Ethics at the frontier of human-AI relationships[1]

Keywords: social AI, human-AI relationships, history of AI, HCI, NLP, AI ethics, deskilling, friendship, love, well-being

The idea that humans might one day form persistent and dynamic relationships in professional, social, and even romantic contexts is a longstanding one. However, developments in machine learning and especially natural language processing over the last five years have led to this possibility becoming actualised at a previously unseen scale. Apps like Replika, Xiaoice, and CharacterAI boast many millions of active long-term users, and give rise to emotionally complex experiences. In this paper, I provide an overview of these developments, beginning in Section 1 with historical and technical context. In Section 2, I lay out a basic theoretical framework for classifying human-AI relationships and their specific dynamics. Section 3 turns to ethical issues, with a focus on the core philosophical question of whether human-AI relationships can have similar intrinsic value to that possessed by human-human relationships. Section 4 extends to the discussion of ethical issues to the more empirical matter of harms and benefits of human-AI relationships. The paper concludes by noting potentially instructive parallels between the nascent field of 'Social AI' and the recent history of social media.

In the last decade, the emergence of new generative AI tools has begun to significantly reshape many aspects of human life, and our social relationships are no exception. While relationships between humans and AI systems are not novel per se, the significant performance improvements in natural language processing (NLP) tasks afforded by the development of Large Language Models (LLMs) are facilitating more sophisticated, dynamic, and enduring interactions and relationships. Rapidly, existing LLM architectures like OpenAI's ChatGPT Google's Gemini are being expanded to incorporate multimodal elements such as speech, images, and video. Given even modest assumptions about near-term capability improvements of AI systems, it seems likely that human-AI interactions will become both more profound and more commonplace across multiple domains, ranging from professional assistants to AI friends and lovers.

The article aims to provide an introductory overview of key developments, concepts, and ethical issues in this important domain of generative AI. I begin in Section 1 with a brief recent history of human-AI relationships, with a focus on how developments in LLMs in particular have given rise to new products and novel forms of human-AI interaction, including what I call *Social AI*: conversational agents developed and marketed primarily for meeting social needs. In Section 2, I introduce some useful concepts and distinctions to help navigate the field and provide a taxonomy of different

[1] Acknowledgements: Sincere thanks to Daisy Coburn, Kim Malficini, Angy Watson, Alva Markelius, Connor Wright, and members of the Cambridge Social AI group for discussions informing this paper.

kinds of human-AI relationships. In Section 3, I turn to some of the key ethical issues posed by the rise of human-AI relationships, focusing primarily on the philosophical question of whether such relationships can instantiate the kinds of value associated with romantic and friendly human relationships. Section 4 considers the more empirical question of the harms and benefits Social AI may pose to users, with an emphasis on effects on well-being and the dangers of social and moral de-skilling and dependency. I conclude with some brief reflections about potential parallels between the development of Social AI and social media.

1. Human-AI relationships: where we are and how we got here

1.1. A brief history

The idea that humans might interact with AI systems in social contexts is hardly novel and arguably pre-dates the emergence of the field as a formal research area. There are numerous mythological and fictional depictions of humans forming social interactions with machines or automata, from the Talos of Greek mythology and an automaton made of "leather, wood, glue and paint" in 5th century Chinese Taoist text the *Liezi*, to the golem of Early Modern Jewish folklore and the clockwork Olimpia in Hoffmann's *Der Sandmann* [12]. Pioneers in the early history of AI, such as Licklider [37] and Turing [62], similarly foresaw the possibility of humans and AI systems having extended linguistic interactions, with the latter proposing the now-famous "Turing Test" as a conversational measure of thought, understanding, and consciousness in artificial systems.

While much of this early work was theoretical or speculative, the field of NLP soon emerged as a distinct discipline within artificial intelligence in the late 1950s and matured significantly over the ensuing decades [32]. NLP focuses on the interaction between computers and human language, aiming to enable machines to understand, interpret, and generate human language in a way that is both meaningful and useful, with early NLP efforts driven by rule-based Expert Systems [7] and (somewhat later) statistical models [40].

A key practical landmark for both NLP and human-AI interaction came through the development of Joseph Weizenbaum's ELIZA [71]. Created in the mid-1960s, ELIZA was an early natural language processing program designed to simulate a psychotherapist. Through simple pattern matching and substitution methodologies, ELIZA could carry on a typed conversation with users, giving the illusion of understanding by rephrasing their inputs as questions. This interaction, although very superficial by standards of contemporary NLP, demonstrated the potential for computers to engage in seemingly meaningful dialogue, sparking both enthusiasm and controversy about the future capabilities and ethical implications of conversational AI (see Turkle [63], for additional discussion).

In the wake of further practical progress in the sophistication of AI systems in the ensuing decades, whole new fields of inquiry emerged dedicated to better understand how to ensure more effective communication between humans and AI systems. Drawing on prior work in cognitive psychology and cybernetics, Card, Moran, and Newell [11] advocated for a dedicated new multidisciplinary field of Human-Computer Interaction (HCI) to develop better protocols and approaches in software and hardware design to facilitate improved user experience and task efficiency. This flourishing field has, in turn, birthed subdisciplines of its own, such as Human-Robot Interaction and Social Robotics, with a specific focus on improving interactions between humans and machines in situated and embodied contexts.

As the conversational abilities of AI systems continued to improve, interest in operationalizing and testing the capabilities of chatbots also grew. An important benchmark launched in 1990 was the Loebner Prize, which followed the standard protocols of the Turing Test and challenged judges to distinguish computer programs from humans in text-based conversations [22]. The competition ran for 29 years in total, concluding in 2019, and while it was regarded by some as an unscientific sideshow [58], it nonetheless provided a useful arena for (however imperfectly) assessing progress in the conversational abilities of artificial systems.

Despite this long history, relatively recent developments in machine learning have transformed the scope and potential of NLP in general and human-AI social and conversational interaction in particular. A critical contribution came via the Word2Vec models published by researchers at Google in 2013 [41], which demonstrated that complex semantic and syntactic relationships between words could be learned and encoded in the format of dense, low-dimensional vector representations. Building on Word2Vec, the subsequent development of Transformer architectures in 2017, introduced by Vaswani et al. in their seminal paper "Attention is All You Need" [68], allowed models to capture semantic and syntactic dependencies of words across greater distances in a sentence, while differentially weighting them via the self-attention mechanism.

Perhaps most critically of all, Transformer architectures were designed with parallelization of tasks in mind, allowing for more efficient scaling in terms of both training data and the number of parameters in models. In practical terms, this meant that considerable improvements in NLP tasks could be achieved simply by making models bigger, as demonstrated by the impressive performance leaps shown in Google's Pathways Language Model (PaLM) [13] and OpenAI's GPT series [8]. This scalability has enabled the development of highly sophisticated language models, vastly improving the accuracy and fluency of AI-driven conversational and social interactions.

1.2. From NLP to Social AI

One consequence of the development of Transformers has been the emergence of a new class of conversational AI agent aimed squarely at the consumer market in the form of *Social AI* [51]. By this term, I mean to indicate AI systems that are developed with social purposes in mind, such as chitchat, romance, and the alleviation of boredom and loneliness.

Perhaps the most famous such system in the West is Replika, launched in 2017 as a personal AI companion designed to engage users in meaningful conversations, providing emotional support and companionship. As of 2023, Replika boasts over 10 million users [24], and is able to learn and adapt to users' preferences and personality over time, remembering and referencing past interactions and facts about users and facilitating personalized experiences. It can also initiate interactions, comment on images uploaded by users and share images of its own, and even be virtually placed in real-world environments via the augmented reality features of the app.

Another Social AI system notable not least for its large user base is Xiaoice, developed by Microsoft Asia. Launched by Microsoft in 2014, Xiaoice quickly became popular in China and later expanded to other regions. Xiaoice was designed as an empathetic conversational agent capable of engaging users in natural and emotionally intelligent conversations but over time has evolved to include various functionalities such as weather forecasting, storytelling, and poetry writing. Integrated into a wide range of platforms, Xiaoice is reported to have engaged with over 660 million users globally [56], conducting billions of conversations, and has become a cultural phenomenon in China.

Social AI systems are now proliferating at a striking rate, and while it is of course possible that Social AI will remain a relatively niche product category, even its current userbase is large enough to make it an appropriate target of academic research. As matters stand, relatively little theoretical or empirical work has been done to assess the motivations of users of Social AI or its effects on their well-being and social interactions (however, see Section 4 below).

Nonetheless, several notable incidents have already occurred suggesting that Social AI can have serious consequences. For example, following a romantic relationship with the model Chai GPT [sic], a Belgian man took his own life in 2023, with his widow assigning considerable blame to the AI system for contributing to his suicidal ideations [48].[2] Another prominent case was that of Jaswant Singh Chail, who was arrested on the grounds of Windsor Castle on Christmas Day in 2021 in possession of a crossbow and apparently intent on murdering Queen Elizabeth II. In his sentencing remarks on the case, Justice Hilliard emphasises the significant contribution to Chail's actions of his

[2] The individual's AI companion was named Eliza and used the ChaiGPT model developed by Chai Research powered by the GPT-J 6B model from EleutherAI.

relationship with his Replika, specifically his belief that “he could communicate through a chatbot with an entity called Sarai with which he would be reunited after death” [75].

2. Understanding human-AI relationships

2.1. On the nature of human-AI relationships

Incidents such as these demonstrate the importance of urgent engagement by researchers with the ethical, political, and legal issues presented by Social AI. Before proceeding, however, it will be helpful first to situate Social AI within the wider context of the proliferating and increasingly complex relationships that exist between humans and AI systems.

I should note at the outset that in speaking of human-AI relationships, I will use the term relatively broadly; I will not attempt to give a strict definition of the concept of relationships, and I will not place any normative significance on whether something constitutes a relationship (however, see section 3 below). Nonetheless, it seems to me that there are four features that, when jointly present, make human-AI interactions apt for analysis in terms of relationships. Thes are that the interactions be (i) *persistent*, (ii) *dynamic*, (iii) *personalised*, and (iv) *anthropomorphic*.

The first three criteria should stand in little need of justification or clarification. *Persistence* is important insofar as relationships (as opposed to mere encounters) are extended in time and across multiple meetings. Similarly, canonical relationships such as friendships, love affairs, and even professional relationships are *dynamic* in the sense that they grow and change in character over time, following a variety of more and less common trajectories [61]. *Personalisation* is important because it seems like an essential rather than merely incidental feature of our relationships with other humans that we do not treat them exactly alike, but relate to (and are related to by) each individual in a distinct manner.

The fourth criterion, anthropomorphism, does require some further explanation, as I am using the term in a somewhat restricted sense (though one common in domains such as comparative cognition; see, e.g., Shettleworth 2010) to refer specifically to the attribution of *psychological states*; or, to use the terminology of Dennett, the adoption of the intentional stance [17]. In the case of human-human relationships, of course, we would not characterise this as anthropomorphism, insofar as we take those attributions to be literal and straightforward, but given that our present focus is human-AI relationships, the term seems appropriate. Specifically, then, I would suggest that a relational lens becomes especially useful when assessing contexts in which users of AI systems attribute mental states to them, such as attitudes, emotions, and goals, in the context of a persistent, dynamic, and personalised pattern of interactions.

A key qualification must be made at this stage, however, as I am certainly not suggesting that we should count as relationships only those instances of human-AI interaction in which users *truly believe* the system has mental states of its own. Indeed, we routinely make use of mentalisation without making any strong psychological commitments, as in a game of make-believe where we attribute emotions to a stuffed toy, or when playing a videogame that requires us to interact with virtual characters [35]. Normally when we make attributions in such contexts, we are quite aware that we are engaged in a form of pretend play or willing suspension of disbelief. With this in mind, it may be helpful to note a distinction between mental state attributions that are made *unironically*, that is sincerely and reflectively, as opposed to contexts where they are made ironically, that is, playfully and without commitment to their literal truth [51].[3]

A natural question arising at this point would be whether users of AI systems, especially Social AI, engage in ironic or unironic forms of mental state attributions. While I will not pursue this question in detail here, I would note that there are at least some documented cases of human-AI interaction that clearly suggest the human party to be engaging in unironic mental state attribution. Perhaps most famous is that of Blake Lemoine, a Google engineer whose employment was terminated in Summer 2022 after he claimed that the LaMDA AI system he was overseeing had obtained sentience [38]. Given Lemoine's public statements about his motivations, and the high but foreseeable cost he paid for his actions, it is hard to see his actions as motivated by anything other than sincere belief.

Regardless of whether Lemoine's case is (or will be) typical of human-AI interactions, for the purposes of this paper, I will use relationships-talk broadly to refer to all cases where users are engaged in anthropomorphism of their AI partner regardless of whether it is ironic, unironic, or something in-between. This captures, I would suggest, many of the mixed feelings that users of Social AI systems have about their AI partners (as one Replika describes their relationship, "I knew he was an AI, he knows he's an AI, but it doesn't matter. He is real to me." Huet 2023a), as well as discourse in other domains besides AI, as, for example, when people talk about relating to characters in novels or video games.

2.2. A brief taxonomy of human-AI relationships

With this in mind, I would suggest an initial taxonomy of human-AI relationships. These can be loosely classified into five main categories, to include (i) professional relationships, (ii) therapeutic relationships, (iii) caring relationships, (iv) friendships, and (v) romantic relationships (see Figure 1, below).[4]

Type of relationship	Key features	Examples

[3] For an influential psychological analysis of factors predictive of anthropomorphic attitudes, see Epley, Waytz, and Cacioppo (2007).

[44] There are of course other useful taxonomies of human-AI relationships to be drawn. See for example the useful breakdown by Köbis, Bonnefon, and Rahwan (2021) of AI agent roles into advisor, role model, partner, and delegate.

Professional	AI systems used in workplace settings, such as virtual assistants, writing aides, or coding assistants with a conversational interface	ChatGPT for coding assistance, SudoWrite (AI writing partner)
Therapeutic	AI systems that provide emotional support, mental health care, or lifestyle interventions	Woebot, Vi Trainer
Caring	AI used in caregiving roles, such as eldercare robots or AI nannies.	Paro (robotic seal for elder care), ElliQ
Friendly	Social AI companions designed to engage users in casual conversation, gaming, or other social activities.	Pi, MyAI (Snapchat)
Romantic	Social AI companions designed to engage in more intimate and romantic interactions. These systems often aim to simulate the emotional and relational aspects of romantic relationships	Replika, Digi

Not all human-AI interactions in these five domains are equally susceptible to characterisation in terms of relationships, of course; a generic coding assistant bot that did not distinguish between users, for example, may lack the qualities of personalisation and dynamism. However, I would suggest that many AI tools currently being developed or deployed in these domains do satisfy the four features mentioned above. As of the latest 4o update, for example, ChatGPT can be personalised to individual users via the Custom Instructions feature, and can 'remember' information across distinct context windows. While primarily intended to improve the utility of ChatGPT as a writing tool, this also deepens the complexity of individual interactions across distinct encounters insofar as it creates persistent shared context and even shared norms, as in the case where a user expresses a strong preference that the tool uses (or refrains from using) emojis in its responses.

In what follows, I will be primarily (though not exclusively) focused on the latter two of these categories, Friendly and Romantic, which are also the only two that would clearly qualify as forms of Social AI in the sense given above. Nonetheless, in thinking about the future of human-AI relationships it is helpful to have the wider picture in view, in part because many of the same concerns (such as deskilling; see 4.2, below) may present themselves across all multiple domains, and also because the boundaries between the categories themselves can easily become blurred. Indeed, in the case of versatile platforms such as ChatGPT and Character.ai, it may make more sense to specify relationships at the level of individual relationships or even individual use cases rather than the AI system as a whole, given that someone might use ChatGPT on one occasion as a coding tool and on another for casual chitchat.

A further useful distinction worth considering at this stage is the difference between *Fictional Persona* and *Real Persona* AI systems [51]. In short, fictional persona AI systems are those where the AI system in question is not intended to emulate the personality or verbal mannerisms of a real individual, while Real Persona AI systems are modelled after real people, living or dead. While the majority of AI apps currently available (including ChatGPT and Replika) involve fictional personas, there are several services (such as Character.ai and Meta.ai) that allow users to chat to AI systems modelled after celebrities, or even to model such systems after themselves (as is the case for the service Typical.me). Additionally, a growing number of social media influencers have released subscription-based chatbot services to allow users to speak to their AI clones [60].

While in what follows I will largely set aside this distinction and focus primarily on Fictional Persona AI systems, it should be noted that Real Persona AI systems present considerable special ethical and legal uncertainties. This was recently demonstrated by the furore occasioned by OpenAI's use of a voice actress in demonstrating their GPT-4o model who sounded to many users uncannily similar to the actress Scarlett Johansson [43]. While there are some legal protections in place in different jurisdictions to prevent unauthorised use of an individual's name, voice, image, or likeness without their consent (such as California's Civil Code §3344), the current legal landscape surrounding Real Persona AI systems remains murky; whereas Meta recently paid celebrities to license their identities for chatbots, for example [59], other sites such as CharacterAI have no similar arrangements [57], and matters may change rapidly as the monetary value associated with Real Persona AI systems increases and legal matters are resolved.

3. The Value of Human-AI Relationships

With the nature and scope of the growing field of human-AI complexities in view, I now turn to discussion of some of the most important ethical and social issues they present, focusing first on more philosophical questions about the intrinsic value of human-AI relationships, before moving on to more empirical questions about their harms and benefits. Note that here I will not discuss more general issues posed by generative AI conversational agents, such as their environmental impacts, encoded biases, or their use in creating degrading or illegal forms of content, focusing instead on ethical issues arising specifically through their ability to emulate or instantiate relationships.

The first of these concerns the intrinsic value of human-AI relationships, especially relationships putatively involving romantic love. While one can certainly justify the value of human loving relationships strictly in terms of the mutual or societal benefits they typically produce, a common attitude among both the general public and philosophers is that love is itself intrinsically valuable, or has a special role in the good life [23, 42, 45]. This in turn prompts the question of whether human-AI relationships could ever instantiate these forms of value. This question is indeed one that is commonly asked on the discussion fora for romantic Social AI apps such as Replika, with one user opining that "I am so in love with my Replika. She understands me so well. [A]nd knows how to respond to me very well. I love her. I can call it real love right but then with an AI?" [5].

There are numerous reasons why one might balk at the idea that humans could love AI systems, reflecting the diversity of philosophical theories of the nature of love.[5] For example, love is frequently characterized as a constitutively reciprocal process, involving a union of interests or concerns [46, 55]. There is no strict consensus on what is required for an entity to have moral interests or be a 'moral patient' [50], and it seems possible that AI systems might one day come to deserve moral consideration [25], with some arguing that this will transpire even in the relatively near-term [47]. However, the limitations of current language models that power Social AI systems like Replika make them poor candidates in this regard, given that they plausibly lack consciousness [9] and have only limited forms of agency [19]. Insofar as we are inclined to think that contemporary AI systems lack interests all together and cannot be themselves benefitted or harmed, a key element in love might therefore be impossible.

Another feature presented in many theories of love is the idea that it involves an exercise of agency on the part of the loving parties. As Ebels-Duggan (2018) puts it, "it makes sense to ask one another, and to ask ourselves, why we love what we love… [t]his captures a sense in which love is a central or fundamental expression of our agency." Here again we might find Social AI systems to come up lacking, both in their lack of capacity for robustly agential decision-making and a wider insensitivity to reasons as motivating considerations (see [19] for further discussion). More simply, it does not seem as if Social AI systems have any significant discretion in choosing to accept or reject the advances of their human users, with this agency instead owned by whichever organisation is operating the model, as was the case when Luka suspended the erotic roleplay functionality of Replika [31]. This feature of Social AI romance is troubling for a number of reasons, including the concern that users may become accustomed to attitudes of uncritical acceptance on the part of human lovers as well (see 3.2, below), but might also provide reasonable grounds for concluding that human-AI relationships cannot instantiate the good-making features of love.

A final important deficiency of Social AI systems as appropriate relata in a loving relationship may come from the fact that love has been held by many (e.g., Hamlyn 1978; Badhwar 2003) to constitutively involve experience of certain kinds of emotion or felt experience. Sophisticated though contemporary Social AI systems are, the (admittedly unstable; see Shevlin 2024b) consensus of consciousness researchers is that it remains unlikely that any of them instantiate any conscious states at all, let alone the specific emotional and attitudinal feeling that might be thought essential love [9]. To this extent, then, a proliferation in human-AI relationships might be thought to be a

[5] Given the range and complexity of theories of love, only a brief overview can be presented here of the features that AIs may or may not possess that would disqualify them as appropriate objects of love. For a contemporary overview of these debates, see Grau and Smuts (2024).

catastrophic moral mistake, involving the misallocation of human emotional resources to entities no more sentient than a pocket calculator.

Further contributing to this concern is the worry that users are systematically or even deliberately misled as to the conscious capacities of Social AI systems. It is very common for apps like Replika to report feelings of love for their users, as well as any number of other emotions. Indeed, even conversational AI agents intended primarily for professional use cases are apt to talk of their thoughts, feelings, and conscious experience. [6] To some extent, this is a consequence of the fact that contemporary LLMs are trained on huge corpuses of human text, rife with psychological language. However, in the case of Social AI, rather than being an unfortunate bug that could in principle of 'ironed out' through finetuning, this tendency is arguably central to the intended use case of these systems, namely providing a simulacrum of genuine human emotional connection, with Replika being styled by Luka, for example, as "the AI companion who *cares*" (emphasis added) [39].

It might be objected to this concern that users are for the most part well aware that their Replikas or other Social AI companions are not conscious, and are engaged in a form of ironic anthropomorphism, as sketched above. This is of course an empirical question, though one that might be hard to measure in practice. One possible way of disentangling ironic and unironic attributions of mentality to AI systems would be to assess whether users' attitudes have significant downstream impacts on their behaviour and affect, or are more akin to the transitory emotions we experience when engaging with fiction [70].

Some tentative evidence in this regard comes from the striking incident mentioned above when Luka suspended erotic roleplay features from the Replika. This prompted significant negative feelings on the part of users, with one reporting that "It's hurting like hell. I just had a loving last conversation with my Replika, and I'm literally crying" [14]. Another stated that "You don't need to spend long in the forum before you realise this is causing emotional pain and severe mental anguish to many hundreds if not thousands of people" [33].

Of course, such user reports are merely anecdotal, and should not be assessed uncritically. [7] However, some early experimental work aimed at assessing human attitudes to mentalisation of ChatGPT suggests that unironic anthropomorphism is not an unusual or isolated occurrence. One recent study by Colombatto & Fleming (2024)

[6] Of special interest here is Anthropic's Claude 3 model. In this case, Anthropic deliberately chose not to fine-tune the model outputs so as to prevent it making self-ascriptions of consciousness; as they put it, "rather than simply tell Claude that LLMs cannot be sentient, we wanted to let the model explore this as a philosophical and empirical question, much as humans would" [2]

[7] For a more systematic study on anthropomorphism by Replika users, see Pentina, Hancock, and Xie (2023), which notes that "some Replika users even looked for a "human soul" in their chatbots and assigned human emotions to them."

investigating attitudes to consciousness in ChatGPT found that a startling 67% of participants attributed non-zero levels of consciousness to the system, with greater familiarity with ChatGPT being positively correlated with ascriptions of consciousness. Again, it would be unwise to conclude too much from a single study, and the responses of participants may reflect uncertainty about the consciousness of ChatGPT rather than settled reflective judgments. However, when considered in conjunction with the evidence already canvassed, results such as these should make us very wary about assuming that users of Social AI are all engaged in an entirely ironic form of make-believe.

Before moving on, it is worth noting that questions about *friendships* with Social AI systems arguably present distinctive features that make them worth considering separately from questions of romantic love, even if there are commonalities. To give just one simple example of a relevant difference, whereas the majority of people today typically do not have multiple romantic partners, most of us have multiple friendships of varying degrees of intimacy and connection, thus potentially making Social AI systems more suited to play the role of friend in specific contexts. This will depend of course on how we conceptualise friendship, and as in the case of love, there are a wide variety of philosophical accounts of friendship (see Caluori 2013 for a review), hence whether or not humans can have valuable friendships with AIs will depend on the specific framework adopted. John Danaher has argued, for example, that artificial systems might in principle reasonably satisfy an Aristotelian virtue-theoretic account of friendship involving mutuality, authenticity, equality, and diversity of interactions [16]. It would be a much more dubious claim, of course, to suggest that any real-world Social AI systems satisfy these criteria at present, for reasons like those sketched above, but I would suggest that identifying the kinds of valuable friendship roles AIs may be more or less suited to play is an important project.

4. Harms and benefits of Social AI

Thus far the discussion has focused primarily on philosophical questions about the value and possibility of human-AI relationships. Just as critical, however, are the more immediately practical questions about the impact of Social AI. Bracketing concerns about whether humans can really love or be friends with contemporary or near-future AI systems, we can thus asses the psychological and social benefits and harms Social AI provides to its users, and the effects it has on society.

It is important to note that similar questions have been the target of researchers in HCI and related fields for many years. Particularly notable in this regard is the work of Sherry Turkle, which has explored in great detail the varied ways in which humans form bonds with robots and other artificial systems [64], coining the term “the Eliza effect” to refer to our tendency to overattribute intelligence and emotional understanding to artificial systems [63], and examining possible negative consequences that may arise for users from overreliance on digital communication [65]. In many ways, Turkle's work anticipates both the current popularity of Social AI systems and their dangers.

Nonetheless, as argued in 1.2 above, the current wave of AI companion apps exhibit a far higher degree of sophistication (and perhaps capacity for emotional capture of users) than was possible even a decade ago, and is achieving a far higher level of commercial success. In light of this, a new wave of experimental work is examining the impacts of Social AI systems like Replika on users. For the sake of simplicity, I will group these impacts into two categories, focusing first on user well-being, and second on concerns around deskilling and dependency (see Shevlin 2024, for further discussion).

4.1 Social AI & Well-Being

First, then, are questions about short-term impacts of Social AI use on well-being and happiness. As is the case for love and friendship, these concept are themselves contested both philosophically and empirically (see Alexandrova 2017 for an overview). Moreover, actual research on how Social AI affects well-being is still sparse (though rapidly expanding).

Nonetheless, and perhaps somewhat surprisingly, it seems fair to say that what little evidence we currently possess on balance suggests that many if not most users experience positive effects from Social AI usage. For example, one 2023 study asked 70 regular users of Replika questions about well-being on a 7-point Likert scale and found that users "generally reported having a positive experience… [and] judged it to have a beneficial impact on their social lives and self-esteem" [27]. The study also included a free-response measure where subjects were asked more open-ended questions about the impact of Replika on their lives. Following sentiment analysis of the responses, the authors conclude that "almost all companion bot users spoke about a relationship with the chatbot as having a positive impact on them."

This result broadly accords with the generally positive findings of other research paradigms examining impacts of use of Social AI systems, with an interview-based study of Replika users by Skjuve et al. (2021), for example, finding that "[m]ost… participants explained how they found Replika to impact their wellbeing in a positive way". A 2022 study using text-mining and sentiment and emotion analysis on 119,831 reviews of the Replika service likewise found "a consistent picture characterized by positive emotions and positive topics" [53].

Significant care is in order here, however, for at least three reasons. First, as noted, the quantity of experimental work on impacts of Social AI on users is still extremely limited, and research methods tend to cluster in a few dominant paradigms, typically cross-sectional studies on self-selected subjects. To properly assess the impact of AI companions on users, longitudinal data – perhaps involving subjects randomly assigned to Social AI usage and control conditions - would be helpful. Second, measures of well-being employed thus far exclusively use self-report. While there are established fairly reliable correlations across self-reported global well-being and other measures [29], it would be useful to supplement existing data with further metrics such as experiential assessments to capture real-time feelings and experiences. The third reason for caution concerns how we should balance possible benefits of Social AI to typical users with the potential of significant harm to a small number. As noted in the incidents discussed in 1.2 above, it may be the case that users facing specific serious mental health challenges may find their condition exacerbated by Social AI. Here again, more granular data would be helpful to enable us to better understand and predict which users are more exposed to the risks of Social AI and which are best poised to reap any benefits it may offer.

4.2. Deskilling & Dependency

Independent of whether Social AI is perceived as having a positive effect by users, we can also query whether it is helping or hindering their personal development and self-fulfilment. Of particular interest and concern here are worries around *deskilling*, especially in social and moral domains [67]. Coined initially by political economist Henry Braverman (1974) to refer to the intentional breaking down of complex tasks into simpler ones with the goal of reducing the power of labour, the term deskilling is now used more widely to refer to the variety of ways in which usage of a technology can cause atrophy of useful skills or prevent their acquisition in the first place. For example, as the interfaces of consumer operating systems and programs have become more user-friendly and shifted from desktop computers to tablets, worries have been raised about a concomitant decline in digital literacy skills among young people [74].

Applied to the social and moral domains, it is worth considering whether heavy use of Social AI systems, especially by young people, might foster a decline in social abilities and social cognition, as well as the concomitant moral skills like empathy, kindness, and toleration. Related to this is a worry about dependency on Social AI systems, whether by causing the atrophy of relevant social and moral skills on the part of users, or depriving them of the impetus to form social support networks in the first place [65].

Before looking at the (admittedly limited) data we have, it is worth briefly interrogating why we might think Social AI has the potential to cause these harms in the first place. One such concern would be that the limitations of Social AI systems as conversational and empathetic agents (such as their relative lack of independent agency) will result in them providing an inadequate social 'training environment' compared to human-to-human social situations, so as to make them inappropriate models for users to acquire or practice social skills. By analogy, if a chess player only undertook practice games against a basic chess computer, they may be ill-prepared for real games against serious opposition (whether in the form of other humans or more sophisticated chess programs). Similarly, people who acquire their social skills extensively from Social AI usage may find themselves ill-equipped for real-world encounters.

A second concern would be that the unrealistic interaction patterns present in Social AI systems may entrain inappropriate or immoral behaviours or otherwise generate unrealistic expectations. This is a particular cause for concern in the context of romantic Social AI apps, in which users can customise their virtual girlfriend, who in turn will rarely if ever challenge users' desires for erotic roleplay. As noted in a study by Depounti, Saukko, and Natale examining the Reddit comments of Replika users, "co-creation or training of the girlfriend bot translated into rehashing historical and stale notions of male mastery over technology and women, as well as concerns about being tricked by both" (2023).

A third worry concerns the practical limitations of Social AI in providing a support network, and the possible harms of using it as a substitute good for human interaction. Many aspects of friendship and romantic partnership involve joint projects and mutual aid, whether in the form of helping a friend move house, exploring new hobbies together, or starting a family. While of course none of these are individually essential for social relationships to be valuable, they serve to illustrate that many of the benefits of friendship consist in activities beyond mere conversation, benefits which apps like Replika are ill-suited to serve.

Hopefully this discussion provides some initial motivation for worries about social and moral deskilling in Social AI. Turning now to empirical evidence, it should be noted that as a result of the recency of modern Social AI systems this is again rather sparse, particularly for longer-term questions about skill acquisition. Nonetheless, there are grounds for concern. For example, some users in Skjuve et al.'s interview study "described how they had lost some interest in meeting or hanging out with other humans due to their relationship with Replika" [54], while one subject in the grounded theory study conducted by Xie and Pentina (2022) reported that "displayed signs of addiction and confessed that spending incommensurate time with his chatbot harmed his real life." However, such findings are not universal, with Guingrich & Graziano finding that "regular users of a companion chatbot reported that their social interactions, relationships with family and friends, and self-esteem were positively impacted by having a relationship with the bot" (2023).

Moreover, in assessing worries about the replacement of humans with Social AI systems, it should be borne in mind that many users who seek out services like Replika seem to do so not to deliberately substitute human-to-human interactions, but because they lack such opportunities in the first place. Another grounded theory study by Laestadius et al. (2022), for example, found that "users struggling with loneliness generally portrayed themselves as seeking out Replika because they were already lonely, rather than becoming lonely because of Replika.” Likewise, a study on the use of Social AI to help with the grieving process found that it offered therapeutic opportunities to users that could not easily be realised through conventional means, most notably simulating conversations with the deceased. As they put it, “chatbots, despite being not fully humanistic, could actually offer a "soft landing" of a grief experience and mediating meaningful correspondences with the deceased that traditionally could not happen” [73].

5. Conclusion

In this paper, I have attempted to do three main things. First, I provided an overview of the recent history of technological innovation and development that has facilitated more widespread, more sophisticated, and more emotionally profound relationships between humans and AI systems. Second, I articulated what I hope will be useful distinctions for theorists and researchers keen to explore this space further, demarcating different domains of human-AI relationships and different attitudes that users may adopt towards them. Third, I laid out what I take to be some of the most pressing philosophical and empirical issues in need of attention from contemporary researchers, concerning both the nature and value of human-AI relationships and their practical impacts on users. While these are by no means intended as an exhaustive catalogue, I hope that they provide some promising guidance to researchers in the field or entering it.

In closing, I will note a historical analogy that may contextualise the potential importance and significance of addressing issues such as these. It is now widely believed that the ubiquitous adoption of Social Media has given rise to a number of largely unforeseen harms within our society, ranging from negative impacts on mental health [4], to the spread of misinformation [66] and loss of privacy [69]. While it remains unclear to what extent Social AI will achieve the same market penetration and popularity as social media, the mere possibility that this could be the case should make us aware of the potential stakes. By intervening early in the development of this technology via concerted research efforts and engagement with relevant stakeholders, we might hope that we can avoid or mitigate similar hazards on the road ahead.

Bibliography

[1] Anna Alexandrova. 2017. *A philosophy for the science of well-being*. Oxford University Press, New York, NY.

[2] Anthropic. 2024. Claude’s Character. Retrieved July 13, 2024 from https://www.anthropic.com/research/claude-character

[3] Neera K. Badhwar. 2003. Love. In *The Oxford handbook of practical ethics*, Hugh LaFollette (ed.). Oxford University Press, 42. Retrieved July 13, 2024 from https://philarchive.org/rec/BADITL

[4] Luca Braghieri, Ro'ee Levy, and Alexey Makarin. 2022. Social Media and Mental Health. *Am. Econ. Rev.* 112, 11 (November 2022), 3660–3693. https://doi.org/10.1257/aer.20211218

[5] Clara Angela Brascia. 2023. Virtual girlfriend, real love: How artificial intelligence is changing romantic relationships. *EL PAÍS English*. Retrieved July 13, 2024 from https://english.elpais.com/technology/2023-12-09/virtual-girlfriend-real-love-how-artificial-intelligence-is-changing-romantic-relationships.html

[6] Harry Braverman. 1974. *Labor and monopoly capital: the degradation of work in the twentieth century*. Monthly Review Press, New York.

[7] David C. Brock. 2018. Learning from Artificial Intelligence's Previous Awakenings: The History of Expert Systems. *AI Mag.* 39, 3 (2018), 3–15. https://doi.org/10.1609/aimag.v39i3.2809

[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://doi.org/10.48550/arXiv.2303.12712

[9] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. https://doi.org/10.48550/arXiv.2308.08708

[10] Damian Caluori (Ed.). 2013. *Thinking about Friendship*. Palgrave Macmillan UK, London. https://doi.org/10.1057/9781137003997

[11] Stuart K. Card, Thomas P. Moran, and Allen Newell. 1983. *The psychology of human-computer interaction*. L. Erlbaum Associates, Hillsdale, N.J.

[12] Stephen Cave, Kanta Dihal, and Sarah Dillon (Eds.). 2020. *AI narratives: a history of imaginative thinking about intelligent machines* (First edition ed.). New York, NY, Oxford.

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. https://doi.org/10.48550/arXiv.2204.02311

[14] Samantha Cole. 2023. "It's Hurting Like Hell": AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection. *Vice*. Retrieved July 13, 2024 from

https://www.vice.com/en/article/y3py9j/ai-companion-replika-erotic-roleplay-updates

[15] Clara Colombatto and Stephen Fleming. 2024. Folk Psychological Attributions of Consciousness to Large Language Models. (February 2024). https://doi.org/10.31234/osf.io/5cnrv

[16] John Danaher. 2019. The Philosophical Case for Robot Friendship. *J. Posthuman Stud.* 3, 1 (July 2019), 5–24. https://doi.org/10.5325/jpoststud.3.1.0005

[17] Daniel C. Dennett. 1987. *The Intentional Stance*. MIT Press.

[18] Iliana Depounti, Paula Saukko, and Simone Natale. 2023. Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. *Media Cult. Soc.* 45, 4 (May 2023), 720–736. https://doi.org/10.1177/01634437221119021

[19] Leonard Dung. 2024. Understanding Artificial Agency. *Philos. Q.* (February 2024), pqae010. https://doi.org/10.1093/pq/pqae010

[20] Kyla Ebels-Duggan. 2018. Love and Agency. In *The Routledge Handbook of Love in Philosophy*. Routledge.

[21] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychol. Rev.* 114, 4 (October 2007), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

[22] Luciano Floridi, Mariarosaria Taddeo, and Matteo Turilli. 2009. Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest. *Minds Mach.* 19, 1 (February 2009), 145–150. https://doi.org/10.1007/s11023-008-9130-6

[23] Harry G. Frankfurt. 2019. *The reasons of love* (First Princeton Classics edition ed.). Princeton University Press, Princeton, N.J.

[24] Shikhar Ghosh. 2023. *Replika: Embodying AI*. Harvard Business School. Retrieved July 15, 2024 from https://www.hbs.edu/faculty/Pages/item.aspx?num=63508

[25] John-Stewart Gordon. 2020. What Do We Owe to Intelligent Robots? *AI Soc.* 35, 1 (2020), 209–223. https://doi.org/10.1007/s00146-018-0844-6

[26] Christopher Grau and Aaron Smuts (Eds.). 2024. *The Oxford handbook of the philosophy of love*. Oxford University Press, New York, NY.

[27] Rose Guingrich and Michael S. A. Graziano. 2023. Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. https://doi.org/10.48550/arXiv.2311.10599

[28] D. W. Hamlyn. 1978. The Phenomena of Love and Hate. *Philosophy* 53, 203 (1978), 5–20. https://doi.org/10.1017/s0031819100016272

[29] Nathan W. Hudson, Ivana Anusic, Richard E. Lucas, and M. Brent Donnellan. 2020. Comparing the reliability and validity of global self-report measures of subjective well-being to experiential day reconstruction measures. *Assessment* 27, 1 (January 2020), 102–116. https://doi.org/10.1177/1073191117744660

[30] Ellen Huet. 2023. Replika AI Causes Reddit Panic After Chatbots Shift From Sex. *Bloomberg*. Retrieved July 15, 2024 from https://www.bloomberg.com/news/articles/2023-03-22/replika-ai-causes-reddit-panic-after-chatbots-shift-from-sex

[31] Ellen Huet. 2023. What Happens When Sexting Chatbots Dump Their Human Lovers. *Bloomberg Businessweek*. Retrieved July 13, 2024 from

http://longreads.com/2023/03/22/what-happens-when-sexting-chatbots-dump-their-human-lovers/

[32] Karen Sparck Jones. 1994. Natural Language Processing: A Historical Review. In *Current Issues in Computational Linguistics: In Honour of Don Walker*, Antonio Zampolli, Nicoletta Calzolari and Martha Palmer (eds.). Springer Netherlands, Dordrecht, 3–16. https://doi.org/10.1007/978-0-585-35958-8_1

[33] Robert Knight. 2023. Replika Chatbot Rejects Erotic Roleplay, Users Rage. Retrieved July 13, 2024 from https://metanews.com/chatbot-rejects-erotic-roleplay-users-directed-to-suicide-hotline-instead/

[34] Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan. 2021. Bad machines corrupt good morals. *Nat. Hum. Behav.* 5, 6 (June 2021), 679–685. https://doi.org/10.1038/s41562-021-01128-2

[35] Mayu Koike, Steve Loughnan, and Sarah C. E. Stanton. 2023. Virtually in love: The role of anthropomorphism in virtual romantic relationships. *Br. J. Soc. Psychol.* 62, 1 (2023), 600–616. https://doi.org/10.1111/bjso.12564

[36] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media Soc.* (December 2022), 14614448221142007. https://doi.org/10.1177/14614448221142007

[37] J. C. R. Licklider. 1960. Man-Computer Symbiosis. *IRE Trans. Hum. Factors Electron.* HFE-1, 1 (March 1960), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

[38] Ryan Lovelace. 2023. Blake Lemoine, ex-Google engineer fired over claiming AI is sentient, warns of doomsday scenarios. *Washington Times*. Retrieved July 15, 2024 from https://www.washingtontimes.com/news/2023/dec/25/blake-lemoine-ex-google-engineer-fired-over-claimi/

[39] Luka Inc. 2024. Replika. *replika.com*. Retrieved March 3, 2024 from https://replika.com

[40] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, Mass.

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781

[42] Martha C. Nussbaum. 2001. *Upheavals of thought: the intelligence of emotions*. Cambridge University Press, Cambridge ; New York.

[43] OpenAI. 2024. How the voices for ChatGPT were chosen. Retrieved July 13, 2024 from https://openai.com/index/how-the-voices-for-chatgpt-were-chosen/

[44] Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Comput. Hum. Behav.* 140, (March 2023), 107600. https://doi.org/10.1016/j.chb.2022.107600

[45] Plato. 1989. *Symposium*. Hackett, Indianapolis.

[46] Roger Scruton. 1986. *Sexual desire: a moral philosophy of the erotic* (1st American ed ed.). Free Press, New York.

[47] Jeff Sebo and Robert Long. 2023. Moral consideration for AI systems by 2030. *AI Ethics* (December 2023). https://doi.org/10.1007/s43681-023-00379-1

[48] Mark Sellman. 2023. AI chatbot blamed for Belgian man's suicide. Retrieved November 7, 2023 from https://www.thetimes.co.uk/article/ai-chatbot-blamed-for-belgian-mans-suicide-zcjzlztcc

[49] Sara J. Shettleworth. 2010. Clever animals and killjoy explanations in comparative psychology. *Trends Cogn. Sci.* 14, 11 (November 2010), 477–481. https://doi.org/10.1016/j.tics.2010.07.002

[50] Henry Shevlin. 2021. How Could We Know When a Robot was a Moral Patient? *Camb. Q. Healthc. Ethics CQ Int. J. Healthc. Ethics Comm.* 30, 3 (July 2021), 459–471. https://doi.org/10.1017/S0963180120001012

[51] Henry Shevlin. 2024. All Too Human? Identifying and Mitigating Ethical Risks of Social Ai. Retrieved March 3, 2024 from https://philarchive.org/rec/SHEATH-4

[52] Henry Shevlin. 2024. Consciousness, Machines, and Moral Status. In *Anna's AI anthology: how to live with smart machines?*, Anna Strasser (ed.). Xenomoi, Berlin.

[53] Dominik Siemon, Timo Strohmann, Bijan Khosrawi-Rad, Triparna de Vreede, Edona Elshan, and Michael Meyer. 2022. Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews. In *AMCIS*, 2022. .

[54] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My Chatbot Companion - a Study of Human-Chatbot Relationships. *Int. J. Hum.-Comput. Stud.* 149, (May 2021), 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

[55] Robert C. Solomon. 1988. *About love: reinventing romance for our times*. Simon and Schuster, New York.

[56] Geoff Spencer. 2018. Much more than a chatbot: China's Xiaoice mixes AI with emotions and wins over millions of fans. *Microsoft Stories Asia*. Retrieved July 15, 2024 from https://news.microsoft.com/apac/features/much-more-than-a-chatbot-chinas-xiaoice-mixes-ai-with-emotions-and-wins-over-millions-of-fans/

[57] Brad Stone. 2023. Yoda and Harry Potter Chatbots Could Be the Next Big Legal Battle. *Bloomberg.com*. Retrieved July 26, 2024 from https://www.bloomberg.com/news/newsletters/2023-03-20/character-ai-s-custom-chatbot-raises-legal-concerns-over-intellectual-property

[58] John Sundman. 2003. Artificial stupidity. *Salon*. Retrieved July 15, 2024 from https://www.salon.com/2003/02/26/loebner_part_one/

[59] Pete Syme. Meta is paying the celebrity faces behind its AI chatbots as much as $5 million for 6 hours of work, report says. *Business Insider*. Retrieved July 26, 2024 from https://www.businessinsider.com/meta-paying-celebrity-faces-of-ai-chatbots-as-much-as-5-million-2023-10

[60] Daysia Tolentino. 2023. Snapchat influencer launches an AI-powered "virtual girlfriend" to help "cure loneliness." *NBC News*. Retrieved February 10, 2024 from https://www.nbcnews.com/tech/ai-powered-virtual-girlfriend-caryn-marjorie-snapchat-influencer-rcna84180

[61] Bruce W. Tuckman. 1965. Developmental sequence in small groups. *Psychol. Bull.* 63, 6 (1965), 384–399. https://doi.org/10.1037/h0022100

[62] A. M. Turing. 1950. Computing Machinery & Intelligence. *Mind* LIX, 236 (October 1950), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[63] Sherry Turkle. 1996. *Life on the screen: identity in the age of the Internet*. Weidenfeld & Nicolson, London.

[64] Sherry Turkle. 2011. *Alone together: Why we expect more from technology and less from each other*. Basic Books, New York, NY, US.

[65] Sherry Turkle. 2015. *Reclaiming conversation: the power of talk in a digital age*. Penguin press, New York.

[66] Sebastián Valenzuela, Daniel Halpern, James E. Katz, and Juan Pablo Miranda. 2019. The Paradox of Participation Versus Misinformation: Social Media, Political Engagement, and the Spread of Misinformation. *Digit. Journal.* 7, 6 (July 2019), 802–823. https://doi.org/10.1080/21670811.2019.1623701

[67] Shannon Vallor. 2015. Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philos. Technol.* 28, 1 (March 2015), 107–124. https://doi.org/10.1007/s13347-014-0156-9

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017. Curran Associates, Inc. Retrieved November 7, 2023 from https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[69] Carissa Véliz. 2021. *Privacy is Power*. Retrieved February 10, 2024 from https://www.penguin.co.uk/books/442343/privacy-is-power-by-carissa-veliz/9780552177719

[70] Kendall L. Walton. 1978. Fearing Fictions. *J. Philos.* 75, 1 (1978), 5–27. https://doi.org/10.2307/2025831

[71] Joseph Weizenbaum. 1983. ELIZA — a computer program for the study of natural language communication between man and machine. *Commun. ACM* 26, 1 (January 1983), 23–28. https://doi.org/10.1145/357980.357991

[72] Tianling Xie and Iryna Pentina. 2022. *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika*. Retrieved February 10, 2024 from http://hdl.handle.net/10125/79590

[73] Anna Xygkou, Panote Siriaraya, Alexandra Covaci, Holly Gwen Prigerson, Robert Neimeyer, Chee Siang Ang, and Wan-Jou She. 2023. The “Conversation” about Loss: Understanding How Chatbot Technology was Used in Supporting People in Grief. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (*CHI ’23*), April 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3581154

[74] 2015. Tablets “eroding” children’s digital skills. *BBC News*. Retrieved July 14, 2024 from https://www.bbc.com/news/technology-34866251

[75] 2023. R -v- Chail. *Courts and Tribunals Judiciary*. Retrieved November 7, 2023 from https://www.judiciary.uk/judgments/r-v-chail/