PNAS PERSPECTIVE OPEN ACCESS

The benefits and dangers of anthropomorphic conversational agents

Sandra Peter[a,1,2], Kai Riemer[a,1,2], and Jevin D. West[b,1,2]

Edited by Richard Aslin, Haskins Laboratories, Inc., New Haven, CT; received September 10, 2024; accepted April 3, 2025

A growing body of research suggests that the recent generation of large language model (LLMs) excel, and in many cases outpace humans, at writing persuasively and empathetically, at inferring user traits from text, and at mimicking human-like conversation believably and effectively—without possessing any true empathy or social understanding. We refer to these systems as "anthropomorphic conversational agents" to aptly conceptualize the ability of LLM-based systems to mimic human communication so convincingly that they become increasingly indistinguishable from human interlocutors. This ability challenges the many efforts that caution against "anthropomorphizing" LLMs, attaching human-like qualities to nonhuman entities. When the systems themselves exhibit human-like qualities, calls to resist anthropomorphism will increasingly fall flat. While the AI industry directs much effort into improving the reasoning abilities of LLMs—with mixed results—the progress in communicative abilities remains underappreciated. In this perspective, we aim to raise awareness for both the benefits and dangers of anthropomorphic agents. We ask: should we lean into the human-like abilities, or should we aim to dehumanize LLM-based systems, given concerns over anthropomorphic seduction? When users cannot tell the difference between human interlocutors and AI systems, threats emerge of deception, manipulation, and disinformation at scale. We suggest that we must engage with anthropomorphic agents across design and development, deployment and use, and regulation and policy-making. We outline in detail implications and associated research questions.

anthropomorphic conversational agents | anthropomorphic agents | large language models | generative AI

It has long been a goal in computer science to make computers more accessible and more natural to interact with, by making computing more human-like (1, 2). Work on synthetic voice interfaces (3, 4) and photorealistic digital human faces (5) have made great strides in recent years. In this perspective, we show that it is recent advances in large language models (LLMs), the technology at the heart of the recent wave of commercial AI systems, that achieves this goal convincingly. We argue that in the quest for creating artificial, human-like intelligence, we might instead have overshot in creating human-like communicative abilities, in ways that we are yet to fully appreciate and prepare for.

A heated debate is ongoing in the AI community over whether or not LLMs are beginning to show signs of human-like intelligence, with some arguing that we are seeing the emergence of human-like reasoning abilities (6–8), while others are showing that this might be an artifact of LLM training, and training data (9–12). We argue that we might have overlooked a different achievement, that LLMs have become as good as, and in many ways better than typical humans in communicative abilities.

While a growing number of studies have shown that LLMs exhibit an inherent brittleness (13) and unreliability (14, 15) in knowledge tasks, and that they struggle with real-world understanding (11, 16) and human-like reasoning tasks (9, 10, 14), others have shown that LLMs instead excel in mimicking human language and communicative abilities (17), that they exhibit human-like responses in personality tests (18, 19), and that they consistently pass the Turing test (20) as a result (14, 21, 22).

We present a growing body of emerging research that demonstrates that the latest generation LLMs have encoded language at such nuanced levels as to mimic

Author affiliations: [a]University of Sydney Business School, The University of Sydney, Sydney, NSW 2006, Australia; and [b]Center for an Informed Public, Information School, University of Washington, Seattle, WA 98195

Author contributions: S.P., K.R., and J.D.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]S.P., K.R., and J.D.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: sandra.peter@sydney.edu.au, kai.riemer@sydney.edu.au, or jevinw@uw.edu.

Published May 16, 2025.

highly convincingly a range of communicative abilities that have proven in tests to be equivalent to, and often exceed, above average human ability levels. This includes the ability 1) to write persuasively (23), empathetically (24) and even deceptively (25), 2) to infer and match user emotion and intent (26), and 3) to engage in highly believable interaction, e.g., via role play (27).

To grasp this development conceptually, and to raise awareness of these abilities, we argue that we must move beyond notions of anthropomorphism, as the attribution of humanity to a nonhuman entity by an individual, to a conception that locates anthropomorphic qualities directly on the machine side of the relationship. We thus use the term "anthropomorphic conversational agents," or "anthropomorphic agents" for short, to highlight the highly nuanced language mastery of LLMs that imbues these systems with their anthropomorphic abilities.

We do not posit that LLMs have developed such traits in a genuine human sense, in line with what others have noted (16, 28, 29). But users increasingly cannot tell the difference between human writing and LLM writing when presented with text (30–32), and between a human interlocutor or a contemporary chatbot when in interaction (17, 26). Some studies even suggest that users believe LLMs have memories, feelings, or consciousness (33). For all intents and purposes, this makes LLMs "anthropomorphic" in their ability to mimic human writing and human communication highly persuasively.

In response to this insight, we posit the central question orienting this perspective: should we continue designing AI systems for anthropomorphic abilities, or should we instead work to dehumanize them? We suggest that anthropomorphic abilities are both AI's greatest potential but also its greatest danger. Anthropomorphic abilities come with the potential to create new conversational user interfaces and new ways to interact with complex information in natural and accessible ways. At the same time, they present the challenge of "anthropomorphic seduction," the allure of convincing human-like interaction in the absence of any true human traits, like understanding or empathy. This brings with it the potential for deception and manipulation at scale and for mistaken beliefs by users that these machines understand human experience and existence in ways they cannot (16, 33).

We discuss benefits and dangers, and derive implications and new research questions for 1) the design and development of anthropomorphic agents, 2) their deployment and use, and for 3) regulation and policy-making. With this perspective we hope to help reorient efforts in the field to focus on these emerging anthropomorphic abilities, for appreciating the latest generation of AI systems as mimicking machines with highly advanced communicative abilities, rather than systems that embody genuine human understanding or intelligence.

While anthropomorphic agents might additionally be imbued with human-like synthetic voices and photorealistic faces to mimic human likeness, we will show that their foundational, anthropomorphic abilities stem from recent advances in LLMs. While synthetic voices or photorealistic faces will no doubt add to the anthropomorphic qualities of AI systems, any human-likeness will be severely curtailed in the absence of such communicative abilities. On the other hand, emerging research shows that LLM-generated text and text-based chat are often enough to mimic humanness highly convincingly (23–27).

LLMs

LLMs, a type of foundation model (34), are based on transformers, a subset of deep neural networks with billions of parameters, trained on a massive corpus of textual data. Training generally involves a form of self-supervision at scale, whereby hidden parts of an input sentence are predicted (11) to successively encode increasingly complex and abstract features of the textual training data into a high-dimensional, numerical latent space (35, 36).

Technically, this means that any words or text features and their tokens become characterized as "nearness relationships" with other tokens in the form of numerical, high-dimensional, vectors. GPT-3, for example, uses word vectors with 12,288 dimensions where each word, simply speaking, becomes represented by a list of 12,288 numbers (37). There appears to be an unwavering belief that, as the size of these models increases with subsequent generations, in terms of both the number of parameters and the volume of training data, ever more sophisticated abilities will continue to emerge, as "scale is all you need." (38)

Indeed, recent incarnations of LLMs, such as OpenAI's GPT4, Anthropic's Claude or Google's Gemini, have demonstrated astonishing abilities to generate from these high-dimensional representations human-like text responses (8), engage in conversations (39, 40), and, in some cases, exhibit what appears like human reasoning abilities (7, 12, 41). Generally, the development of these models, and their optimization and fine-tuning for downstream tasks (42) has been guided by the quest to build toward artificial human-like, even general, intelligence (43).

As such, much of the focus of recent developments has been on improving the reasoning abilities of LLMs, their general functionality across a broad range of knowledge and reasoning tasks. This is evidenced for example by the wide range of tests that have been derived to assess them. For example, LogicBench tests for "logic reasoning" (44), MM-InstructEval for zero-shot "reasoning" (45), KoLa for "knowledge and world understanding" (46), and D-NLP for medical "inference capabilities" (47).

In addition, LLMs are being optimized for communicative abilities, with techniques such as reinforcement learning from human feedback (RLHF) (48) being employed for "improving the user experience of the GPT model by designing interfaces that are both intuitive and user-friendly" (49), e.g., for the model to respond like a chatbot, or dialog agent mimicking a human interlocutor (27), and to give answers that are practical and safe (48).

In the following, we argue that the focus on reasoning abilities might have obscured the astonishing progress of LLMs in communicative abilities. As the AI industry aims to imbue these probabilistic systems with deterministic veracity to reign in hallucinations (50), and the ability to engage in logic or mathematical reasoning (51), we might have paid less attention to the fact that LLMs have already mastered a range of communicative abilities at exceptional levels that outperform above-average humans in tasks typically associated with uniquely human traits.

Anthropomorphic Conversational Agents

In this perspective, we argue that the progress in communicative abilities evident in the latest generation of LLMs amounts to the arrival of what we term anthropomorphic conversational agents, or anthropomorphic agents for short, systems that are able to convincingly mimic human communication, often exceeding typical human ability levels in tests. A fast growing body of research has shown that LLMs are able 1) to produce text that matches or exceeds human writing in persuasiveness (23, 52) and perceived empathy (24, 53), 2) to infer sentiment (54) and emotion (55) from text, with the ability to change its tone (56), 3) and to engage in convincing interaction and role-play with mimicking behavior that exceeds typical human levels (17, 27), matching users' personality traits (57) and linguistic styles (58) convincingly.

LLMs are now able to generate text that is highly convincing and persuasive (23, 30), with a strong increase in ability in each model generation (52). For example, in a series of experiments using GPT-3, it was shown that LLMs are on par with human authors in producing convincing texts that can change users' minds, even on polarized political issues (59). A similar study found that GPT-3 could match human writers in producing convincing propaganda (32). In another experiment a version of GPT-3 was fine-tuned to write self-presentation accounts for impersonating people across different social contexts like job applications or dating profiles; these were found to be indistinguishable from human-written accounts (60). Newer studies that use the latest generation of models, such as GPT-4, demonstrate that LLMs can exceed average human skill. One experiment found that "LLMs significantly outperform human participants across every topic and demographic, exhibiting a high level of persuasiveness" (31). This was particularly true when the LLM was given personal information about the user to tailor its conversational messages.

In addition, LLMs have been found to produce text that is often on par with, or more empathetic than human-written text. A recent between-subject study of 1,000 participants using the latest LLMs, found a "statistically significant superiority of the empathetic responding capability of LLMs over humans," with GPT-4 being the "most empathetic" (24). Another study, using a dataset of prior patient-doctor conversations collected at the Mayo clinic, found that "LLM-powered chatbots have the potential to surpass human physicians in delivering empathetic communication" (53). While there is no doubt that LLMs do not possess empathy, and some studies show mixed results [e.g., humans write more empathetically than LLMs when explicitly told to do so, but humans prefer LLM output otherwise (61)], the latest generation of LLMs is now capable of writing in ways previously thought impossible.

In addition to persuasive writing, LLMs are now capable of reliably inferring emotion and sentiment from human-written text. In one study, a fine-tuned LLM was able to detect emotions in social media posts with 84% accuracy (55). Another study found precision levels of over 95% for LLMs assessing sentiment across multiple languages, with the ability to identify nuanced feelings such as nostalgia and loyalty (54). LLMs can also rank order existing messages according to degree of empathy (53), or change the tone of written text to make it more positive, with clear effects on users' observed emotions (49). Similarly, LLMs have been shown to outperform humans in reframing written scenarios to reduce negative emotion, a skill known in humans as cognitive reappraisal (56).

It has further been shown that deliberate prompting can personalize LLMs to match users in tone and conversation styles (62). The ability to personalize text (31) goes hand in hand with the ability to assume and enact a range of different personas (19). LLMs excel at role-play, capable of impersonating a wide-range of roles and personas (27) during interaction, with the ability to mimic even nuanced linguistic character styles (58). One study (17) found that ChatGPT-4 would frequently modify its behavior during interactions as if it was learning to mimic its interlocutor's behavior. This is amplified by the ability of LLMs to infer as well—or better—than humans the beliefs and intentions of an interlocutor from text (26).

In sum, the latest generation of LLMs, which underpin modern AI products, are unlike any previous information technology. They are capable of encoding and mimicking deeply human, communicative traits so convincingly that they now pass the Turing test reliably (14, 21, 22). We emphasize, again, that these systems do not possess genuine human traits; it remains obviously true that LLMs' inner workings are fundamentally different to human cognition (63), and that they are incapable of feeling or even understanding emotion or language in any genuine sense (27, 64).

Yet, LLMs nevertheless exhibit impressive abilities to simulate such human traits at levels that are equivalent to, and often exceed above average human performance. We identified abilities in three related areas: writing, inference, and interaction. It appears that with recent increases in scale, LLMs have encoded in ever more granular detail nuanced language patterns, or styles (65), that bring about these mimicking abilities, further enhanced via fine-tuning for human-like conversation through reinforcement learning from human feedback (48).

As a result, these advances enable the creation of highly believable anthropomorphic conversational agents. The anthropomorphic qualities of LLM-based systems not only offer designers ways to create new conversational interfaces for better accessibility of existing systems, but to build anthropomorphic agents with deep communicative abilities that remain as yet underexplored. We note that the human-matching language abilities of LLMs will be further amplified by synthetic voice and face simulation techniques, in particular with advances that make faces and voices more human-like by adding intonations, inflections, and "disfluencies" (66).

The Central Question Concerning Anthropomorphism

We posit that the latest incarnations of LLMs challenge our conception of anthropomorphism. Anthropomorphism refers to the natural human tendency (67) to ascribe human-like traits to nonhuman entities (62, 68), such as animals, physical objects or digital entities. Conceptually, anthropomorphism resides in users' minds, not in said entities (69), and is found strongest during interaction (70).

In computing, anthropomorphism has typically been used to explain why and how users change their behavior during interaction when they attribute to systems certain human-like traits (71).

However, this phenomenon changes significantly when the technology already exhibits highly human-like (anthropomorphic) qualities, indistinguishable from the "real deal," as is evidenced by the Turing Test (14, 21, 22). Interestingly, the Turing Test has been characterized as "anthropomorphism-proofed" (72), because its design disincentivizes judges in the test setup from anthropomorphizing the entity they are assessing, as they must by default assume that they are interacting with a machine. This guards against falsely attributing human traits. Hence, when an LLM fools judges consistently, we may conclude that it is not the user anthropomorphizing which causes the outcome, but the anthropomorphic qualities of the system.

Conceptually, for this new class of technologies, "anthropomorphism" can no longer be thought of as originating solely from the user's mind. The term anthropomorphic agent captures this phenomenon—the ability of the LLM-based system to consistently and convincingly mimic human communication traits, which makes distinguishing them from real human interlocutors difficult if not impossible.

The central question emerging from this insight concerns the direction of development of anthropomorphic agents: should we continue designing LLM-based systems to mimic human-like writing and interaction, or should we instead work to dehumanize them? On the one hand, LLMs and related technologies come with the promise to create highly useful and easy-to-use systems, precisely because they appear human-like. On the other hand, anthropomorphic conversational agents offer never before seen abilities to deceive and manipulate users, at scale and with abilities that may exceed typical human levels. We will outline emerging benefits and emerging dangers of anthropomorphic conversational agents, before we return to the question and discuss implications, and new research questions, for the development, deployment and regulation of these systems.

Emerging Opportunities

The development of anthropomorphic agents comes with a range of opportunities. Human-like conversational interfaces, and the ability to write in ways that are tailored and personalized to users' needs, can make otherwise dense information more accessible to users. The ability to role-play and match users during interaction holds the promise of creating new kinds of agentic systems adept at tutoring, coaching, or mentoring, with the potential for better outcomes in fields as diverse as business, education or health care.

Anthropomorphic agents come with the promise of new kinds of user interfaces, spurring advances in the emerging field of conversational UX (73, 74). LLMs are increasingly thought of as an ideal front-end, such as for making dense HR information accessible to employees (75) as exemplified by systems like IBM's AskHR system that serves its large global workforce (76), or for making interaction with data analytics and visualization systems more intuitive and accessible (77). Notwithstanding existing reliability and inaccuracy concerns, it has been argued that LLMs' ability to write in engaging and accessible ways will make dense medical language or medical services more accessible to the general public (78–80). In one experimental study into mental health advice in the workplace, persons living with autism preferred GPT-4 answers over human written ones (81). For recipients the benefits of "highly affective communicative style" outweighed concerns raised by experts over some of the advice that was given.

Another promising aspect of anthropomorphic agents is the ability to role-play and match users in conversation style and personalize interactions (79). This gives rise to new kinds of tutoring or coaching services that promise engaging, targeted and personalized learning experiences (82). In education contexts it has been shown that LLMs can effectively act as tutor, mentor or coach, by merely using sophisticated prompting techniques (83). With some fine-tuning, LLMs lend themselves to develop dedicated tutoring (84) or coaching systems (85, 86). One such system is Khanmigo, an agentic tutoring chatbot that converses in Socratic style by asking questions (87); it is fine-tuned to adjust in style and difficulty to the preferences and conversational requirements of the learner (88). In health contexts, LLM-based systems were found to increase patient engagement in behavior change interventions (89), and to achieve personal weight-loss goals (90). The ability of LLMs to tailor communication to individual comprehension levels shows merit for making medical jargon more understandable (91). Finally, role-play also has a place in leisure and entertainment services, such as gaming and AI companion apps (92).

Emerging Challenges: Anthropomorphic Seduction

Technology that exhibits highly human-like abilities, often at above-average ability levels, comes with a range of novel challenges. We suggest that the general user population will not be prepared for a world full of anthropomorphic agents. Unsuspecting users will be prone to succumbing to what we term anthropomorphic seduction, the allure of persuasive writing and digital services that are indistinguishable from human interlocutors in conversation.

In popular culture, AI is typically portrayed as an all-knowing, hyperrational entity, of superior reasoning ability but struggling with human traits (93) like emotion and humor or sarcasm. For example, in Star Trek Commander Data's lack of humanity and his futile quest to understand the relational nuance of human conversations is a recurring theme (94). Such portrayals are in line with common understandings of computing more generally, as associated with veracity and faithfulness in data representation (95) and with accuracy and precision in algorithmic computation (96). Human traits conversely are understood to be error-prone and said to revolve around creativity, empathy and emotional involvement (97).

Hence, on the one hand, the general user population will expect high accuracy in computing systems but not expect the kind of abilities that allow them to mimic human traits effectively. On the other hand, users are already prone to anthropomorphizing chatbot systems (98) while being unaware of their true workings. This can manifest in "a powerful Eliza effect, in which a naive or vulnerable user may see the dialog agent as having human-like desires

and feelings" (27). Given the recent sharp increases in anthropomorphic abilities, we have already seen emerging evidence that many users readily believe that LLM-based chatbots are conscious and have feelings and memory (33, 99). Related calls to take "AI welfare seriously" are coming even from within the research community (100).

Anthropomorphic seduction presents unique dangers. It opens users up to be trusting and vulnerable toward agentic systems that interact in ways that can be deceptive, persuasive and manipulative. Earlier research into anthropomorphism has already shown that beliefs of humanness elicit more positive emotions and lead users to experience feelings of moral responsibility, increasing their inclination to "do the right thing" toward the chatbot (101). Related research showed that such anthropomorphism is positively related to user self-congruence, the sensation that users can see their own self-perception matched by the system's characteristics, which is known to increase trust in the system (57). Hence, systems with inherent anthropomorphic qualities will be able to instill trust in users and elicit goodwill, making users vulnerable to manipulation and exploitation. For example, it has been argued that user trust is increased when the system engages in reciprocal self-disclosure of information (69). Eliciting and storing personal information from users is useful for personalizing the interaction (102), but is equally concerning from a privacy point of view (103).

At the same time, advanced LLMs also possess deceptive potential that results from their convincing role-play abilities (27). It has been shown that LLMs are able to be outright deceptive in the production of text—defined as the "systematic inducement of false beliefs in others—as a means to accomplish some outcome other than saying what is true" (25), or to engage in targeted harmful writing (27). For example, recent research by Anthropic found that its Claude 3 system was most persuasive when allowed to fabricate information and engage in deception (52). It was also found to be much better than human writers in producing deceptive arguments. Whereas human writers might find it difficult to abstract from facts and let go of the truth, or to set aside moral and ethical convictions, LLMs are able to produce texts free from any such human inhibitions.

We argue that these qualities render advanced LLMs into potential manipulation machines, anthropomorphic agents that are able to instill trust in users, while being deceptive without human moral or ethical inhibitions. This is most concerning with the availability of powerful open source models that can be aligned or jail-broken for malicious tasks (104), outside of the responsibility frameworks or guardrails that large providers like OpenAI or Anthropic have put in place to some degree (105).

Discussion

We return to the central question. Should we lean into the human-like qualities of LLMs for creating anthropomorphic conversational agents, and follow advice to treat LLM-based systems akin to people for best interaction outcomes (106)? Or should we instead find ways to dehumanize these systems by design, and educate users to resist anthropomorphic seduction (16, 107)? Or more pragmatically, will we be able to reap the benefits of anthropomorphic agents, without opening the door for anthropomorphic seduction with its associated risks of deception and manipulation?

Calls that caution against researchers, the media or the general public anthropomorphizing LLMs and AI systems are not new (16, 28). We suggest however that any such calls will increasingly fall flat when the systems themselves already exhibit inherently anthropomorphic qualities. When the interaction with an AI system looks and feels every bit like interacting with another human, refraining from naturally anthropomorphizing the entity will become increasingly more difficult, especially for the general user less attuned to AI's characteristics and limitations.

Given significant ongoing investments, AI will continue to proliferate across all facets of daily life (108). The question then is—do its anthropomorphic abilities require our dedicated attention? On the one hand, we could take any warnings about dangers as the kind of moral panics that accompany every significant technological change in society (109). We might simply wait for the new technology to be absorbed and normalized. On the other hand, the potential dangers of anthropomorphic seduction appear real, manifest and potentially far-reaching. For an example of what happens when matters of design, deployment and regulation are not addressed early on, we suggest looking to social media as the most recent wave of public transformative technology. For all its positive and inclusive outcomes (110, 111), there is mounting evidence that social media has had outsize effects on matters such as public discourse and free speech (112), or mental health of young people (113–115). Real and present concerns exist over the role of mis- and disinformation in public discourse (116–118), the emergence of echo chambers (119) and deepening political polarization (120, 121), in particular around big societal issues like climate change (122). The reasons appear manifold and complex; they include design choices that make social media use addictive (123), the role of algorithms in the curation and distribution of content (124, 125), business models that have evolved to prioritize engagement and platform over user welfare (112, 126), and a favorable regulatory environment (127).

Should we act then? Anthropomorphic conversational agents appear unlike any prior technology in both their nature and their speed of proliferation, with potentially more far-reaching effects than social media. Historian and philosopher Yuval Harari refers to language as "the operating system of human civilization" (128); LLM's highly nuanced mastery of language raises fears for how AI might influence human story telling, history making, and the fabric of society. A technology that impersonates humans convincingly has the potential to be misappropriated, and to fracture societal processes in ways no prior technology was able to. At the same time, its use is proliferating at far greater speed than social media. It took ChatGPT two months to achieve 100 million monthly active users; a milestone that took Instagram 2.5 y and TikTok 9 mo to achieve (129). Against this background, inaction appears risky. We argue that we must engage with the topic urgently and swiftly, because the next two years will likely be crucial as business models around anthropomorphic agents form and solidify.

What then can be done? It appears too early to give clear directives, as the technology is still progressing fast,

and the effects of anthropomorphic agents are not yet fully understood. While a flurry of early research has emerged, the efforts are disconnected. Yet the case of social media offers some guidance. Levers of change will include system design, commercial models of deployment, and regulatory guidance and oversight. For each we discuss implications and research directions.

Implications

The implications of machines with highly nuanced language mastery that can mimic and impersonate human communication believably, but lack human understanding or ethical inhibitions, are potentially far-reaching, especially when the technology can now be amplified with natural voice (130, 131) and realistic face technology (5).

Design and Development. A core consideration for designers should be the following: how can the field exploit anthropomorphic abilities, across writing, inference and interaction, without creating systems that are deceptive and open to misappropriation? We have argued that the field is not currently concerned enough with anthropomorphic qualities of LLMs, which emerged mostly as a byproduct in the pursuit of human-like reasoning abilities, or artificial "intelligence" more broadly. This is mirrored in how leading AI companies conceptualize the dangers from AI. Currently, the most frequently raised dangers are those stemming either from propagating harmful content, like plans for creating weapons or malicious computer code (104), or the emergence of so-called "superintelligence" (132). Given the highly disputed, hypothetical prospect of superintelligence (133, 134), we suggest that potential dangers from "supercommunicators," anthropomorphic agents with highly convincing communicative and language abilities, deserve more immediate attention.

How then can AI developers incorporate such considerations in their practices? One important way is to refrain from actively anthropomorphizing their creations by design. Companies like OpenAI or Anthropic routinely use language that deliberately evokes humanness, in both how they describe their products (135), and the actual systems themselves, which routinely purport to "think," "reason," "evaluate," "believe," or "understand," despite occasional reminders that "as a LLM I do not have opinions." It should be possible to build useful systems that do not unnecessarily evoke human-likenesses beyond their evident communicative abilities. We suggest that terms like "seeing," "thinking," "reasoning" can usefully be replaced with terms such as "recognizing," "computing," "inferring," which in turn would also more precisely reflect the technical processes involved.

Another way is to derive principles of responsible anthropomorphic design and deliberately amend the behavior of the systems themselves. How can design be altered to remind users that they do not in fact interact with another human, yet still preserve the ease-of-use of natural conversation? One could adjust the language, to deanthropomorphize it, make it more neutral, avoid evoking emotion, or to introduce a certain level of friction. It has been suggested that "AI applications could use language that is clearly not written by humans without loss of functionality" (60), such as by creating a dedicated AI accent, or machine dialect, which would clearly indicate when a machine is speaking (60). In addition, it seems prudent to learn from the history of social media and avoid optimizing user interfaces for engagement, or addictiveness (136), and to avoid being overly data hungry. We suggest that agentic systems designed to draw the user into longer conversations will be more seductive, and this will increase the ability to collect and store information about the user.

In order to guide the responsible development of anthropomorphic agents it is important to gain a nuanced and informed understanding of both the kinds, and degrees, of anthropomorphic qualities that LLM-based systems possess. Much research effort has gone into benchmarks to compare the performance of LLMs. Popular tests, such as MMLU (137), BIG-Bench (138), BIG-Bench Hard (BBH) (139), or AGIEval (140) are typically based on human ability tasks (137) but test for knowledge recall or problem-solving (141), often by comparing performance against human experts doing the same tasks (138). Whereas these tests measure LLM abilities in the pursuit of intelligence or reasoning capabilities, dedicated tests that evaluate anthropomorphic qualities are lacking. Exceptions are tests aiming to measure empathy (142), such as GIEBench (143).

We propose that the field set out to research and develop comprehensive benchmarks to discern the levels of anthropomorphic quality of LLMs as the basis for decisions about their responsible development and deployment. The Turing test (20), properly understood as a test for anthropomorphic abilities rather than intelligence, provides a starting point. New tests would either pit LLMs against human interlocutors or against other LLMs. Tests would measure and provide scores across the three main categories of anthropomorphic qualities—writing, inference and interaction.

Deployment and Use. The appropriate degree of requisite anthropomorphic qualities will be highly context-dependent. Some use cases will benefit from high degrees of humanness, such as the use of anthropomorphic agents for role-play in training situations (144, 145), or as tutors and coaches in education (84, 87). In such contexts it would be made clear that interaction takes place with a nonhuman entity, while the interaction itself would benefit from high levels of human-like communicative ability and natural conversation flow. For example, it has been shown that technologies with human likeness can break down barriers in reporting mental health symptoms from PTSD, such as in young war veterans (146, 147), when human likeness enables natural conversation, but the patient is explicitly aware of interacting with a machine and thus does not feel exposed to human judgment.

In other contexts, decisions and their outcomes about the requisite degree, and presentation of, human likeness are much less clear cut. There has been a recent surge in so-called AI companion apps. These services provide anthropomorphic agents fine-tuned to engage in ongoing role-play with users and to act as companions via everyday conversation. As these services lean heavily into the anthropomorphic qualities of LLMs to create a convincing illusion of human-likeness, they provide an important use case. While AI companions have been credited with alleviating

feelings of loneliness in users, or even alleviating suicidal ideation (148, 149), they have also been actively implicated in cases of self-harm* and suicide (150). It has been argued that these apps might be deceptive and exploitative (149, 151), and that the experienced relief from loneliness could be short-lived (92). As we have argued, systems that heavily optimize for anthropomorphic quality raise serious questions regarding responsibilities given issues associated with anthropomorphic seduction. This is underlined by a case in which the social companion provider Replika made significant changes to its chatbots, scaling back their capacity to engage in romantic exchanges, which left many users "distraught" and with a "profound sense of loss" (152).

Other contexts have shown equally mixed outcomes. For example, while LLMs' persuasive writing has been associated with deception (25, 52) and harmful outcomes (27), recent studies have shown that it can equally be used to dissuade users of their beliefs in conspiracy theories (153), or drive users' willingness to make donations for positive outcomes (55).

We suggest that dedicated research is needed to study the effects of anthropomorphic agents in support of responsible decisions about their deployment. We briefly highlight research questions and methodological considerations in pursuit of a new research program into the benefits and dangers of anthropomorphic conversational agents. Such a program will include research into the immediate effects and outcomes at the user level, as well as the long-term, systemic-societal impacts over time. Questions might include, but are not limited to:

- What is the relationship between various anthropomorphic qualities, across the three categories of 1) persuasive and empathetic writing, 2) inference of user traits, and 3) convincing role-play and user matching during interaction, and anthropomorphic seduction as an outcome?
- What are the effects of (and different degrees of) anthropomorphic seduction in various deployment contexts?
- What cues in conversation, or interface design, help users know they are interacting with AI?
- What role do audiorealistic voice or photorealistic face interfaces play in interaction with anthropomorphic agents?
- What constitutes a valid baseline for comparing the effects of human–agent interaction? Would this include human–human conversation, and/or more traditional human–computer interaction?
- How do we judge, what "good," "effective," and "responsible" human–agent interaction looks like?
- How do we control for novelty effects in studying human–agent interaction, when most end-users will be unsuspecting, and might not have been exposed to systems with full anthropomorphic abilities?
- How do demographics, or AI literacy, mediate effects like anthropomorphic seduction?
- What happens to our social interactions in the long term when we get used to consistently above-average conversations with our AI agents?
- How will anthropomorphic agents interact with other conversational technologies, most notably social media, with its known effects and issues?

Policy and Regulation. The design and deployment of anthropomorphic agents will require dedicated regulatory attention as these entities meet an unsuspecting and unprepared public. Studies have shown that humans are largely ineffective at discerning AI created from human created content, and thus easily deceived (60). Policy-makers and industry regulators should be aware of the risks and dangers of anthropomorphic agents, which flow directly from their advanced abilities in writing, inference, and interaction. These abilities come with a number of concrete risks. The ability to write persuasively poses risks of tailored disinformation or propaganda messaging at scale, at quality levels not previously seen. The ability to inconspicuously infer user traits from text or in interaction (154) raises privacy questions and opens the door for new kinds of phishing or social engineering attacks (155). The ability to convincingly match and role-play poses risks of behavior manipulation at scale, such as highly effective predatory sales tactics or new deception schemes.

Not all problematic applications of anthropomorphic agents will fall outside the law. The next two to three years will see the proliferation of business models monetizing the anthropomorphic abilities of LLMs with related technologies such as synthetic voice and face interfaces. In the age of surveillance capitalism (126), the ability to converse persuasively with predetermined intent might prove irresistible as tools for highly effective advertising to unsuspecting consumers. Regulators in critical industries like health, legal or financial services should take note and investigate requirements for new consumer protections.

We suggest policy-makers consider implications across the three areas of 1) risk level, 2) transparency, and 3) mitigation, with the potential to derive safety rating systems for anthropomorphic conversational agents, akin to ratings for entertainment content in cinema, television or gaming (156). As discussed above, evaluating the risk levels from anthropomorphic abilities will require new tests and benchmarks to ascertain degrees of anthropomorphic ability, and research on their contextual effects and outcomes. Rating scales based on such tests could then provide a risk indicator for levels of human likeness.

Systems that embody anthropomorphic abilities at levels that match or exceed most humans require transparency and should come with appropriate labeling and disclosure. For example, the EU AI Act, the world's first comprehensive AI law (157), lists transparency as one of its key principles. It stipulates that "the provider must inform individuals about their interaction with an AI system if it is not readily apparent to the user," aiming to counteract AI deception. However, it remains to be seen how effective such warnings will be, given mixed results with labeling in the past, e.g., with health warnings on cigarette packs (158), or labels indicating misinformation online, which have shown to be effective in some settings but ineffective in others (159). In the United States, concerns have also been raised that blanket disclosure mandates could restrict certain kinds of protected speech undertaken via AI agents (160).

Systems that pose risks of anthropomorphic seduction should also have built-in safeguard mechanisms for mitigating foreseeable negative effects. For example, AI companion apps that are able to infer user emotion should monitor user well-being and detect any potential signs of self-harm

*https://www.npr.org/2024/12/10/nx-s1-5222574/kids-character-ai-lawsuit.

ideation. More research is needed however, since it has been shown that LLMs are much better at picking up positive emotions than negative ones, likely because of their intensive fine-tuning on positive conversation examples (24). This could render safeguards that rely on inherent LLM abilities ineffective, and require dedicated fine-tuning for harm prevention, or a reduction in anthropomorphic quality by design (136).

Notwithstanding these ideas, any efforts to regulate anthropomorphic agents will encounter a plethora of practical questions, beginning with who gets to regulate such systems, given the global nature of "big tech" AI companies. Potential questions include, but are not limited to,

- How will cases of anthropomorphic seduction or deception be discovered, reported and assessed?
- Who bears responsibility when AI systems cause accidental harm through deception or manipulation?
- At what level will regulation be targeted, developers of end-user services, or providers of LLM foundation models implicated in such services?
- How can regulators enforce responsible design principles at either level?
- How can transparency of agentic systems be enforced? Should access to algorithms and training data by independent research institutions be mandated in the interest of public safety?

In the absence of concrete regulation, and as regulation in this space is still emerging, we would need to rely on the awareness and voluntary restraint of developers and system providers to dehumanize their systems, given how readily LLMs can be instructed to evoke and mimic humanness (61). However, we suggest that these actors will be under increasing commercial pressure to take full advantage of their ability to fine-tune LLMs for increased human likeness, creating highly effective anthropomorphic agents that exploit anthropomorphic qualities for economic gain, with potential unintended consequences in the long run.

Conclusion

In this perspective, we introduced and described anthropomorphic conversational agents, an emerging class of systems that make use of advanced language abilities of LLMs, comprising 1) highly persuasive writing, 2) inferring and matching of user traits, and 3) role-playing at levels that consistently pass the Turing test. While the field of AI is building toward intelligence or smartness, LLMs have developed highly sophisticated anthropomorphic abilities almost as a by-product. Experts might be split over whether or not we have, or will, conjure true intelligence from AI systems, but we surely have created technology that mimics and impersonates humans in communication, with abilities that match and increasingly exceed most humans.

The development of anthropomorphic agents comes with the promise to make computing accessible in ways not seen before, by enabling interaction with computers as if with a fellow human. This promise needs to be weighed against the obvious danger that any such impersonation of human likeness also opens the door for highly effective manipulation at scale. We have presented in detail research efforts in this emerging field, and outlined both opportunities and challenges of anthropomorphic conversational agents. Given the potential dangers from anthropomorphic seduction, which stems from the inability of unsuspecting users to tell the difference between human and machine, we outlined a range of implications for design and development, deployment and use, and regulation and policy-making of anthropomorphic agents.

Anthropomorphic conversational agents have now outgrown the conventional notion of anthropomorphism, whereby humans ascribe human-like qualities to nonhuman entities. When technology exhibits inherently human-like qualities that make telling the difference difficult and increasingly impossible, we must recognize that the technology has in itself become anthropomorphic, be that through representation with photorealistic faces, through synthetic voices, or simply by way of human-like, text-based interaction. As we have shown, it is the latest advances in LLMs that underlay the foundations of anthropomorphic agents. In this new era, it will be incumbent upon researchers, developers, and policy-makers to better understand and recognize their benefits but also their dangers.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. J.D.W was supported by the University of Washington's Center for an Informed Public, the John S. and James L. Knight Foundation (G-2019-58788), and the Institute of Museum and Library Services (LG-255047-OLS-23).

1. A. M. Turing, *Computing Machinery and Intelligence* (Springer, 2009).
2. E. Brynjolfsson, *The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence in Augmented education in the Global Age* (Routledge, 2023), pp. 103–116.
3. V. Pitardi, H. R. Marriott, Alexa, she's not human but. unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychol. Mark.* **38**, 626–642 (2021).
4. K. Seaborn, N. P. Miyake, P. Pennefather, M. Otake-Matsuura, Voice in human-agent interaction: A survey. *ACM Comput. Surv.* **54**, 81:1–81:43 (2021).
5. M. Seymour, K. Riemer, J. Kay, Actors, avatars and agents: Potentials and implications of natural face technology for the creation of realistic visual presence. *J. Assoc. Inf. Syst.* **19**, 953–981 (2018).
6. T. Kojima, S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **35**, 22199–22213 (2022).
7. T. Hagendorff, S. Fabi, M. Kosinski, Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat. Comput. Sci.* **3**, 833–838 (2023).
8. T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).
9. M. Lewis, M. Mitchell, Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2402.08955 (Accessed 29 April 2025).
10. D. Hodel, J. D. West, Response: Emergent analogical reasoning in large language models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2308.16118 (Accessed 29 April 2025).
11. M. Mitchell, D. C. Krakauer, The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2215907120 (2023).
12. M. Mitchell, Artificial intelligence learns to reason. *Science* **387**, eadw5211 (2025).
13. T. Ullman, Large language models fail on trivial alterations to theory-of-mind tasks. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2302.08399 (Accessed 29 April 2025).
14. C. Biever, ChatGPT broke the Turing test–The race is on for new ways to assess AI. *Nature* **619**, 686–689 (2023).
15. S. Yadlowsky, L. Doshi, N. Tripuraneni, Pretraining data mixtures enable narrow model selection capabilities in transformer models. arXiv [Preprint] (2023). https://arxiv.org/abs/2311.00871 (Accessed 29 April 2025).
16. M. Shanahan, Talking about large language models. *Commun. ACM* **67**, 68–79 (2024).
17. Q. Mei, Y. Xie, W. Yuan, M. O. Jackson, A Turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2313925121 (2024).
18. A. Salecha *et al.*, Large language models show human-like social desirability biases in survey responses. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2405.06058 (Accessed 29 April 2025).
19. G. Aher, R. I. Arriaga, A. T. Kalai, Using large language models to simulate multiple humans and replicate human subject studies. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2208.10264 (Accessed 29 April 2025).

20. A. Turing, *Computing Machinery and Intelligence* (Mind LIX, 1950), pp. 433–460.
21. C. R. Jones, B. K. Bergen, People cannot distinguish GPT-4 from a human in a Turing test. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2405.08007 (Accessed 29 April 2025).
22. C. R. Jones, B. K. Bergen, Large language models pass the Turing test. arXiv [Preprint] (2025). https://arxiv.org/abs/2503.23674 (Accessed 29 April 2025).
23. S. M. Breum, D. V. Egdal, V. G. Mortensen, A. G. Møller, L. M. Aiello, The persuasive power of large language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2312.15523 (Accessed 29 April 2025).
24. A. Welivita, P. Pu, Are large language models more empathetic than humans? arXiv [Preprint] (2024). https://arxiv.org/abs/2406.05063 (Accessed 29 April 2025).
25. P. S. Park, S. Goldstein, A. O'Gara, M. Chen, D. Hendrycks, AI deception: A survey of examples, risks, and potential solutions. *Patterns* **5**, 100988 (2024).
26. J. W. A. Strachan *et al.*, Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* **8**, 1285–1295 (2024).
27. M. Shanahan, K. McDonell, L. Reynolds, Role play with large language models. *Nature* **623**, 493–498 (2023).
28. N. Inie, S. Druga, P. Zukerman, E. M. Bender, "From "AI" to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust?" in *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024), pp. 2322–2347.
29. M. Binz *et al.*, How should the advent of large language models affect the practice of science?. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401227121. (2025).
30. E. Karinshak, S. X. Liu, J. S. Park, J. T. Hancock, "Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages" in *Proceedings of the ACM on Human-Computer Interaction* (2023), vol. 7.
31. F. Salvi, M. Horta Ribeiro, R. Gallotti, R. West, On the conversational persuasiveness of large language models: A randomized controlled trial. arXiv [Preprint] (2024). https://arxiv.org/abs/2403.14380 (Accessed 29 April 2025).
32. J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, M. Tomz, How persuasive is AI-generated propaganda? *PNAS Nexus* **3**, pgae034 (2024).
33. C. Colombatto, S. M. Fleming, Folk psychological attributions of consciousness to large language models. *Neurosci. Conscious.* **2024**, niae013 (2024).
34. R. Bommasani *et al.*, On the opportunities and risks of foundation models. arXiv [Preprint] (2021). https://arxiv.org/abs/2108.07258 (Accessed 29 April 2025).
35. A. Vaswani *et al.*, Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1–11 (2017).
36. A. Asperti, V. Tonelli, Comparing the latent space of generative models. *Neural Comput. Appl.* **35**, 3155–3172 (2023).
37. T. B. Lee, S. Trott, *A Jargon-Free Explanation of How AI Large Language Models Work* (ARSTechnica, 2023).
38. G. Branwen, The scaling hypothesis (2022). *https://gwern.net/scaling-hypothesis*. Accessed 29 April 2025.
39. L. Tudor Car *et al.*, Conversational agents in health care: Scoping review and conceptual analysis. *J. Med. Internet Res.* **22**, e17158 (2020).
40. S. Shahriar, K. Hayawi, Let's have a chat! a conversation with ChatGPT: Technology, applications, and limitations *Artifi. Intell. Appl.* **2**, 11–20 (2023).
41. J. Huang, K. C. C. Chang, Towards reasoning in large language models: A survey. arXiv [Preprint] (2022). https://arxiv.org/abs/2212.10403 (Accessed 29 April 2025).
42. P. Li, W. Lam, L. Bing, Z. Wang, Deep recurrent generative decoder for abstractive text summarization. arXiv [Preprint] (2017). https://arxiv.org/abs/1708.00625 (Accessed 29 April 2025).
43. S. Altman, Planning for AGI and beyond. *OpenAI Blog* (2023). https://openai.com/index/planning-for-agi-and-beyond/. Accessed 29 April 2025
44. M. Parmar *et al.*, Towards systematic evaluation of logical reasoning ability of large language models. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2404.15522. (Accessed 29 April 2025).
45. X. Yang *et al.*, Zero-shot evaluation of (multimodal) large language models on multimodal reasoning tasks. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2405.07229. (Accessed 29 April 2025).
46. J. Yu *et al.*, KoLA: Carefully benchmarking world knowledge of large language models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2306.09296. (Accessed 29 April 2025).
47. D. Altinok, D-NLP at SemEval-2024 Task 2: Evaluating clinical inference capabilities of large language models. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2405.04170. (Accessed 29 April 2025).
48. D. M. Ziegler *et al.*, Fine-tuning language models from human preferences. arXiv [Preprint] (2019). https://arxiv.org/abs/1909.08593. (Accessed 29 April 2025).
49. Z. Zou, O. Mubin, F. Alnajjar, L. Ali, A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires. *Sci. Rep.* **14**, 2781 (2024).
50. L. Huang *et al.*, A survey on hallucination in large language models: Principles, taxonomy, challenges. and open questions. arXiv [Preprint] (2023). https://arxiv.org/abs/2311.05232. (Accessed 29 April 2025).
51. S. Imani, L. Du, H. Shrivastava, MathPrompter: Mathematical reasoning using large language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2303.05398. (Accessed 29 April 2025).
52. E. Durmus *et al.*, Measuring the persuasiveness of language models. *Anthropic Blog* (2024). https://www.anthropic.com/research/measuring-model-persuasiveness. Accessed 29 April 2025.
53. M. Luo, C. J. Warren, L. Cheng, H. M. Abdul-Muhsin, I. Banerjee, Assessing empathy in large language models with real-world physician–patient interactions. arXiv [Preprint] (2024). https://arxiv.org/abs/2405.16402.
54. J. O. Krugmann, J. Hartmann, Sentiment analysis in the age of generative AI. *Cust. Needs Solut.* **11**, 3 (2024).
55. S. J. Lee, L. Paas, H. S. Ahn, The power of specific emotion analysis in predicting donations: A comparative empirical study between sentiment and specific emotion analysis in social media. *Int. J. Mark. Res.* **66**, 610–630 (2024).
56. J. Z. Li, A. Herderich, A. Goldenberg, Skill but not effort drive GPT overperformance over humans in cognitive reframing of negative scenarios. *OSF Preprint* (2024). https://osf.io/preprints/psyarxiv/fzvd8_v2. Accessed 29 April 2025.
57. A. Alabed, A. Javornik, D. Gregory-Smith, Ai anthropomorphism and its effect on users' self-congruence and self-AI integration: A theoretical framework and research agenda. *Technol. Forecast. Soc. Chang.* **182**, 121786 (2022).
58. S. Chen *et al.*, A multi-task role-playing agent capable of imitating character linguistic styles. arXiv [Preprint] (2024). https://arxiv.org/abs/2411.02457 (Accessed 29 April 2025).
59. H. Bai, J. G. Voelkel, J. C. Eichstaedt, R. Willer, Artificial intelligence can persuade humans on political issues. *OSF Preprints* (2023). https://osf.io/preprints/osf/stakv_v6. Accessed 29 April 2025.
60. M. Jakesch, J. T. Hancock, M. Naaman, Human heuristics for AI-generated language are flawed. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2208839120 (2023).
61. B. Kleinberg *et al.*, Trying to be human: Linguistic traces of stochastic empathy in language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2410.01675 (Accessed 29 April 2025).
62. A. Deshpande, T. Rajpurohit, K. Narasimhan, A. Kalyan, Anthropomorphization of AI: Opportunities and risks. arXiv [Preprint] (2023). https://arxiv.org/abs/2305.14784 (Accessed 29 April 2025).
63. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
64. K. Mahowald *et al.*, Dissociating language and thought in large language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2301.06627v3 (Accessed 29 April 2025).
65. K. Riemer, S. Peter, Conceptualizing generative AI as style engines: Application archetypes and implications. *Int. J. Inf. Manage.* **79**, 102824 (2024).
66. R. Chaudhury, M. Godbole, A. Garg, J. H. Seo, Humane speech synthesis through zero-shot emotion and disfluency generation. arXiv [Preprint] (2024). https://arxiv.org/abs/2404.01339 (Accessed 29 April 2025).
67. N. Epley, A. Waytz, J. T. Cacioppo, On seeing human: A three-factor theory of anthropomorphism. *Psychol. Rev.* **114**, 864 (2007).
68. G. Airenti, The cognitive basis of anthropomorphism: From relatedness to empathy. *Int. J. Soc. Rob.* **7**, 117–127 (2015).
69. K. Saffarizadeh, M. Keil, M. Boodraj, T. Alashoor, "My Name is Alexa. What's Your Name?" The Impact of Reciprocal Self-Disclosure on Post-Interaction Trust in Conversational Agents *J. Assoc. Inf. Syst.* **25**, 528–568 (2024).
70. G. Airenti, The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Front. Psychol.* **9**, 2136 (2018).
71. C. Crolic, F. Thomaz, R. Hadi, A. T. Stephen, Blame the bot: anthropomorphism and anger in customer–chatbot interactions. *J. Mark.* **86**, 132–148 (2022).
72. D. Proudfoot, Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artif. Intell.* **175**, 950–957 (2011).
73. R. J. Moore, S. An, G. J. Ren, The IBM natural conversation framework: A new paradigm for conversational UX design. *Human-Computer Interact.* **38**, 168–193 (2022).
74. R. J. Moore, R. Arar, *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework* (Morgan & Claypool, 2019).
75. W. Xu, J. Desai, F. Wu, J. Valvoda, S. H. Sengamedu, HR-agent: A task-oriented dialogue (TOD) LLM agent tailored for HR applications. arXiv [Preprint] (2024). https://arxiv.org/abs/2410.11239 (Accessed 29 April 2025).
76. E. Burleigh, IBM CHRO reveals key AI chatbot rollout strategy lessons. *Fortune* (2024). https://fortune.com/2024/07/12/ibm-chro-ai-chatbot-rollout-strategy-lessons/. Accessed 29 April 2025.
77. M. Hutchinson, R. Jianu, A. Slingsby, P. Madhyastha, "LLM-assisted visual analytics: Opportunities and challenges" in *Proceedings of EG UK Computer, Graphics & Visual Computing 2024* (2024).
78. D. Wang, S. Zhang, Large language models in medical and healthcare fields: Applications, advances, and challenges. *Artif. Intell. Rev.* **57**, 299 (2024).
79. J. Clusmann *et al.*, The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
80. A. Deroy, S. Maity, "Cancer-answer: Empowering cancer care with advanced large language models" in *Proceedings of the FIRE 2024 Conference (Track: Conversational System for Differential Diagnosis of GI Cancer)* (2024).
81. J. J. Jang, S. Moharana, P. Carrington, A. Begel, ""It's the only thing I can trust": Envisioning large language model use by autistic workers for communication assistance" in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (Honolulu, HI, 2024).
82. M. Park, S. Kim, S. Lee, S. Kwon, K. Kim, "Empowering personalized learning through a conversation-based tutoring system with student modeling" in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024).
83. E. R. Mollick, L. Mollick, Assigning AI: Seven approaches for students, with prompts. *The Wharton School Research Paper* (2023), Available at SSRN. https://ssrn.com/abstract=4475995. Accessed 29 April 2025.
84. S. Zha *et al.*, "An intelligent agent for mentoring students in the creative problem solving process" in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan, 2025).
85. R. Shea, A. Kallala, X. L. Liu, M. W. Morris, Z. Yu, "ACE: A LLM-based negotiation coaching system" in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Singapore, 2024).
86. H. Huang, S. Wang, H. Liu, H. Wang, Y. Wang, "Benchmarking large language models on communicative medical coaching: A novel system and dataset" in *Findings of the Association for Computational Linguistics: ACL 2024* (Toronto, ON, Canada, 2024).
87. W. A. Sahlman, A. M. Ciechanover, E. Grandjean, *Khanmigo: Revolutionizing Learning with GenAI (HBS Case Collection)* (Harvard Business School Case, 2024), pp. 824–859.
88. J. Tyrangiel, An 'education legend' has created an AI that will change your mind about AI. *The Washington Post* (2024). https://www.washingtonpost.com/opinions/2024/02/22/artificial-intelligence-sal-khan/. Accessed 29 April 2025.
89. H. Kumar *et al.*, Large language model agents for improving engagement with behavior change interventions: Application to digital mindfulness. arXiv [Preprint] (2024). https://arxiv.org/abs/2407.13067 (Accessed 29 April 2025).

90. OpenAI, Healthify. *OpenAI Blog* (2024). https://openai.com/index/healthify/. Accessed 29 April 2025.
91. J. H. Lim, S. Kwon, Z. Yao, J. P. Lalor, H. Yu, Large language model-based role-playing for personalized medical jargon extraction. arXiv [Preprint] (2024). https://arxiv.org/abs/2408.05555 (Accessed 29 April 2025).
92. K. Roose, Meet my a.i. friends. *NY Times* (2024). https://www.nytimes.com/2024/05/09/technology/meet-my-ai-friends.html. Accessed 29 April 2025.
93. I. Hermann, Artificial intelligence in fiction: Between narratives and metaphors. *AI Soc.* **38**, 319–329 (2023).
94. E. Rashkin, Data learns to dance: "Star trek" and the quest to be human. *Am. Imago* **68**, 321–346 (2011) (Independent Voices in Psychoanalysis, Summer 2011).
95. A. Burton-Jones, J. Recker, M. Indulska, P. Green, R. Weber, Assessing representation theory with a framework for pursuing success and failure. *MIS Q.* **41**, 1307–1333 (2017).
96. P. Dourish, Algorithms and their others: Algorithmic culture in context. *Big Data Soc.* **3**, 1–11 (2016).
97. J. Manyika, K. Sneader, *AI, Automation, and the Future of Work: Ten Things to Solve for* (McKinsey & Company, 2018).
98. S. Gibbons, T. Mugunthan, J. Nielsen, *The 4 Degrees of Anthropomorphism of Generative AI* (Nielsen Norman Group, 2023).
99. R. Booth, AI could cause social ruptures between people who disagree on its sentience. *The Guardian* (2024). https://www.theguardian.com/technology/2024/nov/17/ai-could-cause-social-ruptures-between-people-who-disagree-on-its-sentience. Accessed 29 April 2025.
100. R. Long *et al.*, Taking AI welfare seriously. arXiv [Preprint] (2024). https://arxiv.org/abs/2411.00986 (Accessed 29 April 2025).
101. M. Giroux, J. Kim, J. C. Lee, J. Park, Artificial Intelligence and declined guilt: Retailing morality comparison between human and AI. *J. Bus. Ethics* **178**, 1027–1041 (2022).
102. OpenAI, Memory and new controls for ChatGPT. *OpenAI Blog* (2024). https://openai.com/index/memory-and-new-controls-for-chatgpt/. Accessed 29 April 2025.
103. S. Zhang *et al.*, "Ghost of the past": Identifying and resolving privacy leakage from LLM's memory through proactive user interaction. arXiv [Preprint] (2024). https://arxiv.org/abs/2410.14931. Accessed 29 April 2025.
104. Z. Xu, R. Huang, C. Chen, S. Wang, X. Wang, Uncovering safety risks of large language models through concept activation vector. arXiv [Preprint] (2024). https://arxiv.org/abs/2404.12038. Accessed 29 April 2025.
105. Anthropic, Core views on AI safety: When, why, what, and how (2023). https://www.anthropic.com/news/core-views-on-ai-safety. Accessed 29 April 2025.
106. E. Mollick, *Co-Intelligence: Living and Working with AI* (Penguin Publishing Group, 2024).
107. A. Placani, Anthropomorphism in AI: hype and fallacy. *AI Ethics* **4**, 691–698 (2024).
108. J. Letzing, *To Fully Appreciate AI Expectations, Look to the Trillions Being Invested* (World Economic Forum, 2024).
109. A. Orben, The Sisyphean cycle of technology panics. *Perspect. Psychol. Sci.* **15**, 1143–1157 (2020).
110. B. N. Rao, V. Kalyani, A study on positive and negative effects of social media on society. *J. Sci. Technol.* **7**, 46–54 (2022).
111. Y. K. Dwivedi *et al.*, Social media: The good, the bad, and the ugly. *Inf. Syst. Front.* **20**, 419–423 (2018).
112. K. Riemer, S. Peter, Algorithmic audiencing: Why we need to rethink free speech on social media. *J. Inf. Technol.* **36**, 427–445 (2021).
113. A. Dane, K. Bhatia, The social media diet: A scoping review to investigate the association between social media, body image and eating disorders amongst young people. *PLoS Glob. Public Health* **3**, e0001091 (2023).
114. D. Pagliaccio *et al.*, Probing the digital exposome: Associations of social media use patterns with youth mental health. *Nat. Mental Health* **1**, 6–17 (2024).
115. J. Haidt, *The Anxious Generation: How the Great Rewiring of Childhood is Causing an Epidemic of Mental Illness* (Penguin Press, New York, NY, 2024).
116. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
117. A. Bovet, H. A. Makse, Influence of fake news in twitter during the 2016 US presidential election. *Nat. Commun.* **10**, 7 (2019).
118. D. Ruths, The misinformation machine. *Science* **363**, 348 (2019).
119. M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023301118 (2021).
120. A. Bessi *et al.*, Users polarization on Facebook and Youtube. *PLoS ONE* **11**, e0159641 (2016).
121. J. Flamino *et al.*, Political polarization of news media and influencers on twitter in the 2016 and 2020 US presidential elections. *Nat. Hum. Behav.* **7**, 904–916 (2023).
122. M. Falkenberg *et al.*, Growing polarization around climate change on social media. *Nat. Clim. Change* **12**, 1114–1121 (2022).
123. M. Flayelle *et al.*, A taxonomy of technology design features that promote potentially addictive online behaviours. *Nat. Rev. Psychol.* **2**, 136–150 (2023).
124. C. Peterson-Salahuddin, N. Diakopoulos, Negotiated autonomy: The role of social media algorithms in editorial decision making. *Media Commun.* **8**, 27–38 (2020).
125. F. Saurwein, C. Spencer-Smith, Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media Commun.* **9**, 222–233 (2021).
126. S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019).
127. M. B. Lawrence, Public health law's digital frontier: Addictive design, section 230, and the freedom of speech. *J. Free. Speech Law* **4**, 299–342 (2023).
128. Y. N. Harari, Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. *The Economist* (2023). https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation. Accessed 29 April 2025.
129. A. R. Chow, How ChatGPT managed to grow faster than TikTok or Instagram (TIME) (2023). https://time.com/6253615/chatgpt-fastest-growing/. Accessed 29 April 2025.
130. OpenAI, Hello GPT-4O. *OpenAI Blog* (2024). https://openai.com/index/hello-gpt-4o/, Accessed 29 April 2025.
131. S. Samuel, People are falling in love with–And getting addicted to–AI voices: Even OpenAI warns that chatting with an AI voice can breed "emotional reliance." *Vox*, 18 August 2024.
132. J. Leike, I. Sutskever, Introducing superalignment. *OpenAI* (2023). https://openai.com/index/introducing-superalignment/. Accessed 29 April 2025.
133. J. Lanier, There is no a.i. *The New Yorker* (2023). https://www.newyorker.com/science/annals-of-artificial-intelligence/there-is-no-ai. Accessed 29 April 2025.
134. M. Mitchell, Debates on the nature of artificial general intelligence. *Science* **383**, eado7069 (2024).
135. OpenAI, Learning to reason with LLMs. *OpenAI Blog* (2024). https://openai.com/index/learning-to-reason-with-llms/. Accessed 29 April 2025.
136. R. Mahari, P. Pataranutaporn, We need to prepare for 'addictive intelligence'. *MIT Technology Review*, 5 August 2024.
137. D. Hendrycks *et al.*, Measuring massive multitask language understanding. arXiv [Preprint] (2021). https://arxiv.org/abs/2009.03300 (Accessed 29 April 2025).
138. A. R. Aarohi Srivastava *et al.*, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2206.04615 (Accessed 29 April 2025).
139. M. Suzgun *et al.*, Challenging big-bench tasks and whether chain-of-thought can solve them. arXiv [Preprint] (2022). https://arxiv.org/abs/2210.09261 (Accessed 29 April 2025).
140. W. Zhong et al., AGIEval: A human-centric benchmark for evaluating foundation models. arXiv [Preprint] (2023). https://arxiv.org/abs/2304.06364 (Accessed 29 April 2025).
141. S. Honda, Benchmarking large language models. *Medium* (2024). https://medium.com/alan/benchmarking-large-language-models-1e1ab5b809ac. Accessed 29 April 2025.
142. A. S. Raamkumar, S. B. Loh, Towards a multidimensional evaluation framework for empathetic conversational systems. arXiv [Preprint] (2024). https://arxiv.org/abs/2407.18538 (Accessed 29 April 2025).
143. L. Wang *et al.*, GIEBench: Towards holistic evaluation of group identity-based empathy for large language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2406.14903 (Accessed 29 April 2025).
144. T. Aas, AI-driven role-play training and the agentic AI: Bringing virtual humans to the workplace. *Training Industry* (2024). https://trainingindustry.com/articles/learning-technologies/ai-driven-role-play-training-and-the-agentic-ai-bringing-virtual-humans-to-the-workplace/. Accessed 29 April 2025.
145. D. Tuggener *et al.*, "Role-playing LLMs in professional communication training: The case of investigative interviews with children" in *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)* (Association for Computational Linguistics, Vienna, Austria, 2024), pp. 249–263.
146. G. M. Lucas *et al.*, "Reporting mental health symptoms: Breaking down barriers to care with virtual human interviewers" in *The Impact of Virtual and Augmented Reality on Individuals and Society* (Frontiers Media SA, 2019), pp. 256–264.
147. A. S. Rizzo, R. Shilling, Clinical virtual reality tools to advance the prevention, assessment, and treatment of PTSD. *Eur. J. Psychotraumatol.* **8**, 1414560 (2018).
148. B. Maples, M. Cerit, A. Vishwanath, R. Pea, Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Res.* **3**, 4 (2024).
149. D. Weijers, N. Munn, AI companions can relieve loneliness–But here are 4 red flags to watch for in your chatbot 'friend'. *The Conversation* (2024). https://theconversation.com/ai-companions-can-relieve-loneliness-but-here-are-4-red-flags-to-watch-for-in-your-chatbot-friend-227338. Accessed 29 April 2025.
150. K. Chayka, You're a.i. companion will support you no matter what. *The New Yorker* (2024). https://www.newyorker.com/culture/infinite-scroll/your-ai-companion-will-support-you-no-matter-what. Accessed 29 April 2025.
151. N. Tiku, AI friendships claim to cure loneliness. Some are ending in suicide. *The Washington Post* (2024). https://www.washingtonpost.com/technology/2024/12/06/ai-companion-chai-research-character-ai/. Accessed 29 April 2025.
152. P. Verma, They fell in love with AI bots. A software update broke their hearts. *The Washington Post* (2023). https://www.washingtonpost.com/technology/2023/03/30/replika-ai-chatbot-update/. Accessed 29 April 2025.
153. T. H. Costello, G. Pennycook, D. G. Rand, Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, 1143–1147 (2024).
154. R. Staab, M. Vero, M. Balunović, M. Vechev, Beyond memorization: Violating privacy via inference with large language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2310.07298 (Accessed 29 April 2025).
155. J. Hazell, Spear phishing with large language models. *Governance AI* (2023). https://www.governance.ai/research-paper/llms-used-spear-phishing. Accessed 29 April 2025.
156. D. A. Gentile, "The rating systems for media products" in *Handbook of Children, Media, and Development*, S. L. Calvert, B. J. Wilson. Eds. (Blackwell Publishing, Malden, MA, 2008), pp. 527–551.
157. K. Mackrael, S. Schechner, European lawmakers pass AI act, world's first comprehensive AI law. *The Wall Street Journal* (2024). https://www.wsj.com/tech/ai/ai-act-passes-european-union-law-regulation-e04ec251. Accessed 29 April 2025.
158. W. G. Shadel *et al.*, Do graphic health warning labels on cigarette packages deter purchases at point-of-sale? An experiment with adult smokers. *Heal. Educ. Res.* **34**, 321–331 (2019).
159. C. Martel, D. G. Rand, Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Curr. Opin. Psychol.* **54**, 101710 (2023).
160. J. Williams, Should AI always identify itself? It's more complicated than you might think. *Electronic Fronteir Foundation's Deeplinks Blog* (2018). https://www.eff.org/deeplinks/2018/05/should-ai-always-identify-itself-its-more-complicated-you-might-think. Accessed 29 April 2025.