Harvard Data Science Review • Issue 7.2, Spring 2025

One Person Dialogues: Concerns About AI-Human Interactions

Darren Frey[1] **Daniel H. Weiss**[2]

[1]Paris Institute of Political Studies: Sciences Po, Paris, France,
[2]University of Cambridge, Cambridge, England, United Kingdom

The MIT Press

Published on: May 28, 2025

DOI: https://doi.org/10.1162/99608f92.01674a29

ABSTRACT

Potential disruptions to economic, educational, and political affairs have remained at the fore of conversations about the implications of large language models (LLMs); however, remarkably little attention has been paid to the potentially more immediate ethical, psychological, and sociological repercussions of these and similar technologies. In the following, the authors develop a number of concerns about sustained LLM-human interaction by contrasting these with ordinary conversational and social contexts. The foremost among these are ethical, especially potential losses of empathetic capabilities, but the authors note a number of possible related linguistic, behavioral, and cognitive consequences. This work is intended to stimulate further research into these questions and concludes by offering suggestions for related empirical and theoretical analyses.

Keywords: artificial intelligence, ethics, human-computer interaction, cognitive science, behavioral science, dialogue

---

Media Summary

Two years after the first popularly available releases, generative large language models like ChatGPT continue to capture the attention and imagination of the public, from policymakers struggling to come to terms with some of their challenges to marketers and managers eager to cut corners. The technology's potential dangers have remained a part of the popular conversation, though this has tended to focus on doomsday scenarios, like the emergence of a Terminator-style superintelligence, or consequences of its misuse, including concerns about privacy, cybersecurity, and education. However, remarkably little attention has been given to the potential repercussions of using the technology exactly as it was intended to be used. For example, what do we know about the impact that talking with digital agents (like ChatGPT) for a long time might have on the way we talk with other human beings? Might regular interaction with such agents alter the way we think of ourselves and treat others? Although these and similar other questions have yet to be answered, we believe that there are significant potential concerns that deserve to be investigated more directly than they have been to date. By focusing on the differences between how individuals interact with digital agents and how they interact (or should interact) with others, we point to a handful of potential concerns, including a possible loss of empathy or a diminution of our capabilities to listen openly and thoughtfully to others. We conclude by examining the implications of these differences and suggesting future experimental and theoretical research that could help further illuminate these matters.

1. Introduction[1]

Older generations call those who make these objects miracle-makers on account of the stunning nature of the public spectacle.

—Hero of Alexandria, *On Automaton-Making*

The above quote is drawn from a text written 2 millennia ago. It illustrates the use and engineering of a number of machines that were meant to operate largely on their own, often in religious contexts, and were used to inspire awe. Regarding a similar use in the Roman world 3 centuries later, David Bentley Hart (2015) wrote:

> Many temples were specially designed to "assist" the faithful in seeing or hearing the god in whom they had placed their trust. Mechanical devices, optical tricks and combustible chemicals were used to simulate miracles and divine visitations. To give the impression that an idol had been inhabited by a divine spirit and brought to life, a clockwork automaton would be used; a hidden speaking trumpet would produce the voice of an unseen god. (Hart, 2015, p. 64)

Hart goes on, cataloging several other devices used in similar contexts. His focus remains fixed on the instruments themselves and the "tricksters" (Hart, 2015, p. 63) who had made them, just as Hero of Alexandria two millennia ago fixated on the machines and the reception of the 'miracle-makers' who built them. The history of the emergence of these technologies makes for fascinating reading, but it too often neglects a crucial perspective: the point of view of those who used the machines. While attending to the way in which, say, a priestly class exerted its power, historians often elide the way the lives of individuals within these communities were shaped by their encounters with these awe-inspiring machines.

In recent years, the pace of technological development has accelerated such that advances that seemed nearly unimaginable two decades ago are increasingly mundane. Many of these innovations are intentionally structured to mimic, aid, or expand human thought, and their progress has been truly astounding. Whether or not one agrees that ChatGPT broke the Turing test and that new means of assessing AI are necessary, the stream of recent success stories and increasing capabilities has occasioned the reevaluation of broad, popular positions toward technology in general and AI in particular (Biever, 2023).

Large language models (LLMs) are expansions of previous work on neural networks, generally transformers, that build probabilistic models based on large corpora to anticipate associations between segments of texts. Natural language processing researchers (and before that, computationally oriented linguists) have developed predictive language models for decades, but this most recent batch has proved increasingly capable at language generation tasks. In particular, the ability of LLMs to have convincing, sustained, and natural-sounding conversations with their users, often mirroring subtle social and linguistic mannerisms, differentiates them from previous AI in ways that merit special attention.

More specifically, these technologies are so new that we know very little about how they impact our own reasoning and decision-making, and we know even less about how they might impact us interpersonally. Most

ethical reflections on these developments has focused on real and genuine threats, but it tends to primarily address either: (a) technical concerns owing to flaws in the development and production of the models or (b) their intentional misuse. Chief among these are concerns about the proliferation of misinformation, biases in training data sets, and countless intentionally nefarious uses of the technology, including cyber warfare and the like. Yet there is scant engagement with what it means for a significant portion of our imaginative, conversational, and social lives to develop in this new context.[2] In other words, what are the likely psychological, interpersonal, and ethical consequences of using this technology precisely as it is meant to be used?[3]

To address these questions, we begin by sketching a perspective of LLMs that we hope will be behaviorally, ethically, and morally more productive than purely pragmatic or technical discussions. The pragmatic conclusion that technology is a simple tool, neither good nor bad in itself, is so common that it seems like a cultural reflex, an automatic response to the question. However, it obscures reflection and ignores the way we come to be shaped and habituated by our use of the objects in our environments. For a very thorough treatment of this argument and something that anticipates the approach taken here, see Albert Borgmann's (1987) *Technology and the Character of Contemporary Life*. Additional important precursors to our analysis here are Sherry Turkle's (2012) *Alone Together: Why We Expect More From Technology and Less From Each Other* and her (2017) "Empathy Machines: Forgetting the Body." While Turkle focuses less on the question of verbal conversational exchange, her analysis of the risks of human interaction with other types of AI-based "relational artifacts" (Turkle 2012, passim) raises concerns similar to those discussed here.

To anticipate, we aim to characterize aspects of LLM use that have a direct bearing on communicative and social ethics, thus reflecting on the distinctiveness of the sorts of interactions individuals have, and are increasingly likely to have, with digital agents like LLMs. Although we draw on recent experimental findings to support many of our claims, the purpose of this article is to provide a robust theoretical motivation for future research programs, not to provide new empirical results.

We should additionally note that the arguments we make are not particular to the specific, technical structure of LLMs, but can apply to any present or future digital agent capable of engaging in humanlike conversational exchange to the same degree as LLMs. That is, our focus is not on the underlying computational architecture, but on the practical, interactive experience of engaging with humanlike digital agents on a sustained basis, particularly in conversational contexts. Thus, we frame most of our reflections in terms of exchanges with humanlike dialogue agents (HDAs), a class of agent that includes LLMs and similar or potentially more sophisticated technologies. An HDA is any nonhuman agent capable of conversational exchange at least as sophisticated as most popular LLMs. While a competency-based definition or a set of necessary conditions would be preferable, for the present purpose it is enough to imagine a sufficient condition: any agent that one might reasonably suspect would pass a fairly robust Turing test is an HDA.[4]

Our approach will be to first describe features of conversational interactions with HDAs, without considering normative issues or concerns. We will then evaluate normative and other concerns specifically in light of the descriptive reflections. In doing so, we aim to illustrate considerations associated with LLMs that might not be apparent otherwise, especially crucial deviations from normal and ideal ways of conversing.[5] We will conclude with a set of suggestions for future inquiry, research programs that entail both empirical analyses of cognitive and behavioral phenomena as well as bodies of ethical literature that could be instructive as we attempt to develop safe ways of working with these new technologies. To introduce these discussions, we begin with a brief review of some related research.

2. Relevant Existing Research

Previous studies in the field of human–computer interaction provide an empirical basis for concerns that long-term, habitual interaction with HDAs could impact the way that people interact with other human beings. Prominently, for example, the landmark study by Reeves and Nass (1996) presents a series of experiments whose findings point to key ways in which people respond and relate to 'media' technologies, finding that many of these interactions mirror social ones (Reeves & Nass, 1996). Essentially, their method was to run iconic psychological studies, replacing certain subjects in the classical experiments with computers (and other technologies). They find, among many other things, that people: (a) rate computer interactions more positively if the computer itself has framed information positively (Nass et al., 1999); (b) identified more with computers and found them more agreeable if explicitly teamed with them (Nass et al., 1996); and (c) tended to be more cooperative with a computer if prompted with a reciprocal exchange (Fogg & Nass, 1997), a finding that evokes Robert Cialdini's classic work on persuasion (2006).

They describe the results of their research in this way: "[H]uman responses to media are determined by the rules that apply to social relationships and navigating the world. Responses to media are not primarily governed by rules about how to use appliances more akin to a hammer or a car" (Reeves & Nass, 1996, p. 10).[6] Reeves and Nass support their conclusions by making reference to how humans evolved in rich social environments, likely developing heuristics to respond to most anything that exhibits social behaviors as if it were a person (Reeves & Nass, 1996, p. 12).[7] Moreover, they emphasize that these sorts of responses to media occur without people being conscious of them; in fact, it seems to happen even if individuals *consciously believe* that it would be unreasonable to treat technology like a human (Reeves & Nass, 1996, pp. 186–189). On their account, there is not a sharp divide between the ways in which people relate to media-based technology and the ways in which people relate to other human beings.

In the nearly 3 decades since their first findings, the media equation has been extended and criticized in a number of ways. For example, researchers have deployed it to examine how individuals interact with more recent technologies, including evaluating politeness in smartphones use (Carolus et al., 2018), autonomic responses to humanlike robots (Reuten et al., 2018), and prosocial behaviors toward avatars and agents (Felnhofer et al., 2018). Various other researchers have argued that, instead of treating computers (and other

media technologies) as social agents, people often treat them as envoys of their programmers (Souza, 2005) or instantiations of specific social-like situations, but not fully identical to actual people.[8]

While the studies by Reeves and Nass (1996) focus primarily on how people's habits of interacting with other human beings are 'carried over' to their modes of interaction with technology, this itself stems from a cultural context in which the majority of 'person-like' interactions take place with other human beings. By contrast, the rise and prominence of HDAs also raise the possibility that the habituation can also go in the other direction. That is, given that there is not a sharp divide between the two, then it seems probable that a greater amount of time spent engaged in person-like conversation with HDAs can shape conversational habits that can then carry over to personal interaction with other human beings.

Although not explicitly focused on reciprocal conversational exchange, Turkle highlights ways in which interactions with 'human-seeming' media can indeed have significant effects that reshape people's modes of interaction with other human beings (Turkle, 2012). In addition, even beyond the specific features of media technologies, there is a compelling, if often overlooked, line of argument that sustained interaction with machines in a more general sense can itself lead to the dehumanization of others, as perhaps best explored in Borgmann (1987).

This last point also provides an opportunity for us to clarify the extent to which our analysis of HDAs might or might not apply to interactions with previous forms of technology. As our basic point of comparison in this study, we will be focusing on differences between a human being interacting with another human being in an in-person setting, and a human being interacting with an HDA. At the same time, many of the characteristics of HDA–human interactions that we will discuss are representative of relations with many other technologies. For example, when an individual interacts with a microwave oven, they do so *instrumentally*. They are accustomed to the microwave being always available, immediately responsive, compliant, nonjudgmental, not emotionally sensitive, and so on—in contrast to interaction with another human being. As such, many of the characteristics of interactions with HDAs that we will illustrate are not new in themselves, but their introduction into the context of habitual human conversation marks a significant departure from relationships with earlier technologies. Features of interactions with technology that might be less problematic in other settings might well be especially risky when they emerge in the distinctively and intimately human context of *verbal conversational exchange*, a domain that had previously been exclusively interhuman.[9] Interacting with HDAs is characterized by *humanlike dialogue*, by conversational exchange in natural language in a back-and-forth manner, which has not generally been the case with previous technology. The introduction of instrumentalizing structures to such conversational contexts gives rise to distinctive concerns that are less relevant to nonconversational interactions with technological devices.

This aspect also distinguishes HDAs from previous computer-based technologies. While the use of Internet search engines, for example, involves various automating features that can contribute to cognitive effects as a result of sustained usage (as we also address below), such search engines, while constituting a certain type of

'dialogue,' have not involved conversational exchange akin to that of natural human dialogue. Therefore, the dynamics of HDA–human interaction are likely to be different in significant ways from engagements with previous technologies. The same applies to earlier dialogue agents or chatbots, from Weizenbaum's ELIZA (1966) onward. Namely, while these have *attempted* to simulate humanlike dialogue, they were not generally successful in actually engaging in conversation with users in a sufficiently natural, contextually dependent manner. By contrast, our analysis of HDAs focuses on the effects of combining technological instrumentality with *successfully natural* conversational interaction, in a manner that would pass the Turing test. While various of the dynamics that we analyze could apply to earlier chatbots to a more rudimentary extent, previous users would have experienced those earlier dialogue agents as more artificial and stilted, and thus as more clearly nonhuman, with concomitant impact on the ways in which users perceive and treat such agents. By contrast, it is the *humanlike dialogue* capabilities of LLMs that differentiate them, and HDAs as a whole, from previous chatbots, dialogue systems, and the like. This feature means that, experientially, users encounter HDAs much more as a 'who' rather than as a mere 'what.'

Finally, we should note the way in which our analysis of HDA–human interaction relates to human–human interactions that are not in-person, but mediated by technology. For instance, communication between two individuals by text messaging differs in various ways from in-person conversation (e.g., lack of visual conversational cues, tone of voice, body language), and so could share some of the aspects we underscore about HDA–human interaction.[10] Yet, for example, even over text messaging, one needs to be attuned to not hurting the other person's feelings, and one does not automatically receive an immediate, compliant, or nonjudgmental response, in contrast to HDA–human interaction, as we show. In these and other respects illustrated below, our analysis of HDAs departs meaningfully from text messaging between two human beings. Likewise, even in screen-mediated human–human conversations, the fact that one *knows and thinks of* the other person as having a body, and so on (particularly if one has previously met the other person in person) may moderate at least some of the disembodied dynamics linked to the lack of visual or auditory experience discussed below.[11]

Thus, although not mutually exclusive, concerns regarding HDAs and their use differ in important ways from concerns about previous forms of technology and technologically mediated conversations. We highlight these distinctive aspects in order to get a sense of what is most 'new' about LLMs (and other future HDAs), so that empirical studies can be successfully designed to measure and evaluate these potential effects. It is with these concerns in mind that we now turn to a descriptive characterization of HDA interactions.

3. Characteristics of HDA Interactions

3.1. What HDA–Human Interactions Do Entail

The following section is intended to set forth a handful of observations about the ways in which individuals interact with HDAs in order to later motivate a number of behavioral, moral, and ethical reflections in the

sections that follow. These observations are, on the whole, neither controversial nor surprising. However, by characterizing what interactions with HDAs do and do not entail, we hope to shed light on concerns that might get overlooked in other contexts. Beginning with the former, HDAs are: (a) always available, (b) immediately and necessarily responsive to the user, who both initiates and ends the conversation, and they are (c) generally compliant, unless explicitly engineered otherwise.

Unlike human conversation partners, HDAs are essentially always available to an individual. Barring technical downtime or absence of a connection to a device (circumstances that are increasingly rare), one can readily summon the HDA. The practical and conventional barriers to ordinary conversation are replaced with constant availability. For example, the HDA is simply always ready at hand, unlike ordinary conversation partners who are often otherwise occupied or unavailable. Similarly, unlike conversations with neighbors, colleagues, or friends, who one might reasonably avoid contacting, say, in the middle of the night, no such restraint is necessary with an HDA.

Not only is the HDA always available, it must always respond. In this sense, it has no agency at all. Its initial response might be overly generic, hallucinatory, or merely the repetition of a hard-coded disclaimer; however, unlike addressing a human, who can always choose to simply walk away from a conversation, the HDA necessarily responds to the user. Furthermore, that response is nearly instantaneous. Again, unlike ordinary human conversation, in which a dialogue partner might choose to defer responding or take some time to deliberate over her response, the HDA's response is immediate.

There is another, broader structural feature of interactions with HDAs that meaningfully departs from ordinary conversational pragmatics. Discussions with such digital agents are always user-initiated and user-ended. The user accessing the HDA decides when to begin a conversation and decides when it is over. In other words, the user exercises absolute control over the very existence of the interaction. Relatedly, the HDA is generally compliant, except when specifically programmed otherwise. The digital agent is, in this way, a clear subordinate, a tool that serves merely as an extension of the pursuit of the user's own ends. Having briefly considered prominent features of interactions with HDAs, it is now worth examining the contrary.

3.2. What HDA–Human Interactions Do Not Entail

Although interactions with large language models vary, they are generally consistent in that they are not: (a) embodied, (b) emotionally or psychologically sensitive, (c) linked to a specific context or person, nor (d) apparently agenda-prone or judgmental.

In most ordinary (nontechnologically facilitated) conversation, individuals take cues from one another that relate to their specific location in space and time. Apart from intentional nonverbal communication, people adjust their speech to one another given a plurality of physically situating factors. For example, individuals tend to mimic one another's postures or tones rather spontaneously. They also compensate or adjust their speech to the others' physical situation, as is the case when a conversation partner is seemingly hurried,

distracted, or relaxed. More generally, individuals attend to—and are expected to attend to—their conversation partner's specificities, and these are usually embodied.[12]

Similarly, various other situating and personal factors are embodied and typically shape the way in which a conversation with another person takes place. For instance, one's conversation with another might be shaped by questions such as: Is this person older than me, or younger than me? Is this person of the opposite sex? Do they have a different accent from mine? While these various aspects of embodied particularity can in some contexts be linked to prejudice or stereotyping, they are also bound up with aspects of ethical sensitivity: one should take care to shape the conversation in ways that are most appropriate to the particular person with whom one is conversing, and awareness of these embodied features is often crucial for enabling a person to adjust their interactions accordingly.

In the vast majority of ordinary conversational settings, individuals approach one another in specific contexts, and these contexts generally shape the way a conversation unfolds. Consider the contrast, for example, between a quick phone call to a close friend and a hiring interview. Formality, tone, diction, and countless other features of the conversation are context-dependent. Although there is a specific context within which HDA–human conversations take place, these are all subject to the user's determination, as in the case of 'prompt engineering.' It would be just as appropriate to address the HDA informally as formally, in a playful vernacular or in Victorian prose—one would not be addressing the HDA 'inappropriately' either way. In any event, the sorts of adaptations that are crucial to everyday conversation, like adjusting one's self-presentation to the conversational other, are not necessary with an HDA.

Relatedly, one does not consider the feelings of the HDA, nor does one expect the HDA to do the same to the user. A normal part of human development is coming to understand subtle conversational relationships between what one wants to say or what one has to say, and what is appropriate, given the conversation partner's specific feelings. If one's partner is crying, it would be curious, if not inappropriate, to ask them about the weather. As part of socialization is learning how to engage in conversation with other people, we learn to restrain certain conversation impulses and to attend to how what we say might affect the other person emotionally, in a given moment and context.

Likewise, conversational interaction with another person is shaped by our awareness of our conversation partner's previous experiences and the specifics of their present situation. One would typically interact differently with a person who has recently lost a parent, or who grew up in economic conditions radically dissimilar from the speaker, or who enjoys (or does not enjoy) certain types of music, and so on. These different 'life features' make one individual different from another, and, along with feelings and embodiment, combine to constitute each individual as a unique person. Our conversations are structured around persons, and they tend to respond to, and reflect, the full diversity and variety of human experience.

By contrast, interaction with an HDA does not involve an expectation of attunement to the emotions of the addressee, since the HDA is not assumed to have any feelings to be hurt, one does not need to worry about the HDA being insulted by what one says. Thus, there is no need for typical forms of conversational restraint in such interactions. Similarly, in any given conversation with an HDA, there is no need to bear in mind thoughts like, 'This is a unique person who is different from others, and I need to make sure that I am giving due acknowledgment to that.' Rather, an HDA is neither assumed to mind nor notice if addressed in an impersonal or depersonalizing manner, and so there is no need (and also no basis) for one's forms of address to be sensitive in those regards.

Finally, one seemingly attractive feature of HDA–human interactions is that they do not involve judgment or social repercussions in the way most other conversations do. Since one is not addressing a human when talking with an HDA, one can, for example, readily confess ignorance of something one ought to know without fear of being judged. Likewise, one need not worry in such interactions about being judged as rude, annoying, inarticulate, and so on, and one also does not need to worry about one's conversation partner conveying such judgments to other people. Additionally, conversations with HDAs are not contingent on getting to know one another, as is the case in ordinary conversation. Although one might experiment with the HDA to get a sense of the contexts in which it excels, the ordinary process of 'getting to know' the HDA is highly structured, stylized, and user-centered. Abstractly, these tendencies seem consistent with human interactions with other, earlier technologies, like a microwave oven; however, distinctive problems arise, as we will now discuss, when these tendencies emerge in the sphere of conversational exchange.

4. Concerns Linked to Consistent, Long-Term HDA–Human Interactions

The above characterizations are, of course, not the full story. Aspects of the evolution of HDA–human interactions are necessarily speculative at points, yet the previous are relatively safe points from which to begin reflection, as they are constrained to considerations that are distinctive to ways in which individuals *already seem to be interacting* with LLMs.[13] In the following three sections, we consider three sorts of concerns that relate to these observations, beginning with linguistic consequences of prolonged HDA use, then turning to moral and ethical concerns, and concluding with broader cognitive considerations.[14]

4.1. Linguistic and Conversational

Among other linguistic consequences, regular and habitual conversation[15] with a digital agent might impact one's ability to listen actively, to attend to embodied conversational cues, and to appropriately differentiate the distinctiveness of individuals. We will consider each of these in turn.

Human conversations entail countless intricacies that have been classified in many ways, but there are certain features that seem to be consistent across most classifications and cultures. Chief among these is some notion of conversational reciprocity. In certain contexts, reciprocity forms the backbone of an entire linguistic theory;

in others, it serves merely as a limiting or constraining factor, often the consequence of specific social norms or conversational constructs.[16] For the present purpose, we will define conversational reciprocity as a shared, active attention to the needs, desires, and cognitive states of others. While conversations involve differing levels of reciprocity, depending on the context and participants, the desire to be recognized and understood within conversation seems fairly universal.

Part of what enables mutual recognition is constantly attending to the other in their own context. As contexts evolve, dialogues often evolve with them, and such transformations span from the mundane to the extraordinary. For example, consider how the term 'here' changes as the subject changes location, perhaps gesturing in space at different points. More compellingly, contrast, for example, a partner saying to you 'I have a problem' with tears in her eyes with a partner saying to you 'I have a problem' after she has accidentally dropped a pen in a hard-to-reach area.

One could argue that 'disembodied' conversations have been exerting influence on humans since the invention of the telephone, if not before. However, in the case of conversations with other humans, we often still actively imagine their locations or bodies. One of the foremost questions to consider moving forward is whether sustained interaction with HDAs leads to a diminished awareness of the embodiment of others and ourselves. Are we less likely to attend to subtle physical cues, like microexpressions? Might we come to ignore the often useful 'gut feelings' that help us navigate conversations? More practically, could such changes be registered on existing tasks, like the reading the mind in the eyes task?[17]

Although the foregoing might seem to suggest that the primary communicative concerns with digital conversation partners involve their (lack of) physicality, the lack of distinctive personhood of such partners is, though subtler, no less of a worry. Regardless of how cleverly engineered, digital conversation partners remain constructs with no distinctive personhood, no individual personalities, and our interactions with them are, in a certain sense, entirely predictable in a way conversations with humans are not.[18] Though it seems sensible to interrogate the pragmatic differences of such agents—evaluating, for example, in what ways a given model outperforms others—it would be unreasonable and inappropriate to spend a great deal of time and energy getting to know what an agent prefers or desires. Yet, in ordinary conversational contexts, the ability to recognize the distinctiveness of individuals in this way facilitates conversation and serves countless other social and personal ends. These reflections prompt an additional, and considerably more difficult, question: by regularly engaging digital agents that lack these qualities of embodiment and distinctive personhood, and that therefore break the previous habitual connection between conversational exchange and sensitivity to those aspects, will we come to be impoverished listeners, less capable of actively engaging our human conversation partners?

4.2. Moral and Ethical

The initial characterizations of HDA–human interactions directly entail the most prominent moral and ethical dilemmas resulting from them, many of which have been foreshadowed in the discussion of conversations. The structure of the concern, in its most general form, is this: that by regularly assuming the sorts of conversational perspectives associated with HDA–human interactions, individuals might come to treat others, and perhaps themselves, as mere objects to be acted upon. At its most extreme, this could promote or normalize analogues of master–slave relationships; more immediately, it could habituate users to one-sided conversations and the many ethical and moral shortcomings that result from them.

The posture one assumes toward a digital assistant is instrumental, and one naturally treats the HDA as a means to one's own end. The relevant characteristics of this interaction have been illustrated above: the HDA is an object (not a person with their own individual needs and desires) that is always immediately available to execute the user's wishes, an agent perhaps best characterized by its compliance, about which one need not (perhaps should not) care. It is precisely in these ways that the interaction evokes master-slave dynamics. Put bluntly, does habituation to such dynamics of one-sided conversational control and a lack of attention to the conversation partner's individuality and agency run the risk of accustoming or habituating a person to similar modes of interaction in intrahuman conversation and interaction?

Some might claim that these sorts of concerns are not relevant to discussions of HDAs, as individuals have throughout history regularly assumed objectifying and instrumental postures to technology. On this line of thought, one could potentially argue that merely interacting with a machine does not automatically result in increases in social harms. Individuals have interacted with machines for centuries now, without this causing an obvious increase in societal violence or oppression on the interhuman level, and HDAs are just another variety of machine.

Setting aside the assumption that previous historical interactions with technology have not been linked with increases in societal violence and oppression—an assumption directly challenged by Borgmann and others—there are two relevant responses in this case, one clarification, one amplification. First, LLMs, and otherwise architected digital agents, are not simple machines, as previously discussed, and there is no reason to assume that the consequences of human interaction with them will resemble the consequences of interacting with machines more generally, in relation to which one does not sustain dialogue and conversational exchange. Because speech and conversation are so closely bound to core aspects of human cognition, personality, and social relations, the potential social effects need to be assessed carefully, without assuming that these would be equivalent to the effects of operating an automobile or a microwave oven.[19] Attending to the ways in which HDAs are importantly different from previous machines is precisely what is required at this point in the history of technology.

Second, while the extent and variety of HDA–human interactions might vary dramatically, producing diverse social and moral consequences, the 'spilling over' to human others *of any single facet* of the interactions

illustrated above would already be problematic. In other words, there are a limitless range of possible concerns at work here short of the actual reintroduction of slavery that merit attention. The pertinent question becomes: to what extent are these dehumanizing dynamics likely or predictable outcomes, given how individuals engage in HDA consumption?

While there are any number of such dynamics that one might address, one of the foremost is the threat of increasing one-sided conversations. Though this is a concern that could have been addressed in the previous section, it is, we believe, such a profound threat to moral and ethical reasoning that it merits special attention here.

Recall that HDAs are always responsive, that they never initiate conversations, and that the conversational *structure*, and, as consequence, much of the *content* of the conversation, is user-directed. The possible world under consideration now is one in which humans spend countless hours initiating conversations with artificial dialogue partners, growing increasingly accustomed to being able to shape and guide the outcome of these conversations. That this could threaten one's ability to be a responsive, empathetic, or compassionate listener seems undeniable.

In addition, such dynamics could also be self-reinforcing. For various people, it could easily feel 'nice' to be in control in conversation, to have a conversation partner who is 'always interested in me,' and not to have to engage in efforts of attentiveness to the needs or feelings of the other person. Prior to LLMs, it was not possible to have the 'benefits' of conversational exchange without also putting in the work of attentiveness to the other—particularly because, if one did not do so, the other person might very well react negatively, and decide not to engage in further conversation! Now, however, one can have the benefits of conversation without such 'work,' and so this could easily seem more attractive. In turn, habituating oneself to this type of conversation could cause one's abilities to atrophy further, leading one to be even more disinclined to conversations involving such effort, leading to further atrophy, and so on.[20]

Alongside concerns that conversations with HDAs could lead to habits of being one-sided, controlling, or a bad listener, there is a more fundamental concern that conversations with HDAs could impact one's attunement to the particularity of one's conversation partner as a person, which is also a key element of ideal conversation. In conversation with a specific person, there is an ethical obligation not to treat them as a 'faceless abstraction,' but to be attuned to that person in their individuality. One knows what it feels like to be addressed in a way that treats one in an abstractly impersonal or 'objectifying' manner, and how this differs from being addressed in a way that is attuned to oneself as a distinctive individual subject, with particular feelings, life-experiences, and embodied needs and vulnerabilities. Accordingly, one tries to bear this in mind and be attuned to each conversation partner as a distinct individual. In this regard, the distinctive visual recognizability of each person's face can function as a reminder to recognize and uphold the other's distinctiveness in ethical relation as well, and to relate to the other as a personal 'you,' rather than as a depersonalized and dehumanized 'it.'

By contrast, in conversation with an HDA, one is not in a position to be attuned to the particularity and distinctive personality of your partner, since the HDA does not possess such qualities in the first place. In conversation with another human being, it is possible to neglect one's duties of relating to one's conversation partner as a particular individual, but it is also possible to enact one's duties of care and recognition. In HDA–human conversation, due to the structural lack of a particular and distinctive personality, one can neither 'fulfill' nor 'fail to fulfill' such responsibility. Prior to the development of HDAs, every instance of conversational exchange was bound up with the need to be attuned to the particularity of the other person; with HDAs, the connection between conversation and the need for such attunement is broken, and so one can become more habituated to conversation that lacks that sensitivity and attunement, which may then carry over to one's interactions with other human beings.

Moreover, when an HDA engages in conversational exchange with you, it does not treat *you* as a unique person. In ideal conversation, personal recognition is most often a mutual phenomenon, in which each person strives to understand the other as a person, and each person's ability to do so is bound up with the recognition that the other person also has their own particular feelings, life experiences, and embodied distinctiveness, just as I do. Since an HDA lacks these features, it is not in a position to relate to me on the basis of that mutual and empathetic recognition, and thus, in such conversations, I must adapt to conversational interaction in which I am not personally understood or recognized in my embodied particularity. While one would normally experience such interaction as disconcerting and problematic, regular conversation with HDAs over time can lead to a situation in which I become 'less bothered' by this type of depersonalized engagement.[21] Instilling such a habit can be detrimental, since being properly aware of dehumanizing or depersonalizing treatment is important for being able to critically resist individuals or institutions who operate in this manner. In addition, if I become accustomed to being treated in a depersonalizing or dehumanizing manner, this in turn may also lessen my sensitivity to reacting negatively if I observe someone else being treated by others in such a manner. This sensitivity is crucial for upholding just and humane social structures and interpersonal relations, and its potential diminution is therefore a matter of serious concern.

There is ongoing debate about whether, or to what extent, LLMs can be empathetic agents. See, for example, the work of Cuadra et al. (2024) and Welivita and Pu (2024). Although we think there is a meaningful distinction to be drawn between *performing or simulating* and *having* empathy—and we believe that the former and not the latter characterizes HDAs supposed empathetic faculties—such an argument is beyond the scope of the present article. Our concern is the experience and consequence of interacting with such agents (see also Turkle, 2017). If one experiences the HDA as an empathetic-seeming agent, and yet is not expected to respond in kind—as is normally the case in human empathetic contexts—this could have consequences on empathetic faculties in normal human relations. In other words, one might come to expect empathy from conversation partners without feeling a need to reciprocate. Conversely, if one does not experience the HDA as an actually empathetic agent, then one might become habituated to not being treated empathetically by others. While much of our discussion points to the ways in which interaction with HDAs does not require the human

user to exercise habits of empathy *toward* the conversational partner, it is also the case that interactions with HDAs may habituate users not to expect empathetic treatment *from* a conversational partner.

4.3. Cognitive and Behavioral

Though linguistic and communicative concerns are, in a way, clearly cognitive issues, in the following section we conclude by briefly underscoring considerations that are typically addressed by cognitive and behavioral scientists. To anticipate, the discussion begins by recalling particularities of conversations with HDAs that might induce specific and novel concerns, then follows this reflection with considerations of potential impacts to convergent and divergent thinking—especially critical thinking capabilities and creativity, respectively—concluding with broader implications to dual-process approaches to cognition.

Unlike most normal conversational contexts, interactions with HDAs might quickly engender trust with any given user because of how they are structured. Precisely because users do *not* view them as human beings, these agents are often perceived as especially intelligent, perhaps even wise; in addition, their nonhuman status could more readily lead a user to perceive them as nonjudgmental and as having no apparent agenda. For these reasons, a user could readily slip into an especially trusting mode of conversational interaction with the HDA.[22]

By contrast, in interhuman relations, trusting conversations of this sort are generally the purview of close intimate relationships or other contexts specifically designed to ensure psychological safety (religious/confessional circumstances, therapeutic environments, etc.). One normally engages in a 'trusting' way specifically with conversation partners that one knows well, with whom one has had concrete reasons (previous experiences with that person over time, or professional recommendation by someone else trustworthy) that lead one to judge the other to be trustworthy. On the contrary, interaction with a 'random' or otherwise unknown person would usually be accompanied by awareness that the other person, as a human being with potentially selfish interests, might not be trustworthy. Thus, among other concerns, one might wonder whether becoming habituated to conversations with an agent one comes to trust quickly—without that trust arising from actually getting to know the person over time—could threaten the normal, protective awareness and faculties that enable one to negotiate environments with mixtures of friendly and hostile elements. Yet, there are more direct threats to critical thinking capabilities.

Part of what characterizes an individual's capacity to think critically is his or her ability to attend to both explicit (logical structures, evidentiary support) and implicit features (credibility of source, etc.) of a position.[23] Apart from whether HDAs might come to serve as powerful confirmation bias machines, supplying users with endless, ready-at-hand arguments for whatever they *want to be the case*, there is a potentially more significant concern that regularly interacting with them could weaken one's metacognitive abilities.[24] If one habitually engages an agent that comes across experientially, in some sufficiently convincing manner, as a super-intelligence (e.g., capable of generating conclusions on complex topics nearly instantaneously), one might well come to apply less rigorous evaluative standards in this context than elsewhere. If a conversation partner comes

across as 'basically like me,' I would understand myself as more in a position to critically evaluate and assess what they say. By contrast, if a conversation partner consistently displays features that seem 'qualitatively different from me' in terms of certain abilities linked to intelligence, might people be more inclined to assume that such an interlocutor 'knows what they are talking about'? This poses a problem about thinking critically in conversation with the agent itself, but there are external concerns, too, to the extent that such a tendency (of 'giving up' one's critical agency) might become a habit that is generalized to other evaluative contexts.

There are similar concerns related to the cognitive science of creativity. While there is significant disagreement over how best to conceive of creativity, one compelling account suggests that the generation of novel, relevant ideas involves actively inhibiting nonoriginal ones. There is a modest amount of neuroscientific evidence (Cassotti et al., 2016) and a significant amount of behavioral research (Benedek et al., 2012; Cassotti et al., 2016) that supports this position. Assuming the inhibition model of creativity is correct in at least certain cases, then interacting with HDAs on a regular basis could threaten creativity in two ways: (1) it could undermine the users' abilities to generate nonoriginal ideas on their own, because in many cases idea generation could readily be outsourced to the model; and, thus, (2) potentially atrophying the mental musculature (i.e., inhibitory control) required to inhibit nonoriginal ideas.[25]

Investigating such questions essentially broadens and complexifies ongoing research programs evaluating how the availability of a number of contemporary technologies, especially ever-present broadband internet, impacts cognition. For example, there is a tremendous amount of research that analyzes the effect of Internet availability on individuals' memories, from semantic (Nagam, 2023) to autobiographical (Marsh & Rajaram, 2019). Regarding the former, there is empirical work suggesting that even the mere mention of online availability reduces individuals' tendencies to store certain information (Nagam, 2023). Researchers considering the impact of technology on cognition often frame this in terms of 'cognitive offloading' (Fisher et al., 2021), a notion that is in line with broader trends in cognitive science that emphasize the miserliness of our thinking and reasoning tendencies, from social cognition (Fiske & Taylor, 2013) to rational deliberation.[26]

Even in the relatively nondiscursive context of the Internet, individuals often overestimate how well they have learned something because of its availability (Fisher et al., 2021). In fact, when subjects are asked to search for answers to questions in a given domain, they report higher levels of confidence in their responses to questions in entirely unrelated domains, relative to a control group that did not use the Internet (Fisher et al., 2015). Might there be similar metacognitive overestimations in terms of learning and knowledge linked to increases in HDA use?

If some of the most persuasive conclusions in contemporary cognitive science are any indication, this is likely to be the case. Leonid Rozenblit and Frank Keil's groundbreaking work (2002) on the illusion of explanatory depth indicates that such overconfidence estimations are incredibly common. Steven Sloman and Philip Fernbach (2017) summarize—and expand on—many of the findings from the past few decades that indicate that individuals consistently overestimate how well they understand and can explain commonplace phenomena.

Keil, along with Fisher and Goddu (2015), has broadened his initial results to demonstrate that, at points, technology is immediately implicated in these overestimations.

Apart from specific concerns about how consistent use of HDAs might impact critical thinking, creativity, or metacognition, there are broader questions about how it might influence reasoning in general. For example, as suggested above, given our limited cognitive resources, there are often adaptive pressures to not store information that is quickly accessible elsewhere. Interestingly, much of the available research essentially depicts the Internet as a collective repository of information, a giant memory store, available for individuals to access as they please. While HDAs might be similarly construed, the sorts of exchanges users have with them seem more active and collaborative in that they are discursive in a way that Internet searches tend not to be. With HDAs, there is often considerable turn-taking, a back and forth that mirrors the way individuals often deliberate in isolation or in dialogue with other people. A natural extension of these earlier research programs is thus: if we can offload memory, why can we not offload reasoning altogether? Put differently, are there adaptive pressures that might encourage us to 'offload' parts of our reasoning to HDAs systematically? And what, if any, cognitive impact would this have on reasoners?

Furthermore, on most dual-process frameworks, intuitive judgments occasionally need to be overridden or corrected by slower, more deliberate forms of reasoning. While there are considerable individual differences in these monitoring processes, most reasoners seem to register reasoning conflicts (De Neys & Glumicic, 2008). However, does this sort of metacognitive monitoring come under threat if we are regularly conversing with a superintelligent agent that feels like it is doing this work for us? Might a nearly constant reinforcement of our intuitions by an apparent superintelligence lead us to be poorer at monitoring them?[27]

5. Conclusion

The above has developed a set of concerns related to the use of HDAs that were derived from considerations of what interactions with such agents do and do not entail, relative to ordinary conversational contexts. Throughout, the emphasis has been on underscoring potential threats that could be the consequence of sustained use shaping both personal and interpersonal habits, norms, and intuitions. Our basic approach in this analysis has been to describe features of ways individuals interact with HDAs in order to motivate concerns that follow from these fairly naturally. While many of these features apply to interactions with machines more generally, they become particularly problematic when they emerge in the sphere of conversational exchange. The latter was previously specifically connected to conversations with other persons, and was intimately bound up with (ideally) being interpersonally sensitive and aware.

Many of the concerns motivated in this article, especially the cognitive and behavioral ones, could be empirically tested by developing or repurposing experimental paradigms that investigate whether sustained interaction leads to the hypothesized deficits. Among the questions raised above, researchers could seek to determine whether there are losses in empathetic and interpersonal abilities, ranging from deficits in attending

to physical cues in conversation to difficulties imagining others' perspectives, that are linked to HDA use, which should ideally be tested over both short and long terms. Common reasoning and critical thinking tests, like the CRT (cognitive reflection test), could readily be repurposed to evaluate the sorts of consequences of HDA consumption discussed above.

Apart from those raised in the previous sections, broader, second-order concerns merit attention as well. Although we have tended to focus on first-order concerns that might directly impact individuals, wide-ranging group dynamics might be impacted as well. For example, turn-taking—the ability of a set of people to distribute conversation equally—is one of the best predictors of the collective intelligence of groups (Woolley et al., 2010). Researchers could evaluate whether the increase in one-sided conversations sustained with HDAs might impact collective cognition using paradigms like those developed in these canonical contexts.

Globally, there are three crucial directions such research programs must take. They should analyze: (1) overall usage patterns; (2) individual differences in usage and context; and (3) whether the extent or type of use creates the sorts of conversational and ethical concerns evoked here. Part of this research program should seek to clarify what counts as a significant interaction with an HDA (i.e., an interaction sufficient to shape a given target behavior), and then proceed to identify how this varies between individuals and usage contexts in order to develop informed and thoughtful guidelines for HDA usage.

In conjunction with such empirical research, further engagement with previous philosophical and phenomenological analyses of conversation and its relation to core aspects of human personality would also be very useful. Thinkers such as Martin Buber (Agassi, 1999; Buber, 1996), Emmanuel Levinas (1969), Simone Weil (2009), and Jürgen Habermas (1984, 1987), among others, have reflected on the ethics of dialogue, power, and ethical responsibility, and have pointed to the risks associated with depersonalization and a loss of sensitivity to other people, as well as to modes of attention that can enable one to resist such tendencies. Engagement with such thinkers is a vital means of uncovering the range of potential problems that HDA–human interaction can generate. Thus, it can help direct researchers to the particular types of experiments that best identify and detect whether these problems are, in fact, a likely result of such interactions. It is our hope that, in pursuing evaluations of LLMs and HDAs more broadly, by remaining thoughtfully focused on their human impact, we can avoid being so taken in by the spectacle of these agents that we fail to ask the right questions about what interaction with them might be doing to us.

Acknowledgments

The authors wish to thank the editors for encouraging and thoughtful discussion as this manuscript was just being conceived.

Disclosure Statement

The authors have no financial or nonfinancial disclosures to share for this article.

References

Agassi, J. (1999). *Martin Buber on psychology and psychotherapy: Essays, letters, and dialogue*. Syracuse University Press.

Akcaoğlu, M. Ö., Mor, E., & Külekçi, E. (2023). The mediating role of metacognitive awareness in the relationship between critical thinking and self-regulation. *Thinking Skills and Creativity*, *47*, Article 101187. https://doi.org/10.1016/j.tsc.2022.101187

Benedek, M., Franz, F., Heene, M., & Neubauer, A. C. (2012). Differential effects of cognitive inhibition and intelligence on creativity. *Personality and Individual Differences*, *53*(4), 480–485. https://doi.org/10.1016/j.paid.2012.04.014

Biever, C. (2023). ChatGPT broke the Turing test—The race is on for new ways to assess AI. *Nature*, *619*(7971), 686–689. https://doi.org/10.1038/d41586-023-02361-7

Borgmann, A. (1987). *Technology and the character of contemporary life: A philosophical inquiry*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/T/bo23186480.html

Buber, M. (1996). *I and thou* (W. Kaufmann, Trans.). Touchstone.

Carolus, A., Schmidt, C., Schneider, F., Mayr, J., & Muench, R. (2018). Are people polite to smartphones? In M. Kurosu (Ed.), *Human-computer interaction: Interaction in context* (pp. 500–511). Springer, Cham. https://doi.org/10.1007/978-3-319-91244-8_39

Cassotti, M., Agogué, M., Camarda, A., Houdé, O., & Borst, G. (2016). Inhibitory control as a core process of creative problem solving and idea generation from childhood to adulthood. *New Directions for Child and Adolescent Development*, *2016*(151), 61–72. https://doi.org/10.1002/cad.20153

Chaudhary, Y., & Penn, J. (2024). Beware the intention economy: Collection and commodification of intent via large language models. *Harvard Data Science Review*, (Special Issue 5). https://doi.org/10.1162/99608f92.21e6bbaa

Cialdini, R. B. (2006). *Influence: The psychology of persuasion.* Harper Business.

Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., & Landay, J. A. (2024). The illusion of empathy? Notes on displays of emotion in human-computer interaction. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *CHI '24: Proceedings of the CHI*

Conference on Human Factors in Computing Systems (Article 446). ACM. https://doi.org/10.1145/3613904.3642336

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research Findings and Recommendations* (ERIC Report No. ED315423). Institute of Education Sciences. https://eric.ed.gov/?id=ED315423

Felnhofer, A., Kafka, J. X., Hlavacs, H., Beutl, L., Kryspin-Exner, I., & Kothgassner, O. D. (2018). Meeting others virtually in a day-to-day setting: Investigating social avoidance and prosocial behavior towards avatars and agents. *Computers in Human Behavior*, *80*, 399–406. https://doi.org/10.1016/j.chb.2017.11.031

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*(3), 674–687. https://doi.org/10.1037/xge0000070

Fisher, M., Smiley, A. H., & Grillo, T. L. H. (2021). Information without knowledge: The effects of Internet search on learning. *Memory 30*(4), 375–387. https://doi.org/10.1080/09658211.2021.1882501

Fiske, S., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). SAGE Publications.

Fogg, B., & Nass, C. (1997). How users reciprocate to computers: An experiment that demonstrates behavior change. In A. Edwards & S. Pemberton (Eds.), *CHI EA '97*: *CHI '97 Extended Abstracts on Human Factors in Computing Systems* (pp. 331–332). ACM. https://doi.org/10.1145/1120212.1120419

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox*. MIT Press.

Habermas, J. (1984). *The theory of communicative action, Volume 1: Reason and the rationalization of society* (T. McCarthy, Trans.). Beacon Press.

Habermas, J. (1987). *The theory of communicative action, Volume 2: Lifeworld and system: A critique of functionalist reason* (T. McCarthy, Trans.). Beacon Press.

Hart, D. B. (2015). *The story of Christianity*. Quercus.

Kahneman, D. (2011). *Thinking fast and slow*. Penguin.

Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. *Philosophy & Technology*, *36*(2), Article 27. https://doi.org/10.1007/s13347-023-00606-x

Kory-Westlund, J. M., Won Park, H., Grover, I., & Breazeal, C. (2022). Long-term interaction with relational SIAs. In B. Lugrin, C. Pelachaud, & D. Traum (Eds.), *The Handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics: Vol. 2. Interactivity, platforms, application* (Vol. 48, pp. 195–260). ACM. https://doi.org/10.1145/3563659.3563667

Lakin, J. L. (2013). Behavioral mimicry and interpersonal synchrony. In J. A. Hall & M. L. Knapp (Eds.), *Nonverbal communication* (pp. 539–575). De Gruyter Mouton. https://doi.org/10.1515/9783110238150.539

Langer, A., Marshall, P. J., & Levy-Tzedek, S. (2023). Ethical considerations in child-robot interactions. *Neuroscience & Biobehavioral Reviews*, *151*, Article 105230. https://doi.org/10.1016/j.neubiorev.2023.105230

Levinas, E. (1969). *Totality and infinity: An essay on exteriority* (A. Lingus, Trans.). Duquesne University Press.

Luria, M. (2023, April 11). Your ChatGPT relationship status shouldn't be complicated. *Wired*. https://www.wired.com/story/chatgpt-social-roles-psychology/

Marsh, E. J., & Rajaram, S. (2019). The digital expansion of the mind: Implications of internet usage for memory and cognition. *Journal of Applied Research in Memory and Cognition*, *8*(1), 1–14. https://doi.org/10.1016/j.jarmac.2018.11.001

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), Article e2313925121. https://doi.org/10.1073/pnas.2313925121

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Nagam, V. M. (2023). Internet use, users, and cognition: On the cognitive relationships between Internet-based technology and Internet users. *BMC Psychology*, *11*(1), Article 82. https://doi.org/10.1186/s40359-023-01041-5

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678.

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology*, *29*(5), 1093–1109.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Center for the Study of Language and Information; Cambridge University Press.

Reuten, A., van Dam, M., & Naber, M. (2018). Pupillary responses to robotic and human emotions: The uncanny valley and media equation confirmed. *Frontiers in Psychology*, 9, Article 774.

https://doi.org/10.3389/fpsyg.2018.00774

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Saghafian, S., & Idan, L. (2024). Effective generative AI: The human-algorithm centaur. *Harvard Data Science Review*, (Special Issue 5). https://doi.org/10.1162/99608f92.19d78478

Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. Riverhead Books.

Souza, C. S. D. (2005). *The semiotic engineering of human-computer interaction*. MIT Press.

Turkle, S. (2012). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Turkle, S. (2017). Empathy machines: Forgetting the body. In V. Tsolas & C. Anzieu-Premmereur (Eds.), *A psychoanalytic exploration of the body in today's world* (pp. 17–27). Routledge. https://doi.org/10.4324/9781315159683-3

van der Goot, M. J., & Etzrodt, K. (2023). Disentangling two fundamental paradigms in human-machine communication research: Media equation and media evocation. *Human-Machine Communication*, *6*, 17–30. https://doi.org/10.30658/hmc.6.2

Weil, S. (2009). *Waiting on God*. Routledge. https://doi.org/10.4324/9780203092477

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Welivita, A., & Pu, P. (2024). *Is ChatGPT more empathetic than humans?* ArXiv. https://arxiv.org/html/2403.05572v1

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688. https://doi.org/10.1126/science.1193147

Footnotes

1. The authors would like to dedicate this article to Professor Eta Berner, a pioneer in the study of the use of AI in medical decision-making, whose life, work, and values influenced the current authors and many others.

↩

2. There are a handful of good exceptions to this rule, including Kasirzadeh and Gabriel's (2023) work on aligning human and LLM norms, a popular discussion by Michal Luria (2023) on ChatGPT's influence on social roles, and Chaudhary and Penn's (2024) concerns about the burgeoning of a marketplace of intentions. ↩

3. It is worth noting, others have evaluated similar concerns in relation to other forms of specifically social varieties of artificial intelligence, though not in relation to LLMs. Turkle (2012) considers a broad set of social concerns, while Kory-Westlund et al. (2022) evaluate long-term social consequences, and Langer et al. (2023) raise developmental questions. ↩

4. For a good example of a robust assessment, see Mei et al.'s (2024) method. ↩

5. Two clarifications are in order: 1. While the focus of this essay is on concerns specifically occasioned by HDAs, aspects of our criticisms apply to earlier technologies as well, some of which are underscored throughout. 2. Our reference for 'normal' conversation is ordinary, human-to-human conversation, conversation that is not mediated by technology at all, as we clarify and justify below. ↩

6. Along the same lines, they suggest, "People's responses show that media are more than just tools. Media are treated politely, they can invade our body space, they can have personalities that match our own, they can be a teammate, and they can elicit gender stereotypes. Media can evoke emotional responses, demand attention, threaten us, influence memories, and change ideas of what is natural. Media are full participants in our social and natural world" (Reeves & Nass, 1996, p. 251). ↩

7. It is worth noting that this sort of support is consistent with a great deal of contemporary cognitive science, neatly dovetailing with accounts as disparate as Gigerenzer and Selten's (2001) ecological rationality framework, Kahneman's (2011) dual process theory, and Mercier and Sperber's (2017) argumentative theory of reasoning. ↩

8. van der Goot and Etzrodt (2023, p. 18) differentiate these latter approaches from the media equation in this way: "In contrast, the Media Evocation paradigm conceptualizes machines as objects that are betwixt and between former diametrical opposites—such as person versus thing—evoking reflection and negotiation processes about the nature of the object but also about ourselves and human identity." ↩

9. Even if people sometimes talk to their cars or to their pets, this does not constitute a back-and-forth verbal conversation in the same sense, such that interhuman interaction has been the only context in which the latter was available. ↩

10. On concerns regarding the disembodied features of screen-mediated conversations, see Turkle (2012, 2017). ↩

11. By contrast, as we will see, in HDA–human conversations, the disembodied nature of the conversation is not merely contingent but is structurally connected to the disembodied nature of the conversation partner. ↩

12. For a good overview of these findings see J. L. Lakin's (2013) work that reviews a number of aspects of conversational "behavioral mimicry," from postural adjustments to tone, speed of exchange, syntax, and many others. ↩

13. Statistical and experimental research into how individuals actually tend to use LLMs and similar HDAs is perhaps the most pressing of the empirical research programs suggested below. ↩

14. This classification of concerns is meant to illustrate distinct features of research programs that merit consideration, but, to the extent that the linguistic and cognitive observations entail the impoverishment of our conversational and rational capabilities, from our perspective all are ultimately considerations of moral and ethical reflection as well. ↩

15. Here and elsewhere, much of the concerns illustrated will likely be a function of the regularity of individuals' interactions with HDAs. ↩

16. For the former perspective, consider the role intention-formation plays in Gricean pragmatics, in which what ensures conversational coordination is near constant monitoring of the conversation partner, while actively updating one's beliefs. For the latter, consider recent discussions on the challenges common reciprocal constructions create in linguistic analyses. ↩

17. These concerns about 'disembodied' interaction may also apply to various forms of human–human interaction, in proportion to the degree of disembodiment. Thus, the effects of disembodied conversation may be stronger in relation to conversations over text messaging than in relation to conversations over telephone, since the latter at least involves various embodied-particular aspects of the live human voice. Again, on concerns regarding disembodied communication, see Turkle (2012, 2017). ↩

18. In our use, 'personality' is related to personhood, and requires lived, formative experiences in the physical world. Any given digital agent can be architected to have, for example, a sunny disposition or a no-nonsense tone, but these qualities are, we believe, distinct from person-based personalities. For a fuller characterization of what we have in mind, see Section 3.2, paragraph 6. ↩

19. While mostly considering various ways people relate to technological media as social agents, Reeves and Nass (1996) also provide experimental evidence of effects in the other direction. They cite, for example, experiments indicating that interactions with personlike media technologies can also prime and influence the way that people then interact with other human beings. See especially pp. 86–87. ↩

20. Turkle (2012) discusses similar dynamics extensively in relation to other forms of technological interaction. ↩

21. See the studies by Reeves and Nass indicating that personlike interactions with computers can generate a sense of being 'on a team' with the computer, and that this sense of 'being teammates' can lead people to conform to and mirror the attitudes that one's 'teammate' enacts (Reeves & Nass, 1996, pp. 154–160). ↩

22. Previous experiments by Reeves and Nass point to ways in which computers can engage in behaviors—for example, politeness and flattery—that can make their users feel more comfortable in continuing to engage with them (Reeves & Nass, 1996, pp. 33–36, 59–63, 70–72). See also p. 21: "There is good evidence that when people are interviewed about sensitive topics, they are more likely to tell the truth to a computer than to another person." ↩

23. Facione (1990, p. 6) offers a comprehensive definition of critical thinking in line with these observations, suggestion that it is "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation and inference as well as explanation of the evidential conceptual, methodological, criteriological or contextual considerations upon which that judgment was based." ↩

24. Apart from the fairly intuitive association between metacognition and critical thinking, there is recent empirical research directly linking the two (see Akcaoğlu et al., 2023). ↩

25. The specific neuroanatomical substrate most consistently implicated in inhibitory tasks in general, and in creative tasks in particular, is the anterior cingulate cortex. Behaviorally, it is worth noting that interactions of this sort have yet to be investigated, and the converse conclusion might well hold: perhaps by essentially outsourcing generation of nonoriginal ideas to an HDA, the user might essentially preserve inhibitory resources for more demanding contexts, perhaps enabling greater rather than less creativity. ↩

26. Although they come to different conclusions about its origin, impact, and normative standing, both Kahneman's (2011) heuristics and bias program and Gigerenzer and Selten's (2001) notion of an adaptive toolbox hinge on cognitive miserliness. ↩

27. The use of a dual-process framework here is largely pragmatic, as it generally affords testable hypotheses and is a prominent model in the contemporary reasoning literature. However, it is worth noting that the consequences of regular HDA use could be considerably more dire if other reasoning theories better approximate the underlying reality. For example, on the argumentative theory of reasoning, a position developed by Hugo Mercier and Dan Sperber (2017), reasoning is conceived as primarily being at the service of winning arguments and only secondarily truth oriented. Their account makes sense of a number of curious results from the motivated reasoning literature, especially the prevalence of 'my-side bias' and the like. More relevantly, HDAs could essentially accelerate the ability of individuals to build arguments for what they already believe, while in many cases potentially also inserting and reinforcing those cognitive biases shared between HDAs and humans into their arguments (Saghafian & Idan, 2024). ↩

References

- Agassi, J. (1999). *Martin Buber on psychology and psychotherapy: Essays, letters, and dialogue*. Syracuse University Press.

  ↩
- Benedek, M., Franz, F., Heene, M., & Neubauer, A. C. (2012). Differential effects of cognitive inhibition and intelligence on creativity. *Personality and Individual Differences*, *53*(4), 480–485. https://doi.org/10.1016/j.paid.2012.04.014

  ↩
- Biever, C. (2023). ChatGPT broke the Turing test—The race is on for new ways to assess AI. *Nature*, *619*(7971), 686–689. https://doi.org/10.1038/d41586-023-02361-7

  ↩
- Borgmann, A. (1987). *Technology and the character of contemporary life: A philosophical inquiry*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/T/bo23186480.html

  ↩
- Buber, M. (1996). *I and thou* (W. Kaufmann, Trans.). Touchstone.

  ↩
- Carolus, A., Schmidt, C., Schneider, F., Mayr, J., & Muench, R. (2018). Are people polite to smartphones? In M. Kurosu (Ed.), *Human-computer interaction: Interaction in context* (pp. 500–511). Springer, Cham. https://doi.org/10.1007/978-3-319-91244-8_39

  ↩
- Cassotti, M., Agogué, M., Camarda, A., Houdé, O., & Borst, G. (2016). Inhibitory control as a core process of creative problem solving and idea generation from childhood to adulthood. *New Directions for Child and Adolescent Development*, *2016*(151), 61–72. https://doi.org/10.1002/cad.20153

  ↩
- Cialdini, R. B. (2006). *Influence: The psychology of persuasion.* Harper Business.

  ↩
- Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., & Landay, J. A. (2024). The illusion of empathy? Notes on displays of emotion in human-computer interaction. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *CHI '24: Proceedings of the CHI Conference on Human Factors in Computing Systems* (Article 446). ACM. https://doi.org/10.1145/3613904.3642336

↩

- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299.

↩

- Felnhofer, A., Kafka, J. X., Hlavacs, H., Beutl, L., Kryspin-Exner, I., & Kothgassner, O. D. (2018). Meeting others virtually in a day-to-day setting: Investigating social avoidance and prosocial behavior towards avatars and agents. *Computers in Human Behavior*, *80*, 399–406. https://doi.org/10.1016/j.chb.2017.11.031

↩

- Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*(3), 674–687. https://doi.org/10.1037/xge0000070

↩

- Fisher, M., Smiley, A. H., & Grillo, T. L. H. (2021). Information without knowledge: The effects of Internet search on learning. *Memory 30*(4), 375–387. https://doi.org/10.1080/09658211.2021.1882501

↩

- Fiske, S., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). SAGE Publications.

↩

- Fogg, B., & Nass, C. (1997). How users reciprocate to computers: An experiment that demonstrates behavior change. In A. Edwards & S. Pemberton (Eds.), *CHI EA '97*: *CHI '97 Extended Abstracts on Human Factors in Computing Systems* (pp. 331–332). ACM. https://doi.org/10.1145/1120212.1120419

↩

- Habermas, J. (1984). *The theory of communicative action, Volume 1: Reason and the rationalization of society* (T. McCarthy, Trans.). Beacon Press.

↩

- Habermas, J. (1987). *The theory of communicative action, Volume 2: Lifeworld and system: A critique of functionalist reason* (T. McCarthy, Trans.). Beacon Press.

↩

- Hart, D. B. (2015). *The story of Christianity*. Quercus.

↩

- Levinas, E. (1969). *Totality and infinity: An essay on exteriority* (A. Lingus, Trans.). Duquesne University Press.

↩

- Marsh, E. J., & Rajaram, S. (2019). The digital expansion of the mind: Implications of internet usage for memory and cognition. *Journal of Applied Research in Memory and Cognition*, *8*(1), 1–14. https://doi.org/10.1016/j.jarmac.2018.11.001

↩

- Nagam, V. M. (2023). Internet use, users, and cognition: On the cognitive relationships between Internet-based technology and Internet users. *BMC Psychology*, *11*(1), Article 82. https://doi.org/10.1186/s40359-023-01041-5

↩

- Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678.

↩

- Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology*, *29*(5), 1093–1109.

↩

- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Center for the Study of Language and Information; Cambridge University Press.

↩

- Reuten, A., van Dam, M., & Naber, M. (2018). Pupillary responses to robotic and human emotions: The uncanny valley and media equation confirmed. *Frontiers in Psychology*, 9, Article 774. https://doi.org/10.3389/fpsyg.2018.00774

↩

- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

↩

- Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. Riverhead Books.

↩

- Souza, C. S. D. (2005). *The semiotic engineering of human-computer interaction*. MIT Press.

↩

- Turkle, S. (2012). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

↩

- Turkle, S. (2017). Empathy machines: Forgetting the body. In V. Tsolas & C. Anzieu-Premmereur (Eds.), *A psychoanalytic exploration of the body in today's world* (pp. 17–27). Routledge. https://doi.org/10.4324/9781315159683-3

↩

- Weil, S. (2009). *Waiting on God*. Routledge. https://doi.org/10.4324/9780203092477

↩

- Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

↩

- Welivita, A., & Pu, P. (2024). *Is ChatGPT more empathetic than humans?* ArXiv. https://arxiv.org/html/2403.05572v1

↩

- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688. https://doi.org/10.1126/science.1193147

↩