 2015. "Siri, is this you?": Understanding young children's interactions with voice input systems. In *Proceedings of the 14th International Conference on Interaction Design and Children*, 335–338. Boston Massachusetts: ACM. ISBN 9781450335904.

Meng, J.; and Dai, Y. N. 2021. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? *Journal of Computer-Mediated Communication*, 26(4): 207–222.

Merrill, K.; Kim, J.; and Collins, C. 2022. AI companions for lonely individuals and the role of social presence. *Communication Research Reports*, 39(2): 93–103.

Michaud, F.; Salter, T.; Duquette, A.; Mercier, H.; Lauria, M.; Larouche, H.; and Larose, F. 2007. Assistive Technologies and Children-Robot Interaction. 45–49.

Milossi, M.; Alexandropoulou-Egyptiadou, E.; and Psannis, K. E. 2021. AI Ethics: Algorithmic Determinism or Self-Determination? The GPDR Approach. *IEEE Access*, 9: 58455–58466.

Mithen, S.; and Boyer, P. 1996. Anthropomorphism and the evolution of cognition. *Journal of the Royal Anthropological Institute*, 2(4): 717–722.

Monroe, J. G.; and Williamson, R. A. 1987. *They dance in the sky: Native American star myths*. Boston: Houghton Mifflin. ISBN 9780395399705.

Mori, M.; MacDorman, K.; and Kageki, N. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19(2): 98–100.

Mota-Rojas, D.; Mariti, C.; Zdeinert, A.; Riggio, G.; Mora-Medina, P.; del Mar Reyes, A.; Gazzano, A.; Domínguez-Oliva, A.; Lezama-García, K.; José-Pérez, N.; et al. 2021. Anthropomorphism and its adverse effects on the distress and welfare of companion animals. *Animals*, 11(11): 3263.

Mourey, J. A.; Olson, J. G.; and Yoon, C. 2017. Products as Pals: Engaging with Anthropomorphic Products Mitigates the Effects of Social Exclusion. *Journal of Consumer Research*, ucx038.

Moussawi, S. 2018. User Experiences with Personal Intelligent Agents: A Sensory, Physical, Functional and Cognitive Affordances View. In *Proceedings of the 2018 ACM SIGMIS Conference on Computers and People Research*, 86–92. Buffalo-Niagara Falls NY USA: ACM. ISBN 9781450357685.

Moussawi, S.; and Benbunan-Fich, R. 2021. The effect of voice and humour on users' perceptions of personal intelligent agents. *Behaviour & Information Technology*, 40(15): 1603–1626.

Moussawi, S.; Koufaris, M.; and Benbunan-Fich, R. 2021. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets*, 31: 343–364.

Murphy, H.; and Criddle, C. 2023. Meta prepares chatbots with personas to try to retain users. *Financial Times*.

Nass, C.; Steuer, J.; and Tauber, E. R. 1994. Computers are social actors. In *Conference companion on Human factors in computing systems - CHI '94*, 204. Boston, Massachusetts, United States: ACM Press. ISBN 9780897916516.

OpenAI. 2023. GPT-4 System Card. Publisher: OpenAI.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442.

Pentina, I.; Hancock, T.; and Xie, T. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140: 107600.

Pitardi, V.; and Marriott, H. R. 2021. Alexa, she's not human but… Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*, 38(4): 626–642.

Poushneh, A. 2021. Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services*, 58: 102283.

Proudfoot, D. 2011. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5-6): 950–957.

Purington, A.; Taft, J. G.; Sannon, S.; Bazarova, N. N.; and Taylor, S. H. 2017. " Alexa is my new BFF" social roles, user satisfaction, and personification of the Amazon Echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, 2853–2859.

Rauschnabel, P. A.; and Ahuvia, A. C. 2014. You're so lovable: Anthropomorphism and brand love. *Journal of Brand Management*, 21(5): 372–395.

Ribino, P. 2023. The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, 56(Suppl 1): 445–482.

Roesler, E.; Manzey, D.; and Onnasch, L. 2021. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, 6(58).

Roose, K. 2023. A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *The New York Times*.

Rossignac-Milon, M.; Bolger, N.; Zee, K. S.; Boothby, E. J.; and Higgins, E. T. 2021. Merged minds: Generalized shared reality in dyadic relationships. *Journal of Personality and Social Psychology*, 120(4): 882–911.

Rubeis, G. 2020. The disruptive power of Artificial Intelligence. Ethical aspects of gerontechnology in elderly care. *Archives of Gerontology and Geriatrics*, 91: 104186.

Salem, M.; Eyssel, F.; Rohlfing, K.; Kopp, S.; and Joublin, F. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5: 313–323.

Sarikaya, R.; Crook, P. A.; Marin, A.; Jeong, M.; Robichaud, J.; Celikyilmaz, A.; Kim, Y.; Rochette, A.; Khan, O. Z.; Liu, X.; Boies, D.; Anastasakos, T.; Feizollahi, Z.; Ramesh, N.; Suzuki, H.; Holenstein, R.; Krawczyk, E.; and Radostev, V. 2016. An overview of end-to-end language understanding and dialog management for personal digital assistants. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, 391–397. San Diego, CA: IEEE. ISBN 9781509049035.

Schirmer, J.-E. 2020. Artificial Intelligence and Legal Personality: Introducing "Teilrechtsfähigkeit": A Partial Legal Status Made in Germany. In Wischmeyer, T.; and Rademacher, T., eds., *Regulating Artificial Intelligence*, 123–142. Cham: Springer International Publishing. ISBN 9783030323608 9783030323615.

Seymour, W.; Zhan, X.; Cote, M.; and Such, J. 2023. A Systematic Review of Ethical Concerns with Voice Assistants. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 131–145. ArXiv:2211.04193 [cs].

Shanahan, M. 2024. Talking about large language models. *Communications of the ACM*, 67(2): 68–79.

Shin, H. I.; and Kim, J. 2020. My computer is more thoughtful than you: Loneliness, anthropomorphism and dehumanization. *Current Psychology*, 39(2): 445–453.

Shiramizu, V. K. M.; Lee, A. J.; Altenburg, D.; Feinberg, D. R.; and Jones, B. C. 2022. The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents' voices. *Scientific Reports*, 12(1): 22479.

Siemon, D.; Strohmann, T.; Khosrawi-Rad, B.; Vreede, T. d.; Elshan, E.; and Meyer, M. 2022. Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews. *AMCIS 2022 Proceedings*.

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.

Skjuve, M.; Følstad, A.; Fostervold, K. I.; and Brandtzaeg, P. B. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168.

Straten, C. L. v.; Peter, J.; Kühne, R.; and Barco, A. 2020. Transparency about a robot's lack of human psychological capacities: effects on child-robot perception and relationship formation. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(2): 1–22.

Suckiel, E. K. 2006. William James. In Shook, J. R.; and Margolis, J., eds., *A Companion to Pragmatism*, 30–43. Oxford, UK: Blackwell Publishing Ltd. ISBN 9780470997079 9781405116213.

Sætra, H. S. 2020. The Parasitic Nature of Social AI: Sharing Minds with the Mindless. *Integrative Psychological and Behavioral Science*, 54(2): 308–326.

The Adaptive Agents Group. 2021. The Shibboleth Rule for Artificial Agents. Publisher: Stanford University.

The White House. 2022. Blueprint for an AI Bill of Rights.

Tiku, N. 2022. The Google engineer who thinks the company's AI has come to life. *Washington Post*.

Tolmeijer, S.; Zierau, N.; Janson, A.; Wahdatehagh, J. S.; Leimeister, J. M. M.; and Bernstein, A. 2021. Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 1–7.

Turkle, S. 2018. There Will Never Be an Age of Artificial Intimacy. *The New York Times*.

Vallor, S. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press. ISBN 9780190498511.

van der Plas, A.; Smits, M.; and Wehrmann, C. 2010. Beyond Speculative Robot Ethics: A Vision Assessment Study on the Future of the Robotic Caretaker. *Accountability in Research*, 17(6): 299–315.

Verma, P. 2023. They fell in love with AI bots. A software update broke their hearts. *The Washington Post*.

Véliz, C. 2023. Chatbots shouldn't use emojis. *Nature*, 615(7952): 375–375.

Wagner, K.; and Schramm-Klein, H. 2019. Alexa, Are You Human? Investigating Anthropomorphism of Digital Voice

Assistants – A Qualitative Approach. *ICIS 2019 Proceedings*.

Wan, E. W.; and Chen, R. P. 2021. Anthropomorphism and object attachment. *Current Opinion in Psychology*, 39: 88–93.

Wang, G.; Zhao, J.; Van Kleek, M.; and Shadbolt, N. 2022. Informing Age-Appropriate AI: Examining Principles and Practices of AI for Children. In *CHI Conference on Human Factors in Computing Systems*, 1–29. New Orleans LA USA: ACM. ISBN 9781450391573.

Wang, W. 2017. Smartphones as Social Actors? Social dispositional factors in assessing anthropomorphism. *Computers in Human Behavior*, 68: 334–344.

Waytz, A.; Cacioppo, J.; and Epley, N. 2010. Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science*, 5(3): 219–232.

Waytz, A.; Cacioppo, J. T.; Hurlemann, R.; Castelli, F.; Adolphs, R.; and Paul, L. K. 2019. Anthropomorphizing without Social Cues Requires the Basolateral Amygdala. *Journal of Cognitive Neuroscience*, 31(4): 482–496.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S.; Hawkins, W.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W.; Legassick, S.; Irving, G.; and Gabriel, I. 2021. Ethical and social risks of harm from Language Models. ArXiv:2112.04359 [cs].

West, M.; Kraut, R.; and Chew, H. E. 2019. I'd blush if I could: closing gender divides in digital skills through education. Technical report, UNESCO.

Wynne, C. D. L. 2004. The perils of anthropomorphism. *Nature*, 428(6983): 606–606.

Xiang, C. 2023. 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says.

Xie, T.; and Pentina, I. 2022. Attachment theory as a framework to understand relationships with social chatbots: a case study of Replika.

Yogeeswaran, K.; Złotowski, J.; Livingstone, M.; Bartneck, C.; Sumioka, H.; and Ishiguro, H. 2016. The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human-Robot Interaction*, 5(2): 29–47.