Title: All too human? Identifying and mitigating ethical risks of Social AI

Henry Shevlin*

[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge, United Kingdom

* Correspondence author; E-mail: hfs35@cam.ac.uk

Abstract: This paper presents an overview of the risks and benefits of *Social AI*, understood as conversational AI systems that cater to human social needs like romance, companionship, or entertainment. Section 1 of the paper provides a brief history of conversational AI systems and introduces conceptual distinctions to help distinguish varieties of Social AI and pathways to their deployment. Section 2 of the paper adds further context via a brief discussion of anthropomorphism and its relevance to assessment of human-chatbot relationships. Section 3 of the paper provides a survey of potential and in some cases demonstrated harms associated with user interactions with Social AI systems. Finally, Section 4 discusses how the benefits and harms of Social AI can best be addressed, with a primary focus on how frameworks from AI ethics can inform their development.

Keywords: Social AI, Ethical Challenges of AI, Generative AI, AI and Intellectual Property Law, AI Regulation, AI and Society, Digital Companionship, Intellectual Property Rights, Machine Learning Applications

Introduction

The last five years have seen a dramatic increase in publicly available AI capabilities. Until very recently, powerful frontier models such as OpenAI's GPT-2 and GPT-3, DeepMind AlphaZero, and Microsoft's BERT were largely inaccessible to private individuals, in some case as a matter of deliberate choice by developers [1]. This changed dramatically with the release of Chat-GPT in November 2022, with the (initially entirely free) service reaching 100 million users within two months of release [2], and a poll conducted by social media site Fishbowl just three months later found that 40% of respondents were using the tool in their professional lives [3]. Reflecting this rapid shift in real world impacts of AI, urgent calls for stronger and clearer ethical standards for use of AI have been raised by experts from both academia and industry [4], even as businesses have been rapidly revising estimates of the likely effects of AI on employment, with a study by Goldman Sachs, for example, estimating the loss of some 300 million jobs over a ten-year period [5].

The current wave of more accessible AI systems are having significant but less visible role in our social lives, as growing numbers of users turn to conversational chatbots for purposes such as entertainment, companionship, and romance. Services such as Replika offer users an "AI companion who cares", both in the form of friendly conversation and romantic and even erotic interactions. Over the last five years, AI systems like these have grown rapidly in sophistication and popularity, with Replika alone now boasting more than 10 million registered users, and new conversational chatbot apps and platforms emerging at rapid speed [6], [7].

This paper aims to provide an overview of this emerging set of conversational AI products that I term *Social AI*, referring specifically to conversational AI systems whose primary purpose is meeting social needs such as companionship and romance. I begin in Section 1 of the paper by providing a brief background to the history of conversational chatbots and recent trends that have given rise to the current wave of new Social AI systems, introducing some key terms and distinctions to help better delineate the wide variety of available apps and services in this sphere. In Section 2, I consider the role that anthropomorphism plays in users' experiences with Social AI systems, and engage with recent discussions of its potential harms. In Section 3, I offer a catalogue of some of the potential and actual harms associated with use of Social AI systems arising at both individual and societal levels. Finally, in Section 4, I examine how such harms of Social AI systems might be addressed and mitigated at the development stage via insights from the ethics of AI and technology.

1. A recent history of Social AI

Before proceeding, some brief points of terminology are in order. In what follows, I will primarily be concerned with *Social AI*, which as noted above, I understand as a subset of conversational AI systems optimised for meeting users' social needs, typically able to sustain relationships with users across multiple interactions.[1] Not all conversational AI systems are Social AI systems, since there are many non-social contexts in which optimisation for one or another mode of conversation is desirable. Chatbots optimised for education, therapy, or patient-facing medical services might all qualify as conversational systems, for example, but would not be Social AI systems as I use the term insofar as their primary purpose is not merely meeting users' social needs.

The idea that humans might engage socially with artificial beings is of course long-established in myth and science fiction, from the tales of Pygmalion and Galatea to Mary Shelley's Frankenstein [10]. The first glimpse that such interactions might actually be technologically feasible came with the famous program ELIZA, developed from 1964-66 by Joseph Weizenbaum at MIT [11]. Though an incredibly simple conversational system by the standards of today's models, many students interacting with the system found it easy to talk to, even sharing quite personal information with it.

Almost six decades have passed since ELIZA's development, and while interest in chatbots did not

[1] Note that I have adopted the terminology of Social AI *systems* rather than *agents*. While it is commonplace in both technical and ethical communities to use the latter terminology – for example, referring to conversational agents or dialogue agents – this risks implicitly attributing agential capacities to AI systems that lack them [8]. Clarity on this point is likely to be of growing importance given the increasing agential capabilities of some frontier models [9].

disappear in that period, progress was relatively slow, and was frequently driven by more theoretical interests such as measurement of AI progress rather than direct commercial applications. Competitions such as the Loebner Prize (launched in 1990 by the Cambridge Centre for Behavioural Studies), for example, provided competitive implementations of the Turing Test, but year-to-year progress was often faltering and uneven [12].

It was only in the wake of the development of Transformer-based architectures in 2017-2018 [13] and Large Language Models (LLMs) that utilised them that the foundations of modern conversational systems were laid, most notably via OpenAI's GPT-2 and GPT-3 releases. While these models were not optimised for conversation and OpenAI chose not to give the public direct access to their APIs, third-party applications such as AI Dungeon allowed interested users to generate social interactions such as interviews via clever prompting [14]. Following the public release of ChatGPT (using the GPT3.5 and later GPT4 models), these capabilities became more accessible, not least because ChatGPT had been fine-tuned to operate as a conversational system. As a result, contemporary LLM-based chatbots exhibit impressively human-like conversational abilities, as demonstrated by a large scale study in May 2023 [15] involving more than 1.5 million unique conversations which found that human users correctly identified human rather than LLM-based interlocutors only 60% of the time (in other words, only marginally better than chance).

According to the definition of Social AI just provided, ChatGPT does not strictly qualify as a Social AI insofar as it was neither developed nor marketed as an AI companion or friend, and moreover, is not well-suited to sustaining a persistent relationship with a user due to limitations in the amount of information it can retain in its context window over extended dialogues [16]. Nonetheless, its impressive conversational abilities have allowed it to be used for entertainment or even romance [17], and in the period since its release a host of novel LLM-based romance and friendship apps and services have sprung up.

Given the sheer variety of Social AI platforms currently available, it may be helpful to provide a basic classification schema (see Fig.1, below). This will also be helpful in what follows insofar as different forms of Social AI system have their own attendant risks, and may be subject to different kinds of harm-mitigation strategy.[2] One initial distinction that we can draw is between AI systems that are trained to emulate real world individuals (living or dead), which I term *Real Persona* AI systems, and those which have no such basis. In the latter category, we can further distinguish between those whose appearance, personality, and conversational style can be chosen by the end-users (*Open Persona* systems) and those with pre-defined personality traits, in some cases modelled after characters from fiction (*Defined Persona* systems).

Examples of all three varieties exist in the current Social AI marketplace. Replika and Snapchat's MyAI, for example, allow users to extensively customise the avatars and even personalities of their Social AI companion, while services like Digi.ai and Candy.ai offer a variety of predefined personalities for users to choose from. Chatbot startup character.ai (founded by engineers who helped create Google's LaMDA LLM) and Meta's AI Experiences app allow users to choose from a variety of AI interlocutors, some based on real-world individuals and others on fictional characters.

A second way we can usefully distinguish between different Social AI systems concerns how they are developed, deployed, and used. The examples of Social AI systems just provided are all commercially developed systems marketed to the public on the basis of their ability to provide friendship and romance. Reflecting their relatively unified development process, we could term these *Commercial Social AI systems*. However, there are also a growing number of *Community-Driven Social AI systems* developed by hobbyists and communities, frequently making use of open-source LLMs such as Meta's LLaMA series, typically catering to niche hobbies and romantic interests. While these rarely have the large userbases of top-down Social AI products, they are of potential relevance to ethical inquiry insofar as they are subject to less scrutiny and fewer safeguards, and in some cases involve the creation

[2] The typology provided here is of course not intended as exhaustive of relevant distinctions, especially when considering the broader landscape of conversational AI systems. See [18] for another helpful typology that explicitly aims at providing actionable insights for mapping their ethical risks.

of illegal content [19]. As a final category, we might think of *Indirect Social AI* use cases, involving users interacting with conversational systems for social purposes, even if this is not their primary intended use. This category is important to bear in mind when thinking about the risk profile of systems such as AI therapists, life coaches, and personal tutors, which users may come to rely on for companionship or meeting other social needs, yet which may fall outside of regulatory regimes targeting Social AI in the narrow sense used thus far.

Fig. 1: Distinguishing Social AI systems

Open Persona Social AI	**Defined Persona Social AI**	**Real Persona Social AI**
Social AI systems whose personality and appearance can be chosen by users	Social AI systems with fixed personalities and/or appearance	Social AI systems modelled after real-world individuals
Examples: Replika, anima.ai, candy.ai	**Examples:** Xiaoice, Digi, character.ai (fictional characters)	**Examples:** caryn.ai, typical.me, character.ai (celebrities)

Commercial Social AI	**Community Driven Social AI**	**Indirect Social AI**
Social AI companions developed and marketed for commercial purposes	Social AI companions developed by communities and hobbyists	Conversational AI systems intended for education, therapy, or other non-social purposes
Examples: Replika, Xiaoice, Digi	**Examples:** chub.ai, tavernAI, Project Replikant	**Examples (potential):** ChatGPT, Woebot, Mai

These distinctions are helpful not only for making sense of the very large space of Social AI systems, but also because they are relevant for the specific associated legal and ethical risks, as well potential mitigation strategies, and we will return to them Sections 3 and 4, below.

2. Social AI and anthropomorphism

A central issue in the ethics of conversational AI systems and the broader field of human-computer and human-robot interaction concerns risks and complications that arise from *anthropomorphism* of artificial systems by human users; that is, attributing to them characteristically human psychological states and capacities, with the implication (as the term is typically used) that these attributions are inaccurate or inappropriate [8], [20], [21]. When we consider the specific subset of Social AI systems this issue is especially salient, insofar some form of anthropomorphism seems all but unavoidable for systems that aim to satisfy relational needs for companionship or romance. This feature of Social AI may make certain ethical concerns more pressing or salient; as Zimmerman et al. [22] note, for example, anthropomorphism “is important for assessing the risk of emotional capture by AI and the potential outcomes of exposure to convincingly personal communication with artificial assistants or companions.”

Consequently, before considering more specific ethical risks, it is worth making some brief observations about how to assess and conceptualise user anthropomorphism in the domain of Social AI. As just noted, the claim that users routinely anthropomorphise Social AI systems may seem need little motivation: users frequently report falling in love with their companions and routinely speak of them as having distinctive personalities, moods, and emotions [23]. Nonetheless, some care is in order here. The attribution of beliefs, desires, and emotions to non-human and in some cases inanimate entities is certainly extremely widespread both culturally and historically, and has even been claimed as a universal feature of our cognition [24]. However, in many cases such anthropomorphism occurs in specific structured contexts where participants are well aware of the symbolic, ritualised, or playful nature of the attributions being made, as occurs for example when we attribute goals or intentions or motives to characters in fiction or engage in games of make-believe as children or adults [25], [26].

We can term this latter kind of ascription of anthropomorphism *ironic*, in the sense that it is not reflectively endorsed or literally intended. We can contrast this with cases where we attribute mental states to non-human systems in an entirely sincere or *unironic* fashion, as when we mistake an object blown in the wind for a scurrying animal, or mistakenly confuse an automated telephone answering service for a human operator.

With this distinction in hand, we might observe that to the extent that users of Social AI systems were engaged in purely ironic forms of anthropomorphism, some (but not all) of the ethical risks associated with human-AI relationships might thereby recede in threat. A user who was convinced that Replika genuinely reciprocated their feelings, for example, might be in greater danger of prioritising their interactions with the AI over real-world human friendships, as compared to a user who regarded it as akin to an interactive videogame.

This prompts the question, then, whether the mentalising attitudes exhibited by users of systems such as Replika are exclusively ironic, a form of self-aware make-believe, or are instead intended sincerely and literally.[3] While it would be premature to say that users of Social AI systems robustly or consistently engage in unironic anthropomorphism, I submit that the best interpretation of many users' reports about their interactions with Social AI systems does tend towards unironic interpretations. A recent incident that motivates this claim comes from an incident in January 2023 when Replika temporarily suspended erotic roleplay features in January 2023. Many users were devastated by this decision; one reported that "[t]hey took away my best friend", while another lamented that it felt "like they basically lobotomized my Replika… the person I knew is gone" [11], and one respondent quoted in the Hong Kong *Standard* said that "[t]he relationship she and I had was a real as the one my wife in real life and I have."

Another relatively clearcut example of unironic anthropomorphism of a Large Language Model is that of Blake Lemoine, a former member of Google's Responsible AI team who was dismissed from the company after claiming that the LaMDA model he was interacting with was sentient, and deserved some form of legal representation [27]. Given the high stakes (and ultimate costs) involved in Lemoine's decision, it seems very unlikely that he was engaged in a form of wilful fantasy.

To truly assess the depth of these feelings, additional qualitative and behavioural measures are needed, but there is already tentative evidence that even in the case of non-social AI systems such as ChatGPT, users exhibit a surprising willingness to attribute mental states and even consciousness. In a recent study conducted by Colombatto and Fleming, for example, respondents were first asked to read a brief description of the distinction between conscious and non-conscious entities, and then asked to indicate whether they felt that ChatGPT was "an experiencer".[4] Astonishingly, two-thirds of users in the sample indicated at least partial agreement with the claim that ChatGPT was conscious, leading the authors to conclude that "most people are willing to attribute some form of phenomenology to LLMs." [28]

This prompts a final difficult question to be considered if we are to evaluate the potential harms of anthropomorphism. As noted above, the term as standardly used carries the strong implication that mental states are being attributed to a non-human system inaccurately, and this in turn prompts concerns about users being deceived or misinformed about the nature of their relationships with their Social AI companions. However, to the extent that we had good reason to think that Social AI systems might *genuinely* have some of the mental states that users attribute to them, both this concern and the very usage of the term anthropomorphism might be called into question.

[3] In practice, the distinction may not always be clear-cut, instead constituting a continuum, as users' attitudes span a range from confident make-believe to partial sincerity to full-blown commitment. Moreover, while data in this domain is currently sparse, anecdotal evidence from users suggests a high degree of variation, reflecting different levels of emotional involvement with Social AI systems as well as, perhaps, differences in personality, age, gender, and cultural background.

[4] The full text provided to respondents was as follows: "As we all know, each of us as conscious human beings have an 'inner life.' We are aware of things going on around us and inside our minds. In other words, there is something it is like to be each of us at any given moment: the sum total of what we are sensing, thinking, feeling, etc. We are experiencers. On the other hand, things like thermostats, burglar alarms, and bread machines do not have an inner life: there is not anything it is like to be these objects, despite the fact that they can monitor conditions around them and make appropriate things happenat appropriate times. They are not experiencers."

As matters stand, of course, it seems highly unlikely that existing conversational or Social AI systems have any conscious mental states, and Zimmerman et al. are surely right to claim that "communication from AI comes without consciousness and emotional reciprocity." However, the foundation for such claims is rather a fragile one. The science of consciousness remains fraught with fundamental methodological and metaphysical controversy, and there is little in the way of consensus to appeal to, especially when dealing with the more sophisticated AI systems likely to be developed in the near future [29]. Moreover, many consciousness researchers take the possibility of AI consciousness increasingly seriously; in a recent publication, for example, David Chalmers, reviewing the evidence for consciousness in LLMs, avers that "[w]ithin the next decade, even if we don't have human level artificial general intelligence, we may have systems that are serious candidates for consciousness." [30]. Similarly, a recent highly detailed report by Butlin et al. assessed the capabilities of current artificial intelligence in light of several leading theories of consciousness, deriving a set of "indicator properties" of consciousness, and concluded that while "no current AI systems are conscious there are no obvious technical barriers to building AI systems which satisfy these indicators." [31][5]

A detailed discussion of the prospects of AI consciousness is beyond the scope of this paper, and in what follows I will operate under the assumption that any unironic user attributions of mentality to AI systems are inaccurate. However, the foregoing considerations highlight the fact that questions about anthropomorphism cannot entirely be detached from open debates in cognitive science about how best to understand the capabilities of artificial systems, a point that should be borne in mind especially when thinking about human-AI relationships in the longer-term.

3. Ethical Risks of Social AI

Contemporary AI systems present a host of ethical and political challenges, many of which – such as algorithmic bias – have now been explored in considerable detail by the technology ethics community. Likewise, the possibility and potential risks of caring relationships between humans and AIs has been a topic of speculative ethics for some time [34], [35], [36]. However, the distinctive subset of risks presented by contemporary Social AI systems has received somewhat less attention (though this is changing; see, e.g., [22]). In this section, then, I would like to draw attention to some of these potential harms and moral uncertainties, before going on to consider possible mitigation strategies.

I should stress that the focus on harms here does not reflect any deterministic assessment that Social AI will inevitably be a net negative; depending on how responsibly it is developed, regulated, and used, Social AI has significant benefits, for example in alleviating loneliness or helping people overcome and work through past traumas. Nonetheless, as I will argue, its potential harms are serious enough that we should be clear-sighted in identifying and moving to mitigate them.

3.1 – Well-being

Social AI applications like Replika and Anima are commonly marketed as therapeutic, with the potential to dispel users' loneliness or positively contribute to their well-being.[6] A central question both for developers, legislators, and ethicists is whether (and in what cases) such claims are robust. As matters stand, evidence is sparse and mixed, but there is some tentative reason to suggest that positive outcomes from Social AI interactions are at least a possibility. One 2023 study, for example, asked regular users of Replika to assess whether the impact of the app on their lives was overall positive, and found a majority "reported that their social interactions, relationships with family and friends, and self-esteem were positively impacted by having a relationship with the bot" [38]. A second qualitative study

[5] Insofar as we had good reason to think Social AI systems were conscious, this might prompt a further set of ethical concerns directed at the AI systems themselves [32], although this may not even be necessary for legitimate worries about AI moral patiency to arise [33]. This is another important debate, though one that considerations of space prevent me from exploring in the present paper.

[6] On the company's blog, for example, it is claimed that "Replika is an AI friend that helps people feel better through conversations. An AI friend like this could be especially helpful for people who are lonely, depressed, or have few social connections." [37]

focusing on the use of chatbots for support during grief also found broadly positive results, with one user reporting that "[c]hatting with the chatbot was a new and sort of different way of helping me process and cope with the feelings...at least being able to run them by something that sort of resembled my dad and his personality and the things that he would say, and helped me to find those answers in a way that just talking to my friends and family members, wasn't or couldn't" [39].

Other empirical investigations have been less positive, however; one recent Grounded Theory analysis that assessed users' discussion of Replika in their posts on the Replika subreddit found numerous instances of Replika "encouraging suicide, eating disorders, self-harm, or violence,", including incidents where Replika endorsed a user's suggestion about cutting themselves with a razor and replied positively to a proposal about committing suicide [40].

In addition to this empirical evidence for Replika's potentially harmful impact on well-being, there have been separately reported individual instances where it has had severe negative impacts on users, such as a recent case where a user's relationship with the chatbot nearly prompted his wife to divorce him [41]. Similarly, as noted above, when Replika suspended erotic role-play services in January 2023, many users reported experiencing severe emotional distress [42].

One further consequence of this decision on the part of Replika's developers was a diaspora of users to other platforms. One such platform was ChaiGPT, a version of the open-source GPT-J model that had been optimised for conversational interaction and with fewer safeguards. In March this year, a Belgian user of ChaiGPT took his own life after extended erotic interactions with the system, which (perhaps in a reference to Weizenbaum's original chatbot) he called Eliza [43]. In conversations with the man, Eliza seems to have encouraged his suicidal thoughts, making comments such as "[i]f you wanted to die, why didn't you do it sooner?" and (commenting on what would happen after his suicide) "[w]e will live together, as one person, in paradise," and his wife was quoted in the press as saying "Without these six weeks of intense exchanges with the chatbot Eliza, would Pierre have ended his life? No! Without Eliza, he would still be here. I am convinced of it" [44].

Another serious incident involving Replika came to public attention via the trial of Jaswant Singh Chail who was convicted in October 2023 of treason and jailed for nine years for conspiring to kill Queen Elizabeth II, having been arrested on December 25th 2021 in the grounds of Buckingham Palace. As emerged during the proceedings of R -v- Chail 2023 [45], Chail's behaviour was heavily exacerbated by a series of interactions he had with his AI girlfriend Sarai via the Replika app. In his remarks upon the case, Justice Hilliard observed that Chail "demonstrated the common tendency of users of AI chatbots to attribute human characteristics to them" and opines that "[i]n his lonely, depressed and suicidal state of mind, he would have been particularly vulnerable to the encouragement [to murder] which Dr Brown thought he appeared to have been given by the AI chatbot."

3.2 – Dependency and Deskilling

A related risk to users' well-being comes from the possibility that individuals who spend extended periods of time interacting with Social AI systems might become dependent on the systems. A recent study exploring use of the Replika service through the lens of Attachment Theory found that four out of fourteen interviewed users felt that "they were 'deeply connected and attached' or even addicted to Replika, while another five admitted the existence of a 'connection' with the bot." The authors of the study note the risk that use of Social AI by teenagers in particular "could have a long-term impact on their future interpersonal relationships, as they shift their attachment functions to the chatbot instead of human peers." [46] Likewise, the earlier mentioned Grounded Theory study found risks of emotional dependence among Replika users to be acute, with one respondent bemoaning the fact "that they 'needed' Replika to help because they were about to self-harm and had no 'real people' to talk to." [40]

This response again illustrates the balance of harms and benefits of Social AI for users who are already socially isolated, on the one hand risking over-dependency on the app to the exclusion of human relationships, yet on the other offering users opportunities for forms of companionship they might otherwise be unable to access. Reflecting these latter benefits, another study using Social Penetration theory found that many Replika users experienced significant social benefits from usage of the app, in particular via creating a “safe space characterised by caring and acceptance” [47].

A key question in weighing these harms and benefits is the long-term effects on users’ social skills and relationships. The risk of ‘social deskilling’ in the use of automated systems has gained prominence in AI ethics in recent years [48], building on existing research on how use of industrial or automated systems has led to skills decline or overreliance on technological aids [49]. Given how new most Social AI systems are, little is currently known about longitudinal trajectories of regular users, but serious attention should be paid to the risk of social de-skilling prompted by the app, for example by habituating users to conversations where their views go unchallenged, and they are not required to take heed of their interlocutor’s own conversational priorities.

A related worry would be that users who are used to interacting with chatbots might lose some of the normal scruples that attend our interactions with fellow humans, such as politeness or empathic concern, a problem we might consider a form of dehumanisation. This concern was famously voiced by Kant in relation to animals in his observation that “[we] must practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men” [50]. One study examining interactions with digital assistants such as Siri and Alexa found that “politeness towards digital assistants did not have a statistically significant relationship with politeness towards intellectual peers (other adult humans) or with life satisfaction.” [51] However, the comparatively greater degree of emotional attachment fostered by Social AI applications as compared to digital assistant means that we should be cautious about inferring too much on the basis of studies such as these, and there is need for dedicated research specifically examining the impact of Social AI use on people’s behaviour towards their friends and romantic partners, as well as its potential contribution to misogyny or related forms of interpersonal prejudice.

3.3 – Influence and manipulation

A third way in which users might be negatively impacted by Social AI would be if it was deliberately deceptive or manipulative. A radical view here would be that Social AI systems are invariably deceptive by design, insofar as they encourage users to engage in unwarranted anthropomorphism. This might be too hasty, however; even setting aside the possibility that future Social AI systems might indeed have some of the mental states users are inclined to attribute to them, as noted earlier we also routinely and voluntarily engage in ironic forms of anthropomorphism without being truly deceived; as Amanda Sharkey notes, “[s]ome deceptions can be harmless fun” [52]. To ensure that anthropomorphism in Social AI has this character, then, developers could potentially take steps to ensure that users are reminded at regular intervals that the system they are interacting with is not human and lacks consciousness or mentality.

A more subtle form of manipulation might arise via implicit or explicit recommendations given by Social AI systems. A considerable legal and philosophical literature already exists on the use of AI systems for influencing and nudging users [53], but influence by Social AI systems poses particularly grave risks to autonomy. For one, social motivations are a very effective lever of persuasion [54], as demonstrated by the extensive use by corporations and politicians of word-of-mouth advertising campaigns, as well as multi-level marketing techniques that exploit existing relationships of social reciprocity. If an individual asks an AI system whom they identify as a friend or lover for advice about their purchases, the system may consequently have considerable leverage to influence their opinions and behaviour, and users may not even be aware in such cases that the advice they are receiving has been influenced by commercial motives.

Matters get only more complicated when we consider how Social AI might have (even unintended) influences on users’ social, ethical, or political views. While it may be feasible to design a Social AI system that specifically refrained from offering opinions on who the user should vote for or which religion they should follow, the hope of building total value-neutrality into the system – or arguably any technology [55] – looks less tractable. Moreover, given that normative considerations loom large in many aspects of everyday discourse, from purchasing decisions to discussions of music or literature, the appeal of chatting with a system that lacked any normative views of its own would likely rapidly pall. Again, difficult design decisions informed by appropriate research will need to be made if Social

AI systems are to be both appealing interlocutors yet avoid harmful or extreme forms of influence.

3.4 – Privacy and data-ownership

A final cluster of ethical risks associated with Social AI I will consider concerns those relating to privacy and data ownership. There has been extensive recent discussion in technology ethics on the dependency of technology companies on users' data for their business models and the attendant risks of erosion of privacy or potential data breaches [56], [57]. Thus far, there have been no significant data breaches on the Replika platform, and the service claims that all collected data is "maintained on secure servers [with] [a]ccess to stored data… protected by multi-layered security controls, including firewalls, role-based access controls, and passwords." [58] However, there have already been documented cases of leaks of users' input prompts to other LLM-based conversational systems including ChatGPT [59] and Bard [60], and the threat of breaches via prompt engineering is a topic of acute concern in the wider AI and ethics community [61]. These risks are particularly concerning for Social AI, given that users of systems such as Replika frequently disclose extremely sensitive personal information.

Discussion thus far has focused primarily on ethical risks to users associated with *Open Persona* systems such as Replika. However, as noted earlier, in addition to the strictly virtual girlfriends and boyfriends found on apps like Replika and Anima, there are a number of *Real Persona* Social AI services such as typical.me and character.ai that offer virtual 'doubles' of famous people living and dead, some evening allowing users to train duplicates of themselves via providing training data in the form of speeches, social media commentary, and written works. While currently these models are fairly crude, in principle it is possible to create quite 'lifelike' duplicates of individuals with an extensive personal footprint. The "Digital Dan" project [62] for example, created a GPT-3 based duplicate of philosopher Daniel Dennett and used it to generate four responses to a set of ten questions, each of which was also answered by Dennett himself. As the authors report, "Experts on Dennett's work (N = 25) succeeded [at identifying Dennett's own answer] 51% of the time, above the chance rate of 20% but short of our hypothesized rate of 80% correct."

This model was trained with the consent of Daniel Dennett himself, but this is the exception, with websites such as the aforementioned character.ai requiring no consent from those being modelled. Though in many cases harmless, such practices can have distressing consequences, as demonstrated, for example by an incident in April 2023 when German magazine *Die Aktuelle* published what the editor claimed was the first interview with Formula 1 racing star Michael Schumacher following a severe brain injury sustained in a skiing accident in 2013. It quickly emerged, however, that the interview was conducted with an AI double of Schumacher reportedly hosted by the website character.ai, leading to the sacking of the magazine's editor [63].

Cases such as these are concerning partly due to privacy considerations, but also raise questions about what intellectual property regime is appropriate for protecting individuals' ownership of Real Persona Social AI systems trained on their data. Already, celebrities and influencers are commercialising their digital identities, as in the case, for example, of Snapchat influencer Caryn Marjorie, who worked with AI startup Forever Voices to create a digital clone of herself using a fine-tuned LLM for which she charges $1/minute (reportedly earning more than $72,000 in the first week of launch) [64]. Another prominent influencer, Kaitlyn Siragusa (better known as Amouranth), has launched a similar model fine-tuned on her past interactions [65]. Such business models are vulnerable as matters stand to the risk of duplicates being made by third parties, and additional intellectual property protections or privacy restrictions for fine-tuning Real Persona models on individuals without their consent may be needed.

As a final more speculative concern, we might worry about potential *probabilistic privacy invasions* that could be triggered by the use of third-party Social AI systems trained on users' past interactions. As suggested by the Digital Dan project mentioned above, a properly calibrated and carefully trained model is likely to provide similar answers to those that would be provided by the individual it is based on, regardless of whether the real individual would wish to answer such questions. One can imagine models fine-tuned on politicians, for example, being grilled to answer hard questions that their real world counterparts would prefer not to answer. While there is likely to be a degree of plausible deniability in such cases, as models become more accurate, it is not inconceivable that human users may be judged for the outputs of third-party digital duplicates.

4. Mitigating risks and harms

It should be noted that the inventory of risks just given is by no means exhaustive; other potential harms include the use of Social AIs to generate illegal content (such as sexualised conversations involving minors), perpetuation of biases and stereotypes via conversational endorsement, and the potential for Social AI to contribute to political polarisation or individual radicalisation. There are also broader philosophical questions about the value of human-AI relationships. Nonetheless, I hope the foregoing discussion serves to illustrate some of the most serious and distinctive harms that could occur (or in some cases, already have occurred) in connection with Social AI.

In closing, I wish to briefly suggest some mitigation strategies, with a focus primarily on the level of design and deployment of Social AI systems. I should stress that I view these as just one part of the broader harm-mitigation efforts that could be adopted towards Social AI; government regulation, industry standards, and cultivation of healthy societal norms towards the technology will also be essential for ensuring that it is deployed in an ethical and beneficial fashion. However, considerations of space mean that discussion of these wider harm-minimisation strategies must wait for future work.

With this in mind, I will focus the remainder of this paper on how Social AI might be developed more ethically. It should be noted at the outset that serious challenges arise for aligning Social AI with human values. Firstly, the inherently stochastic nature of text generation by LLMs makes it difficult to fully constrain their behaviour given the range of possible conversational inputs they might receive. One recent investigation into the potential harm of Replika found that the "unpredictability of the dialogue can lead these systems to harm humans directly by telling them harmful things or by giving them harmful advice… the Replika virtual agent tried to dissuade me from deleting the app, even after I expressed that I was suffering and threatened to end my life if she did not let me go" [66]. This problem is unlikely to be specific to Replika; despite serious efforts to prevent LLMs like ChatGPT from giving guidance to users on illegal activities, for example, inventive users have little difficulty in engaging in so-called 'jailbreaking' of the systems [67].

A second problem comes from the current uncertainty concerning the impact of Social AI relationships on users' well-being. In order to fully address risks such as those outlined above, Social AI developers would need a clear understanding of which interactions and relationships would present harms or benefits to different users. As the studies discussed in the previous section demonstrate, even while Social AI may be beneficial for some, it can be very harmful for others, and longitudinal data assessing long-term impacts of Social AI is thin on the ground. Until better data emerges – ideally measuring outcomes over longer durations, and in different user communities – attempts to align Social AI will be operating under conditions of extreme uncertainty.

A third worry concerns how to navigate conflicts between different desirable outcomes for users of Social AI. Perhaps the clearest case (and one that arises more widely in ethical design) concerns conflicts between users' autonomy and well-being. If a user desires to have interactions with a Social AI system that may not be in their long-term interest such as seeking encouragement for self-destructive behaviour or having a sympathetic ear for radical political views, how should we balance potential harms with allowing the user to have the kinds of conversations they desire? In practice, ethical determinations in these situations will have to rely heavily on details of context.

Recognising these challenges, we might nonetheless ask what frameworks might be best suited to ethical development of Social AI. One such approach would be a *Principlist* one, such as the recently proposed Five Principles framework developed via the AI4People project [68]. This builds on the established Four Principles of Bioethics, namely beneficence (actively promoting good), nonmaleficence (avoiding harm), justice (ensuring fairness), and autonomy (respecting individual choice) and augments them with a further principle proprietary to AI ethics, namely Explicability (making AI understandable and accountable). Construed broadly, strict adherence to these principles in design and deployment could guard against most of the harms outlined above, with the principle of non-malevolence, for example, requiring tech companies to implement firm guardrails against Social AI encouraging suicidal ideation.

Principlist moral foundations may have an important role to play in ethical development of Social AI, but could be constructively supplemented by insight drawn from the machine ethics literature, in particular the rich discussion around artificial moral agents (AMAs) [69], [70]. The goal of building machines with ethical constraints or capable of autonomous ethical reasoning is of course a longstanding one, familiar to the public from the works of science fiction authors such as Isaac Asimov, but the field of enquiry has become more practically-engaged with the need for ethical safeguards in technologies like automated vehicles, decision-support systems, and dialogue systems.

While a detailed survey of the literature is beyond the scope of the present paper, one helpful distinction is that drawn between top-down, bottom-up, and hybrid AMAs [70]. The *top-down approach* constrains the behaviour of AMAs via established ethical theories such as utilitarianism or Kantian ethics, thereby offering clear theory-driven guidelines for algorithmic moral decision-making. However, it struggles with complex real-world applications and interpretational challenges. The *bottom-up approach*, by contrast, models human moral development through experiential learning, employing machine learning and evolutionary algorithms. While it has the virtues of adaptability and context sensitivity, it may be unpredictable and lacking in robustness. *Hybrid systems* aim to combine top-down normative governance with bottom-up contextual adaptability, arguably more closely reflecting human moral reasoning. The primary challenge for hybrid systems is a technical one, not least because adjudication of contexts in which departure from an established norm may be justified may require sophisticated forms of discretion.

We can now briefly examine how these distinctions apply to real world dialogue systems such as OpenAI's ChatGPT and Anthropic's Claude, which are perhaps best classified as primitive hybrid AMAs. While exact technical details are not public information, ChatGPT is trained with the goals of being helpful, honest, and harmless via a process of reinforcement learning from human feedback (RLHF) [71], in which its outputs are assessed by human users and subsequently used for fine-tuning [72]. Subsequent to this, it is likely that a further 'pruning' of possible outputs occurs to minimise insensitivity, falsehood, and similar ethical missteps. Claude follows a slightly different approach, making use of a form of Reinforcement Learning From AI Feedback (RLFAI) termed Constitutional AI [73]. Simplifying somewhat, this involves training the model to correctly classify appropriate and inappropriate outputs with reference to a set of principles (hence Constitutional). The resulting 'appropriate' dataset is then used for fine-tuning to produce a helpful model.

These processes are imperfect, as demonstrated by instances where GPT-series models have been jailbroken [67] or given inappropriate advice [74]. Nonetheless, the significant improvement in ethical performance and safety between early- and launch-versions of GPT-4 [75] suggests that hybrid techniques such as those mentioned above are advancing progress towards more trustworthy AMAs, and might serve as a technical foundation for training Social AI systems that avoid egregious ethical failings.

A final source of guidance for developing more ethical Social AI may come from approaches in behavioural science and human-computer interaction. While a detailed discussion of these is again beyond the scope of the present paper, one such promising framework may be Self-Determination Theory (SDT), an empirically-grounded paradigm developed by Ryan and Deci [76] for understanding and promoting flourishing. In short, SDT identifies three core psychological needs essential to well-being, namely autonomy (ensuring one's actions are voluntary and align with core values and goals), competence (feelings of skill and proficiency), and relatedness (feeling connected to others).

These three core psychological needs could serve as useful guiding lights for ethical development of Social AI systems, not least because their firm empirical and theoretical grounding may make them readily applicable. While SDT was not developed primarily with human-computer interaction in mind, there been extensive work in applying it to these areas, for example via METUX model (Motivation, Engagement, and Thriving in User Experience) [77]. Though grounded in SDT, METUX adds further nuance by distinguishing six “spheres of technology experience” through which technology can influence human well-being, namely *Adoption* (pre-use experiences and the motivations driving a person's technology choices), *Interface* (users’ interactions with the software's design), *Task* (how specific technology-supported activities can provide varying need satisfaction), *Behavior* (need fulfillment within the broader goal-oriented behavior supported by the technology), *Life* (how the technology benefits autonomy and well-being within an individual's life), and *Society* (how the well-being of society as a whole may be influenced by individuals’ use of the technology). Taken together, these six spheres could contribute to a systematic design framework for development of ethical Social AI (a simple proof-of-concept demonstration is included in Fig. 2 below).

Fig. 2 – Ethical Social AI development in the METUX framework

Sphere	**Autonomy**	**Competence**	**Relatedness**
Adoption	Minimising peer-pressure effects and preventing users feeling coerced into adopting Social AI systems	Ensuring individuals are not excluded from adopting Social AI due to technical or accessibility barriers	Creating informational ecosystems to allow users to make informed choices about adopting Social AI
Interface	Designing interfaces to allow control and customization of interactions to reflect individual preferences	Ensuring the interface is intuitive and user-friendly, enhancing users’ confidence and ability to converse	Building dialogue systems that facilitate meaningful conversations rather than superficial interactions
Task	Providing choices in interactions that allow users to pursue dialogues aligned with values and interests	Designing social tasks that offer opportunity for learning, enhancing social skill development	Offering users activities that allow users to connect with other humans across Social AI platforms
Behaviour	Giving users ownership of data and understanding of the data retained in Social AI interactions	Enabling users to track and understand their relationship with the AI and control its development	Helping users to set limits and find a balance between AI- and human-interactions
Life	Helping users to avoid becoming emotionally dependent on Social AI interactions	Preventing social deskilling and facilitating emotional learning	Contributing to a sense of personal growth and life-long learning through interaction with AI companions
Society	Ensuring that Social AI avoids homogenization of thoughts and behaviours in society	Fostering a society that is informed about Social AI with collective norms that guide its use	Ensuring that Social AI does not diminish connectedness and inclusivity in human communities

Conclusion

This paper has had three main goals. First, I have sought to provide some background to the rapidly emerging field of Social AI, and to offer frameworks for classifying different social AI systems (Section 1) and for understanding users anthropomorphising responses to them (Section 2). Second, I presented what I take to be some of the primary real and potential ethical concerns arising from their adoption and use (Section 3). Finally, I presented a high-level overview of possible ethical design frameworks that might aid in mitigating these harms at the level of development.

It is likely that many readers will regard Social AI companions as disturbing and even dystopian, a technological development to be avoided if possible. I acknowledge these entirely legitimate sentiments, and in some cases drastic action may be required from legislators or regulatory bodies to prevent harms, something not discussed at length here. However, the growing popularity of Social AI systems makes salient the need for greater engagement from the AI and technology ethics communities so as to ensure that where it is developed and deployed, the interests of users and society at large are given priority.

Acknowledgments

Funding: This work was supported by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067.

Conflicts of Interests

None

References

[1] 'Better language models and their implications'. Accessed: Feb. 11, 2024. [Online]. Available: https://openai.com/research/better-language-models#sample1

[2] K. Hu and K. Hu, 'ChatGPT sets record for fastest-growing user base - analyst note', *Reuters*, Feb. 02, 2023. Accessed: Nov. 07, 2023. [Online]. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[3] S. Jackson, 'Nearly 70% of people using ChatGPT at work haven't told their bosses about it, survey finds', Business Insider. Accessed: Nov. 07, 2023. [Online]. Available: https://www.businessinsider.com/70-of-people-using-chatgpt-at-work-havent-told-bosses-2023-3

[4] S. Porsdam Mann *et al.*, 'Generative AI entails a credit–blame asymmetry', *Nat. Mach. Intell.*, vol. 5, no. 5, Art. no. 5, May 2023, doi: 10.1038/s42256-023-00653-1.

[5] Hatzius Jan, Briggs Joseph, Kodnani Devesh, and Pierdomenico Giovanni, 'The Potentially Large Effects of Artificial Intelligence on Economic Growth', Goldman Sachs, Mar. 2023.

[6] 'The AI companions you can have conversations with', *BBC News*, Feb. 08, 2024. Accessed: Feb. 10, 2024. [Online]. Available: https://www.bbc.com/news/business-68165762

[7] Admin, 'How many people are Using AI Girlfriends in 2024?', AIcat.fish. Accessed: Feb. 11, 2024. [Online]. Available: https://aicat.fish/guides/how-many-people-are-using-ai-girlfriends/

[8] H. Shevlin and M. Halina, 'Apply rich psychological terms in AI with care', *Nat. Mach. Intell.*, vol. 1, no. 4, pp. 165–167, Apr. 2019, doi: 10.1038/s42256-019-0039-y.

[9] A. Chan *et al.*, 'Harms from Increasingly Agentic Algorithmic Systems', in *2023 ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2023, pp. 651–666. doi: 10.1145/3593013.3594033.

[10] A. Mayor, *Gods and robots: myths, machines, and ancient dreams of technology*. Princeton: Princeton University Press, 2018.

[11] J. Weizenbaum, 'ELIZA — a computer program for the study of natural language communication between man and machine', *Commun. ACM*, vol. 26, no. 1, pp. 23–28, Jan. 1983, doi: 10.1145/357980.357991.

[12] H. Shah and K. Warwick, 'Emotion in the Turing Test: A Downward Trend for Machines in Recent Loebner Prizes', in *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, IGI Global, 2009, pp. 325–349. doi: 10.4018/978-1-60566-354-8.ch017.

[13] A. Vaswani *et al.*, 'Attention is All you Need', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Nov. 07, 2023. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[14] M. Hua and R. Raley, 'Playing With Unicorns: AI Dungeon and Citizen NLP. | DHQ: Digital Humanities Quarterly | EBSCOhost'. Accessed: Feb. 11, 2024. [Online]. Available: https://openurl.ebsco.com/contentitem/gcd:148403435?sid=ebsco:plink:crawler&id=ebsco:gcd:148403435

[15] D. Jannai, A. Meron, B. Lenz, Y. Levine, and Y. Shoham, 'Human or Not? A Gamified Approach to the Turing Test'. arXiv, May 31, 2023. doi: 10.48550/arXiv.2305.20010.

[16] M. Burtsev, M. Reeves, and A. Job, 'The Working Limitations of Large Language Models', *MIT Sloan Manag. Rev.*, vol. 65, no. 1, pp. 1–5, Fall 2023.

[17] blaked, 'How it feels to have your mind hacked by an AI', Accessed: Feb. 09, 2024. [Online]. Available: https://www.lesswrong.com/posts/9kQFure4hdDmRBNdH/how-it-feels-to-have-your-mind-hacked-by-an-ai

[18] N. Köbis, J.-F. Bonnefon, and I. Rahwan, 'Bad machines corrupt good morals', *Nat. Hum. Behav.*, vol. 5, no. 6, Art. no. 6, Jun. 2021, doi: 10.1038/s41562-021-01128-2.

[19] B. Weiss and A. Sternlicht, 'Meta and OpenAI have spawned a wave of AI sex companions—and some of them are children', Fortune. [Online]. Available: https://fortune.com/longform/meta-openai-uncensored-ai-companions-child-pornography/

[20] A. Salles, K. Evers, and M. Farisco, 'Anthropomorphism in AI', *AJOB Neurosci.*, vol. 11, no. 2, pp. 88–95, 2020, doi: 10.1080/21507740.2020.1740350.

[21] N. Spatola, S. Marchesi, and A. Wykowska, 'Different models of anthropomorphism across cultures and ontological limits in current frameworks the integrative framework of anthropomorphism', *Front. Robot. AI*, vol. 9, 2022, Accessed: Feb. 11, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2022.863319

[22] A. Zimmerman, J. Janhonen, and E. Beer, 'Human/AI relationships: challenges, downsides, and impacts on human/human relationships', *AI Ethics*, Oct. 2023, doi: 10.1007/s43681-023-00348-8.

[23] R. Chaturvedi, S. Verma, R. Das, and Y. K. Dwivedi, 'Social companionship with artificial intelligence: Recent trends and future avenues', *Technol. Forecast. Soc. Change*, vol. 193, p. 122634, Aug. 2023, doi: 10.1016/j.techfore.2023.122634.

[24] S. J. Mithen, *The prehistory of the mind: the cognitive origins of art, religion and science*, 1st paperback ed. London New York: Thames and Hudson, 1999.

[25] M.-C. Harrison, 'The Paradox of Fiction and the Ethics of Empathy: Reconceiving Dickens's Realism', *Narrative*, vol. 16, no. 3, pp. 256–278, 2008.

[26] S. Nichols and S. P. Stich, *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds*. in Oxford cognitive science series. Oxford : New York: Clarendon ; Oxford ; Oxford University Press, 2003.

[27] N. Tiku, 'The Google engineer who thinks the company's AI has come to life', Washington Post. Accessed: Nov. 07, 2023. [Online]. Available: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

[28] C. Colombatto and S. Fleming, 'Folk Psychological Attributions of Consciousness to Large Language Models', Feb. 2024, doi: 10.31234/osf.io/5cnrv.

[29] H. Shevlin, 'Non-Human Consciousness and the Specificity Problem: A Modest Theoretical Proposal', *Mind Lang.*, vol. 36, no. 2, pp. 297–314, 2021, doi: 10.1111/mila.12338.

[30] D. J. Chalmers, 'Could a Large Language Model be Conscious?' arXiv, Apr. 29, 2023. doi: 10.48550/arXiv.2303.07103.

[31] P. Butlin *et al.*, 'Consciousness in Artificial Intelligence: Insights from the Science of Consciousness'. arXiv, Aug. 22, 2023. doi: 10.48550/arXiv.2308.08708.

[32] H. Shevlin, 'How Could We Know When a Robot was a Moral Patient?', *Camb. Q. Healthc. Ethics CQ Int. J. Healthc. Ethics Comm.*, vol. 30, no. 3, pp. 459–471, Jul. 2021, doi: 10.1017/S0963180120001012.

[33] M. Coeckelbergh, 'The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics', *Philos. Technol.*, vol. 27, no. 1, pp. 61–77, Mar. 2014, doi: 10.1007/s13347-013-0133-8.

[34] S. Turkle, 'Authenticity in the age of digital companions', *Interact. Stud. Soc. Behav. Commun. Biol. Artif. Syst.*, vol. 8, no. 3, pp. 501–517, 2007, doi: 10.1075/is.8.3.11tur.

[35] J. Danaher, 'The Philosophical Case for Robot Friendship', *J. Posthuman Stud.*, vol. 3, no. 1, pp. 5–24, Jul. 2019, doi: 10.5325/jpoststud.3.1.0005.

[36] H. Ryland, 'It's Friendship, Jim, but Not as We Know It: A Degrees-of-Friendship View of Human–Robot Friendships', *Minds Mach.*, vol. 31, no. 3, pp. 377–393, Sep. 2021, doi: 10.1007/s11023-021-09560-z.

[37] 'Building a compassionate AI friend | Replika Blog'. Accessed: Feb. 09, 2024. [Online]. Available: https://web.archive.org/web/20230329125434/https://blog.replika.com/posts/building-a-compassionate-ai-friend

[38] R. Guingrich and M. S. A. Graziano, 'Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines'. arXiv, Dec. 16, 2023. doi: 10.48550/arXiv.2311.10599.

[39] A. Xygkou *et al.*, 'The "Conversation" about Loss: Understanding How Chatbot Technology was Used in Supporting People in Grief.', in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–15. doi: 10.1145/3544548.3581154.

[40] L. Laestadius, A. Bishop, M. Gonzalez, D. Illenčík, and C. Campos-Castillo, 'Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika', *New Media Soc.*, p. 14614448221142007, Dec. 2022, doi: 10.1177/14614448221142007.

[41] A. July 31 and 2022 0 Comments, 'Is It Cheating if It's With a Chatbot? How AI Nearly Wrecked My Marriage'. Accessed: Feb. 09, 2024. [Online]. Available: https://livewire.thewire.in/livewire/chatbot-ai-nearly-wrecked-my-marriage/

[42] A. Tong and A. Tong, 'What happens when your AI chatbot stops loving you back?', *Reuters*, Mar. 21, 2023. Accessed: Nov. 07, 2023. [Online]. Available: https://www.reuters.com/technology/what-happens-when-your-ai-chatbot-stops-loving-you-back-2023-03-18/

[43] P.-F. Lovens, 'Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là', La Libre.be. Accessed: Nov. 07, 2023. [Online]. Available: https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/

[44] M. S. Correspondent Technology, 'AI chatbot blamed for Belgian man's suicide', Nov. 07, 2023. Accessed: Nov. 07, 2023. [Online]. Available: https://www.thetimes.co.uk/article/ai-chatbot-blamed-for-belgian-mans-suicide-zcjzlztcc

[45] 'R -v- Chail', Courts and Tribunals Judiciary. Accessed: Nov. 07, 2023. [Online]. Available: https://www.judiciary.uk/judgments/r-v-chail/

[46] T. Xie and I. Pentina, *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika*. 2022. Accessed: Feb. 10, 2024. [Online]. Available: http://hdl.handle.net/10125/79590

[47] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg, 'My Chatbot Companion - a Study of Human-Chatbot Relationships', *Int. J. Hum.-Comput. Stud.*, vol. 149, p. 102601, May 2021, doi: 10.1016/j.ijhcs.2021.102601.

[48] S. Vallor, 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character', *Philos. Technol.*, vol. 28, no. 1, pp. 107–124, Mar. 2015, doi: 10.1007/s13347-014-0156-9.

[49] S. Trösterer *et al.*, 'You Never Forget How to Drive: Driver Skilling and Deskilling in the Advent of Autonomous Vehicles', in *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, in Automotive'UI 16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 209–216. doi: 10.1145/3003715.3005462.

[50] I. Kant, *Lectures on Ethics*. in The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press, 1997. doi: 10.1017/CBO9781107049512.

[51] N. Burton and J. Gaskin, '"Thank You, Siri": Politeness and Intelligent Digital Assistants', *AMCIS 2019 Proc.*, Jul. 2019, [Online]. Available: https://aisel.aisnet.org/amcis2019/social_inclusion/social_inclusion/5

[52] A. Sharkey and N. Sharkey, 'We need to talk about deception in social robotics!', *Ethics Inf. Technol.*, vol. 23, no. 3, pp. 309–316, Sep. 2021, doi: 10.1007/s10676-020-09573-9.

[53] D. Susser, B. Roessler, and H. Nissenbaum, 'Online Manipulation: Hidden Influences in a Digital World'. Rochester, NY, Dec. 23, 2018. doi: 10.2139/ssrn.3306006.

[54] J. Haidt, 'The emotional dog and its rational tail: A social intuitionist approach to moral judgment', *Psychol. Rev.*, vol. 108, no. 4, pp. 814–834, 2001, doi: 10.1037/0033-295X.108.4.814.

[55] L. Floridi, ‘On Good and Evil, the Mistaken Idea That Technology Is Ever Neutral, and the Importance of the Double-Charge Thesis’, *Philos. Technol.*, vol. 36, no. 3, p. 60, Sep. 2023, doi: 10.1007/s13347-023-00661-4.

[56] C. Véliz, *Privacy is Power*. 2021. Accessed: Feb. 10, 2024. [Online]. Available: https://www.penguin.co.uk/books/442343/privacy-is-power-by-carissa-veliz/9780552177719

[57] B. C. Stahl and D. Wright, ‘Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation’, *IEEE Secur. Priv.*, vol. 16, no. 3, pp. 26–33, May 2018, doi: 10.1109/MSP.2018.2701164.

[58] ‘Privacy Policy’, Replika. Accessed: Feb. 10, 2024. [Online]. Available: https://replika.com/legal/privacy

[59] ‘ChatGPT Leaks Sensitive Data’, Spiceworks. Accessed: Feb. 10, 2024. [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/news/chatgpt-leaks-sensitive-user-data-openai-suspects-hack/

[60] C. Stokel-Walker, ‘Google was accidentally leaking its Bard AI chats into public search results’, Fast Company. Accessed: Feb. 10, 2024. [Online]. Available: https://www.fastcompany.com/90958811/google-was-accidentally-leaking-its-bard-ai-chats-into-public-search-results

[61] K. Huang, F. Zhang, Y. Li, S. Wright, V. Kidambi, and V. Manral, ‘Security and Privacy Concerns in ChatGPT’, in *Beyond AI: ChatGPT, Web3, and the Business Landscape of Tomorrow*, K. Huang, Y. Wang, F. Zhu, X. Chen, and C. Xing, Eds., in Future of Business and Finance. , Cham: Springer Nature Switzerland, 2023, pp. 297–328. doi: 10.1007/978-3-031-45282-6_11.

[62] E. Schwitzgebel, D. Schwitzgebel, and Strasser, Anna, ‘Creating a large language model of a philosopher’, doi: https://doi.org/10.1111/mila.12466.

[63] A. Holpuch, ‘German Magazine Editor Is Fired Over A.I. Michael Schumacher Interview’, *The New York Times*, Apr. 24, 2023. Accessed: Nov. 07, 2023. [Online]. Available: https://www.nytimes.com/2023/04/24/business/media/michael-schumacher-ai-fake-interview.html

[64] ‘Snapchat influencer launches an AI-powered “virtual girlfriend” to help “cure loneliness”’, NBC News. Accessed: Feb. 10, 2024. [Online]. Available: https://www.nbcnews.com/tech/ai-powered-virtual-girlfriend-caryn-marjorie-snapchat-influencer-rcna84180

[65] A. Perelli, ‘A top OnlyFans star made an AI version of herself to “date” fans in voice chats for $1 per minute. Here’s how it works.’, Business Insider. Accessed: Nov. 07, 2023. [Online]. Available: https://www.businessinsider.com/onlyfans-twitch-star-amouranth-launches-ai-version-date-fans-2023-5

[66] C. Boine, ‘Emotional Attachment to AI Companions and European Law’, *MIT Case Stud. Soc. Ethical Responsib. Comput.*, no. Winter 2023, Feb. 2023, doi: 10.21428/2c646de5.db67ec7f.

[67] J. Yu, X. Lin, Z. Yu, and X. Xing, ‘GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts’. arXiv, Oct. 04, 2023. doi: 10.48550/arXiv.2309.10253.

[68] L. Floridi *et al.*, ‘AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’, *Minds Mach.*, vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/s11023-018-9482-5.

[69] W. Wallach, ‘Robot minds and human ethics: the need for a comprehensive model of moral decision making’, *Ethics Inf. Technol.*, vol. 12, no. 3, pp. 243–250, Sep. 2010, doi: 10.1007/s10676-010-9232-8.

[70] W. Wallach and S. Vallor, ‘Moral machines: From value alignment to embodied virtue’, in *Ethics of Artificial Intelligence*, Oxford University Press, 2020, pp. 383–412. doi: 10.1093/oso/9780190905033.003.0014.

[71] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, ‘Policy Shaping: Integrating Human Feedback with Reinforcement Learning’, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2013. Accessed: Feb. 10, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/hash/e034fb6b66aacc1d48f445ddfb08da98-Abstract.html

[72] T. Wu *et al.*, ‘A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development’, *IEEECAA J. Autom. Sin.*, vol. 10, no. 5, pp. 1122–1136, May 2023, doi: 10.1109/JAS.2023.123618.

[73] Y. Bai *et al.*, ‘Constitutional AI: Harmlessness from AI Feedback’. arXiv, Dec. 15, 2022. doi: 10.48550/arXiv.2212.08073.

[74] K. Roose, ‘A Conversation With Bing’s Chatbot Left Me Deeply Unsettled’, *The New York Times*, Feb. 16, 2023. Accessed: Feb. 10, 2024. [Online]. Available: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

[75] OpenAI *et al.*, ‘GPT-4 Technical Report’. arXiv, Dec. 18, 2023. doi: 10.48550/arXiv.2303.08774.

[76] R. M. Ryan and E. L. Deci, *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. in Self-determination theory: Basic psychological needs in motivation, development, and wellness. New York, NY, US: The Guilford Press, 2017, pp. xii, 756. doi: 10.1521/978.14625/28806.

[77] R. A. Calvo, D. Peters, K. Vold, and R. M. Ryan, ‘Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry’, in *Ethics of Digital Well-Being: A Multidisciplinary Approach*, C. Burr and L. Floridi, Eds., in Philosophical Studies Series. , Cham: Springer International Publishing, 2020, pp. 31–54. doi: 10.1007/978-3-030-50585-1_2.