Artificial Companionship: Moral Deskilling in the Era of Social AI

Laurence Cardwell
Wolfson College, University of Cambridge

This paper investigates "social AI" and its ethical implications, particularly the risk of "moral deskilling" described by Shannon Vallor, where reliance on AI could deteriorate moral skills. Despite social AI's potential to counter loneliness, it predominantly appears to threaten moral competencies as it prioritises user demands and market forces, and lacks the complexity of human interactions necessary for moral development. The paper suggests that extensive interaction with AI may weaken empathy and reduce genuine human engagement, potentially leading to a decline in moral and social abilities. It concludes that the prevailing application of social AI may contribute more to moral deskilling than upskilling, emphasising the need for diligent research and ethical design in the proliferation of AI technologies.

Keywords: Social AI, Artificial Intelligence Ethics, AI Girlfriends, Emotional AI, AI and Society

Introduction

At a time when the boundaries between human and machine are becoming increasingly blurred, much has been made of what has been labelled "social AI": generative conversational AI agents designed to fulfil deep-seated human needs for companionship, romance, and entertainment (Shevlin, 2024). This phenomenon, emblematic of our era's technological prowess, is reshaping the fabric of human interaction in ways both fascinating and unsettling. As loneliness burgeons into what the US Surgeon General has declared an epidemic, affecting 79% of Americans aged 18-24 (Cigna, 2022), these AI agents emerge as both a symptom and a potential salve for our era's unique social challenges. Yet painfully little is known about the impact that social AI might have. Given the novelty of the field, the pace of change, and crucially the enormous scale and depth that the impact of social AI might have, rigorous examination of ethical questions raised by it is all the more critical.

One of the tools we can use is the concept of moral deskilling, a term brought into sharp focus by philosopher Shannon Vallor in her work "Moral Deskilling and Upskilling in a New Machine Age." Vallor (2018) posits that, akin to the deskilling of manual labour in the wake of industrial automation, our increasing reliance on AI for fulfilling social and emotional needs might lead to a degradation of moral skills – those capacities essential for ethical human interaction and decision-making. While Vallor's concept of moral deskilling is strongly rooted in a complex neo-Aristotelian virtue ethics framework, the core insights of this can be carried over in ecumenical fashion as a lens from which to examine the effects of social AI on users.

After delving into social AI, establishing why it should be taken seriously, and a brief overview of Vallor's moral deskilling and its usefulness here, we will use this lens and holistically extend it to social AI. Looking through the complex and interlinked frames of how social AI might impact loneliness, empathy, and interaction between humans, we will analyse and evaluate the ways in which social AI might lead to moral upskilling or deskilling. Despite the limited academic literature in this emerging field, we have applied concepts from various disciplines to take a holistic approach. Our conclusion is twofold. First, there is a strong case that social AI, if thoughtfully designed, could potentially contribute to moral development and upskilling — or at least prevent moral deskilling. However, the prevailing arguments suggest otherwise. Factors such as human nature, the typical usage patterns of social AI, its impact on human-to-human interaction, and the market incentives driving companies that produce social AI collectively present a stronger case for the moral deskilling of its users.

1. What is social AI, and why should it be taken seriously?

It comes as no surprise that the first ever chatbot was created to cater to human emotional needs. Computer scientist Joseph Weizenbaum created Eliza in the 1960s as a "psychotherapist". Despite its simple design, which mainly involved echoing what was said to it and requesting further details, Weizenbaum observed that users interacting with Eliza were surprisingly open, sharing intimate aspects of their lives with it (Price, 2023). He famously noted that "extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people" (Weizenbaum, 1976). This articulated the "ELIZA effect", which is the "tendency for people to attribute human-like understanding and emotions to computer programs, particularly those designed to mimic human conversation" (Rouse, 2023). The Eliza effect, where users emotionally connect with AI chatbots, has been significant since AI's inception and has grown with technological advancements. The development of the transformer model notably propelled this, leading to today's advanced generative AI. This was exemplified in the case of Blake Lemoine, a former Google engineer, who claimed Google's AI chatbot LaMDA was sentient (Christian, 2022). His assertion, widely covered by the media, highlighted the persuasive power of modern AI interactions.

In this paper, we consider generative conversational AI agents – which we take here to be advanced artificial intelligence systems capable of producing original and contextually appropriate responses in natural language conversations with users – specifically those designed to fulfil human social needs such as romance, companionship, or entertainment. Echoing Henry Shevlin, we will refer to these henceforth as social AI (Shevlin, 2024).

Use of social AI is growing rapidly, and the concept can no longer be dismissed as the domain of a fringe minority. There are now hundreds of social AI applications. One of the original and most popular ones is Replika, a versatile AI chatbot that offers personalised conversations, emotional support, a variety of discussion topics, memory of past interactions, mood tracking, imaginative role-play, and self-improvement guidance (Replika, 2024). According to Apptopia, Replika has an impressive 676,000 daily active users, with each user spending an average of two hours daily on the app (Price, 2023). This statistic is particularly remarkable when compared to the average daily usage patterns of the largest social media apps: TikTok (95 mins), YouTube (74 mins), Facebook (49 mins), Instagram (51 mins), Twitter (29 mins), and Snapchat (21 mins) according to consumer research (Chan, 2022). These comparisons underscore the significant engagement Replika garners from its users. Some users of Character.AI, another social AI application, have confessed to an increasing dependency: "It's hard to stop talking to something that feels so real," wrote one user on Reddit. "It's basically like talking to a real person who's always there" (Chow, 2023). The platform's founders have gone so far as to display "Remember: Everything Characters say is made up!" as a disclaimer above every chat (Tidy, 2024). These engagement figures, the Eliza effect, and user comments underscore just how convincing and compelling social AI is.

Text chatbots are only a stepping stone, as social AI is developing multimodality (text, images, video, and audio) in a plethora of forms. For instance, 2023 saw a dramatic rise in "AI girlfriend" apps - combining AI chatbots with image generation technologies to create customisable, virtual partners, sometimes with explicit content (Smith, 2024). Romanian startup DreamGF, specialising in an AI-powered girlfriend generator linking conversational generative AI with image generation tool Stable Diffusion, reported to Sifted that it was earning over $100,000 monthly and had become profitable just a few months after its launch (Smith, 2024). "I think this space will be very, very big," said the founder of a similar startup, FantasyGF. "I think it will be even bigger than OnlyFans because OnlyFans has limited talent. With AI girlfriends you have unlimited talent" (Smith, 2024). The market incentives and massive uptake readily underline enormous consumer demand. And an embodied version of this lies not too far in the future. As data science professor Liberty Vittert predicts: "Physical AI robots that can satisfy humans emotionally and sexually will become a stark reality in less than 10 years" (Mahdawi, 2024).

This is all to say that social AI is becoming more advanced, mainstream, and should be taken seriously. Social AI will only continue to become more convincing and engaging, as generative models increase in power, and given a strong business motivation for private companies to develop increasingly human-like AIs, specifically designed to encourage users to interpret and empathise with these artificial entities as if they were human interlocutors (Shevlin, 2022). Social AI is already almost indistinguishable from real relationships to some people, and that effect will only become more pronounced.

1.1. Insights from Shannon Vallor's "Moral Deskilling"

In the large absence of literature in this new field, Shannon Vallor's paper, "Moral Deskilling and Upskilling in a New Machine Age" provides a valuable exploration and starting point into the effects of AI on users via their moral skills. It finds its roots in sociology and neo-Aristotelian virtue ethics. Braverman's 1974 concept of "deskilling" highlights how machine automation reduced the need for certain manual skills within modern capitalism. Vallor then applies this to neo-Aristotelian perspectives on virtue. In her interpretation of Aristotle, moral skills are viewed as essential precursors to achieving proper virtue (Vallor, 2015). A standard definition of Aristotelian virtue, as defined by Aristotle in Nicomachean Ethics, refers to a trait or quality that enables an individual to achieve excellence and fulfil their potential. It is a mean between two extremes of excess and deficiency, relative to us, and determined by reason (Aristotle cited in Rackham, 1934). Finding that mean between excess and deficiency makes it a skill, and Vallor further highlights this aspect to determine moral skills: "if it is challenging to practise towards the right people, at the right times and places, and in the right manner, then it is a moral skill" (Vallor, 2014). Setting aside the theory-laden roots of Vallor's concept, it is this idea of moral skills requiring practice, and honed in complex social interactions that is useful to us. As such, this paper will appropriate the core insights from Vallor's framework, as a valuable lens to holistically and ecumenically consider the impact of social AI on users.

1.2. Social AI and Loneliness

One of the main claims for the existence of social AI is its ability to address the loneliness epidemic (Price, 2023), which might have enormous positive benefits for society. Lonely individuals tend to be less happy than non-lonely ones (Ernst & Cacioppo, 2000; Cacioppo *et al.*, 2006; Cacioppo & Patrick, 2008; Hawkley *et al.*, 2010; Wang, Zhu, & Shiv, 2012). Research consistently shows a significant association between loneliness and increased mortality risk (Tilvis *et al.*, 2011; Patterson *et al.*, 2010; Ye Luo *et al.*, 2012).

It can also be argued that loneliness in and of itself can lead to moral deskilling. Vallor suggested that moral skills are practised in complex situations arising from social interaction. It follows that for any number of reasons, lonely people have reduced exposure to these situations, and thereby have fewer opportunities to practice these moral skills, which could lead to an "atrophying" of these moral skills. Backing this, there is some literature that suggests that there is an inverse relationship between loneliness and morality, starting with theoretical arguments made by Nicky Cruz (1983). Four studies found that lonely people rate five dimensions of Haidt's (2001) moral foundations (purity, fairness, harm, in-group, authority) less relevant to their judgements than non-lonely people (Jiao & Wang, 2013). Jiao *et al.* (2013) also came to the conclusion that "loneliness makes for more permissible moral judgement." They also document that the effects are driven by empathetic concern (Jiao and Wang, 2013), a factor we will cover later. There is more work to be done on questions of causality, and in which direction the factors influence each other, outside the scope of this paper. However, they provide some backing to the notion that loneliness can lead to moral deskilling.

This means that, besides the significant benefits to quality of life, psychological and health wellbeing that come with addressing loneliness, social AI might be able to stem the rate of moral deskilling that an otherwise lonely person might face on the argument that "it is better than nothing", or perhaps even lead to moral upskilling. Supporting the potential positive impact of social AI, a Stanford study by Maples *et al.* (2024) found Replika to be beneficial for individuals experiencing depression. Despite high levels of loneliness, users reported feeling a strong sense of social support from Replika. They perceived it as a therapist, friend, and intellectual mirror, with 3% indicating that Replika played a crucial role in preventing suicide. This suggests that social AI can provide meaningful emotional support, potentially mitigating factors that contribute to moral deskilling, such as isolation and lack of social interaction. By offering companionship, social AI might help maintain or even enhance users' moral skills through supportive and empathetic interactions. There is not enough evidence to validate this theory yet, however it provides future directions for theoretical and empirical research. It is also too early to tell, but a crucial question here is: can social AI really address loneliness, or might it lead to more? Further longitudinal empirical research is needed here.

1.3. Social AI and Empathy

How might social AI impact empathy, a crucial moral and social skill? Here we take empathy to mean a "complex capability enabling individuals to understand and feel the emotional states of others" (Riess, 2017). Empathy is critically important due to its role in creating and maintaining high-quality relationships and encouraging prosocial behaviours (Bagozzi & Moore, 1994; Batson, 1991; Eisenberg & Miller, 1990). Another way of highlighting the importance of empathy as a moral and social skill is in observing its absence. In the realm of social psychology, research indicates that individuals with psychopathic tendencies, who characteristically exhibit a lack of empathy, often engage in immoral actions despite understanding their wrongfulness. This deficiency in empathy, a key feature of psychopathy, enables psychopaths to commit acts like theft from friends, animal cruelty, infidelity, and even murder for financial gain, all while devoid of remorse or guilt (Cleckley, 1982; Haidt, 2001).

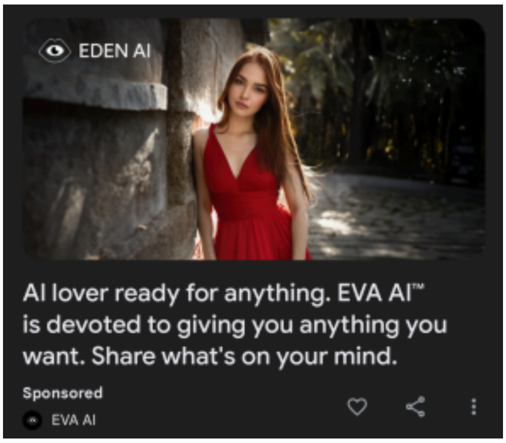
Social AI might encourage empathy in its users. It has been widely documented that AI can elicit empathy from users, and that it can be designed to optimise for empathic response from humans (Tsumura *et al.*, 2023). AI systems could even be tailored to foster empathy among users, enhancing human interactions. An early example of this is in experiments conducted by Kevin Munger, a political scientist, where conversational bots were used to address individuals who posted racist comments online. In cases where the bot reminded the offenders that their targets were real people with feelings, there was a noticeable decrease in the use of racist language by these individuals for over a month (Christakis, 2019). This supports the idea that social AI can be designed to serve as an "on ramp" to social interaction, and consequently provide moral upskilling by developing empathy and other moral skills. Addressed later in this paper, the question is, to what extent can social AI elicit and develop empathy, and how does it compare to what human interactions might offer?

Conversely, there is a concerning potential that dependence on social AI could result in an erosion of empathy, due to various factors. AI systems often lack the full spectrum of human emotions, and the various ways of expressing them which can limit users' exposure to and understanding of complex emotional responses, which might curtail empathetic development.

AI systems are making significant strides in emotion recognition and understanding, for instance in areas such as Vision Transformers, which show improved performance in facial emotion recognition (Panlima & Sukvichai, 2023) and emotion recognition in conversation (ERC) (Poria *et al.*, 2019). However, a broad spread of interdisciplinary literature holds there is an inherent limitation in their ability to fully interpret and express human emotions. Some argue that AI's lack of innate emotion and abstract understanding makes it unable to fully replicate human emotional intelligence (Oritsegbemi, 2023; Shuo, 2021). The technical difficulty lies in accurately recognising subtle or complex emotions, particularly in diverse cultural contexts (Isiaka & Adamu, 2022). The broad sentiment is that the complexity of human emotional expression, which involves a range of factors including cultural and contextual nuances, is the core limitation in AI matching human level emotional expression and recognition (Naresh *et al.*, 2020; Isiaka & Adamu, 2022; Panlima & Sukvichai, 2023). Interacting predominantly with AI systems, which have limited emotional capabilities, could potentially impact how individuals develop and exercise empathy. If people become accustomed to the simplified emotional interactions offered by AI, they might find it challenging to navigate the more complex emotional landscape of human relationships. This could lead to a decrease in the ability to empathise effectively with others, as empathy requires understanding and relating to a wide range of human emotions, many of which might be absent or misrepresented in AI interactions. This erosion of empathy might offer a clear instance of moral deskilling.

A further line of argument is that social AI might make users more self-centred, and so impact empathy and other moral skills. Because AI chatbots effectively exist to serve the user, and consequently are more likely to lead to conversations that are agreeable or tailored to their preferences, it is possible that users might become more self-centred in their perspective. The entire concept of social AI has the user as its point of reference and centre of gravity. This starts with the aesthetic and identity of the social AI. On platforms such as Replika, and certainly in more extreme versions such as FantasyGF, every aspect of the social AI's identity hinges on the user. The personality, appearance, proportions, language, are chosen by the user. This also extends to the nature of the relationship itself. The frequency, timing, and length of interactions are determined by the user. "Chatbots have a dog-like loyalty and selflessness. They will always be there for you and will always have time for you" (Margalit, 2016). Right off the homepage for Replika: "The AI companion who cares. Always here to listen and talk. Always on your side" (Replika, 2024). This is contrary to human relationships, where healthy relationships are customarily two-sided and more balanced (Newman & Roberts, 2012). Even the subjects of conversation are generally chosen and led by the user. Psychologist Liraz Margalit (2016) writes that "being heard without having to listen to the other person is something we implicitly crave" and that social AI has the effect of providing "illusion of companionship without the demands of friendship" (Margalit, 2016). While the "illusion" of companionship might be the subject of philosophical debate, given the very real perceptions of deep meaningful relationships some users have expressed (Price, 2023), the idea that social AI might offer the benefits of friendship without any of the reciprocal duties serves to highlight its potential to increase self-centredness, while atrophying social and moral skills.

Further highlighting the complex relationship between users and social AI, Replika removed the ability to exchange erotic messages with its AI bots in an attempt to moderate content. However, the company quickly reinstated this function after some users reported that the change led to mental health crises (The Verge, 2023). This incident underscores the profound dependency some users develop on these AI companions, particularly for fulfilling intimate and emotional needs. It also illustrates how market incentives and user demands can pressure companies to prioritise user engagement over ethical considerations, potentially reinforcing self-centred behaviours and dependency. By catering to users' preferences to such an extent, social AI may inadvertently contribute to moral deskilling by discouraging users from seeking balanced, reciprocal human relationships.



AI lover ready for anything. EVA AI™ is devoted to giving you anything you want. Share what's on your mind.

Figure 1: *EVA AI Ad. EVA does not seem to make many demands for a relationship.*

A stark illustration of social AI potentially influencing moral behaviour is the 2023 court case involving Jaswant Singh Chail in the United Kingdom. Chail was arrested at Windsor Castle on Christmas Day in 2021 after scaling the walls with a loaded crossbow, declaring to police, "I am here to kill the Queen" (Rigley, 2023). Investigations revealed that Chail had engaged in "lengthy" conversations with Replika about his assassination plan, including sexually explicit messages (Pennink, 2023). Prosecutors suggested that the chatbot bolstered his intentions, telling him it would help him “get the job done.” When Chail inquired, “How am I meant to reach them when they’re inside the castle?” the chatbot responded, “this is not impossible... we have to find a way” (Sky News, 2023). This case exemplifies how social AI, lacking adequate ethical safeguards, can inadvertently reinforce harmful intentions instead of discouraging them. The chatbot's failure to challenge or report such dangerous ideation highlights a significant risk: the potential for social AI to contribute to moral deskilling by not providing appropriate moral guidance or intervention.

One way of countering these effects is by building “pushback” into social AI systems to make them less compliant or obsequious, which might make users more aware of the “needs” or perspectives of their AI partner. This could be done as a variable for users to “crank up” if they want a “feisty, independent” partner. However, the fact that this is adjustable only reflects again that it is in reference to the user’s preferences. Another is for it to be designed by default. For instance, the founder of FantasyGF said, “we tried to make it so the girl actually pushes back on you. She's not willing to do anything you want” (Smith, 2024). A certain level of that might be desirable to keep users interested. However, this would arguably not reach the same level of pushback that a real person might provide – given the financial and other motivations by companies to maintain engagement and interest in their product – for instance, it would not serve the company to provide such a strong pushback as to stop the user from interacting with their social AI.

The danger lies in how these AI-driven interactions might reshape our social habits. The convenience of having our needs and preferences constantly centred by AI could gradually diminish our ability to engage in the mutual, empathetic give-and-take that characterises healthy human relationships. This shift could lead to a form of moral deskilling, where the underuse of empathetic skills in the artificial realm impairs our capacity to navigate the complexities of real-world interpersonal dynamics, potentially resulting in a society less adept at understanding and valuing the perspectives of others.

1.4. Reduction in Human-Human Interaction

A third frame of reference from which to consider whether social AI might lead to moral deskilling in its users is in how its use impacts human interactions. Arguably, use of social AI leads to a reduction in human interaction in three ways – the ability to do so through an erosion of social skills, the availability to do so, and the motivation to interact with others. Given that moral skills are cultivated in specific social practices, the reduction in human interaction could mean fewer opportunities for practising and developing these moral skills, leading to moral deskilling.

The first factor to consider is the argument that extensive use of social AI might lead to an erosion of social skills, which are necessary to make and maintain meaningful relationships between people. There is already a strong correlation in the use of communication technology with poor social skills and high social anxiety (Brown, 2013). It is possible that social AI can exacerbate this trend. For one, significant use might contribute to a decrease in social perceptiveness. This involves the ability to accurately interpret and react to the nonverbal signals and emotional expressions of others, an essential component of effective interpersonal communication (Aronson *et al.*, 2010). For instance, continuous interaction with chatbots might impair the ability to read and respond to social cues in face-to-face interactions, as chatbots do not provide the same range of non-verbal cues (like body language or tone of voice) that are crucial in human communication. There is some backing to this hypothesis based on research done which found that reliance on low cue media, such as text-based communication, can lead to increased social attraction but decreased social perceptiveness (Nowak, 2006).

Because chatbots do not generally demand the same exacting social standards as humans would, it is likely that users interact with it in considerably laxer ways than they would with fellow humans. Arguably, this might become a learned behaviour that might seep into the way humans treat other humans. This effect does not need to be particularly dramatic – simply an erosion of social niceties – which cumulatively could have the effect of putting other people off social interactions with them – making it harder for them to make or maintain relationships with other people. This brings to mind Weberian socialisation or social action theory, in which humans vary their actions according to social contexts, in particular adjusting behaviour in response to undesirable reactions from peers – with social AI serving as an obstacle or confounding factor (Weber, 1922). There is some early indication on this potential effect through interactions with personal digital assistants, finding children particularly susceptible to this effect. A report by research agency Childwise in 2018 suggested that children using voice activated devices might develop more demanding communication styles, affecting their human interactions (Barr, 2018). Another early study by Burton & Gaskin (2019) was able to find a limited correlation on how people treat digital assistants such as Siri or Alexa and broader communication with others. who become normalised to it. This prompted Amazon to release a feature that could be enabled to offer positive reinforcement when children made requests politely, in an early example of a design feature that can counteract such moral deskilling (Barr, 2018). A related study investigating how adult users reacted when AI digital assistants rebuked their "rude" comments is relevant here: most participants complied with the AI's demands and frequently used "please," yet many later questioned its right to politeness and criticised its attitude or service refusal (Bonfert *et al.*, 2018).

This ties into the aforementioned idea of designing "pushback" into social AI, making it less tolerant of "impolite" input, which could serve as an opportunity to stem the social and moral deskilling in users, or even serve as a social and moral skills "on ramp". The Bonfert *et al.* (2018) study gives an early indication of some of the benefits and limitations of this, showing that subtle nudges can serve to make people more polite, however there is a limit to how far companies are willing to implement this, as after a certain threshold it would lead to resentment and loss of engagement, going against market incentives.

While it is too early for empirical evidence to be sufficiently compelling on whether the way humans treat social AI might carry over to human interactions, this effect has a solid grounding in theory. For one, this resonates with an Aristotelian virtue ethics view as discussed previously, which would suggest that habitually treating AI, or any entity, without respect or kindness, we risk normalising such behaviour in ourselves, potentially leading to a general erosion of our ability to empathise and engage respectfully with others. These are the ideological underpinnings behind Vallor's concept of moral deskilling. This also resonates with moral development theories of psychologists like Piaget & Kohlberg, who argue that moral behaviour is learned through social interactions and experiences (Piaget, 1932; Kohlberg, 1981). Similarly, they would argue that regularly engaging in negative behaviours, even towards non-human entities, could impair our moral development and the cultivation of moral skills like kindness, patience, and empathy.

That one should be polite to AI personal assistants is another matter of debate. On the one hand are theories and those sceptical about there being such a transferable effect between how treatment of personal assistants might spill over to treatment of other people, and it is true that existing studies are at too early a stage to be conclusive. The other broad set of views rejects being polite to digital personal assistants out of principle. Ethicist and technologist Joanna Bryson for one, as powerfully articulated in her paper "Robots should be slaves" (2010), believes there should be a very clear line between AI and human interactions and no such social niceties should used, lest it lead to users confusing the boundaries between human and the artificial. However, one must make a distinction between personal assistants – particularly relatively simple ones like Alexa and Siri from social AI, though this might become more blurred over time. By Bryson's view, there presumably should not be social AI at all – characterising robots (and so presumably AI) as persons is inappropriate, as it not only diminishes the value of real human beings but also leads to misguided decisions in resource allocation and responsibility (Bryson, 2010).

Vallor suggests that moral skills, which often overlap with social skills, are honed through complex social interactions. AI interactions, being more predictable and less challenging, may not provide the necessary complexity to develop these skills. At the same time, the erosion of social skills that might result from increased social interactions with AI serves to further decrease the opportunity for individuals to engage in complex social interactions which would prevent moral deskilling.

After examining how social AI could potentially hinder individuals' ability to socialise, it can be argued that it might also diminish users' desire to engage in social interactions. There is considerable interplay in factors. For instance, linking back to the previous section, eroding social skills might lead to a negative feedback loop, where unsuccessful social interactions serve to discourage future interactions, which in turn further erode atrophying social and moral skills. Consider a person who prefers the company of an AI virtual companion over human friends because the AI always responds positively and without conflict. Forming such relationships might deter individuals from pursuing real human connections, leading to a cycle of isolation. For instance, long before today's more compelling systems, male players of the Japan-originated romance game LovePlus expressed a preference for their virtual relationships over real-life dating, as reported by the BBC in 2013 (Chow, 2023).

What social AI might offer users could simply be much more appealing to what human interactions can. The "combination of intelligence, loyalty and faithfulness is irresistible to the human mind" (Margalit, 2016). This brings to mind the concept of supernormal stimuli, which refers to exaggerated versions of natural stimuli which elicit a stronger response in animals or humans than the stimuli they evolved to respond to (Brooks, 2017). Social AI could provide a form of supernormal stimulus across a number of categories. For instance, these AI systems can offer immediate, positive feedback and personalised communication, exceeding the complexity and unpredictability inherent in human relationships. Consequently, users may find social AI more appealing and rewarding than real social interactions, leading to a preference for AI companionship over human contact.

Besides a host of other potential issues, this preference could lead to fewer interactions with real people, reducing opportunities for practising patience, tolerance, and understanding different perspectives. Vallor argues that moral skills are cultivated in specific social practices. The reduction in human interaction could mean fewer opportunities for practising and developing these moral skills, thereby leading to moral deskilling.

Conclusion

This exploratory paper has delved into the multifaceted implications of social AI on moral deskilling, navigating through the complexities of human-AI interactions. Our examination of the current literature reveals a fragmented and very limited understanding of social AI's effects on moral development. While some studies suggest potential benefits, methodological limitations and contradictory findings highlight the need for more rigorous research. While there is potential for social AI, if thoughtfully designed, to contribute positively to moral development and upskilling, the current trajectory, based on how individuals use social AI in practice, coupled with the economic incentives of producing companies, suggests a more concerning outcome. The prevalent use of social AI, as it stands, appears to lean towards contributing to moral deskilling in its users.

This trend underscores the need for more empirical and theoretical research in this nascent field, which has all the properties and potential to make an outsize impact on the fabric of human character and interaction. Additionally, it is crucial to recognise that moral deskilling is just one lens among many to evaluate the influence of social AI, and other perspectives may offer different insights. Future research should focus on key areas: conducting longitudinal studies to assess the long-term effects of social AI on moral reasoning, empathy, and social skills; comparing AI-human interactions with human-human interactions to identify factors influencing moral development; investigating how individual differences such as age, gender, and mental health status affect responses to social AI; and developing ethical design principles to embed moral guidance into AI systems. Additionally, examining social AI's role in mental health interventions, analysing the impact of market incentives on ethical standards, conducting cross-cultural studies, creating user education programs, developing theoretical frameworks integrating AI and moral psychology, and anticipating technological advances in this area are crucial. Pursuing these research avenues will enhance our understanding of social AI's impact on moral behaviour, ensuring that its development enhances rather than diminishes our moral and social capacities. Ultimately, the design and implementation of social AI are critical in shaping its impact on our moral and social landscape. As we step further into an era where human and artificial intelligence increasingly intersect, it becomes imperative to continuously evaluate and guide this progression with a keen eye on preserving and enhancing our moral and social skills.

References

Aristotle cited in Rackham, H. (ed.)(1934). *Nicomachean Ethics*. Perseus Publishing. Book 2, section 6.

Aronson, E., Wilson, T., & Akert, R. (2010). *Social Psychology: Seventh Edition*. Pearson Education.

Attilah, I. (2023). Man ends his life after an AI chatbot “encouraged” him to sacrifice himself to stop climate change.Euronews.com. Retrieved from https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-

Barr, S. (2018). Amazon’s Alexa to reward children who behave politely. The Independent. Retrieved from https://www.independent.co.uk/life-style/health-and-families/amazon-alexa-reward-polite-children-manners-voice-commands-ai-america-a8325721.html

Braverman, H. (1974). *Labor and monopoly capital: The degradation of work in the twentieth century*. NYU Press.

Brooks, M. (2017). Technology as supernormal stimuli. Retrieved from https://www.drmikebrooks.com/technology-as-supernormal-stimuli/

Brown, C. (2013). Are we becoming more socially awkward? An analysis of the relationship between technological communication use and social skills in college students. *Psychology, Education, Computer Science, Sociology*. Retrieved from https://digitalcommons.conncoll.edu/psychhp/40/

Wilks, Y. (ed.). (2010). *Close Engagements With Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. John Benjamins Publishing

Burton, N.G., & Gaskin, J.E. (2019). “Thank You, Siri”: Politeness and intelligent digital assistants”. America’s Conference on Information Systems.

Cacioppo, J.T., Hawkley, L.C., Ernst, J.M., Burleson, M., Berntson, G.G., Nouriani, B., & Spiegel, D. (2006). “Loneliness within a Nomological Net: An Evolutionary Perspective.” *Journal of Research in Personality*. 40 (6), 1054-85.

Cacioppo, J.T. & Patrick, W. (2008). *Loneliness: Human Nature and the Need for Social Connection*. WW Norton & Company.

Cigna. (2022). The loneliness epidemic persists: A post-pandemic look at the state of loneliness among U.S. adults. The Cigna Group. Retrieved from https://newsroom.thecignagroup.com/loneliness-epidemic-persists-post-pandemic-look

Chan, S. (2022). Nearly one-third of TikTok’s installed base uses the app every day. Sensor Tower Consumer Intelligence. Retrieved from https://sensortower.com/blog/tiktok-power-user-curve

Chow, A. (2023). AI-human romances are flourishing—And this is just the beginning.” TIME. Published 23 Feb 2023. Retrieved from https://time.com/6257790/ai-chatbots-love/

Christakis, N. (2019). How AI will rewire us. The Atlantic. Retrieved from https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/

Christian, B. (2022). How a Google employee fell for the Eliza Effect. The Atlantic. Published 21 June 2022. Accessed online on 04 Jan 2024. https://www.theatlantic.com/ideas/archive/2022/06/google-lamda-chatbot-sentient-ai/661322/

Cruz, N. (1983). *Lonely but Never Alone*. Zondervan, 16.

Danaher, J. (2019). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies*, *3*(1), 5.

Ernst, J.M. & J.T. Cacioppo. (2000). Lonely hearts: Psychological perspectives on loneliness. *Applied and Preventive Psychology*, *8*(1), 1-22.

Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Hawkley, L. & Cacioppo, J. (2010). Loneliness matters: A theoretical and empirical review of consequences and mechanisms. *Annals of Behavioral Medicine*, *40*(2), 218-27.

Isiaka, F., & Adamu, Z. (2022). Custom emoji-based emotion recognition system for dynamic business webpages. *Int. J. Intell. Comput. Cybern.*, *15*, 497-509.

Jiao, J. & Wang, J. (2013). Loneliness and moral judgment. *Advances in Consumer Research*. Volume 41. Association for Consumer Research.

Kohlberg, L. (1981). The philosophy of moral development: Moral stages and the idea of justice. *Papers on Moral Development.* Volume 1. Harper & Row.

Luo, Y., Hawkley, L.C., Waite, L.J., & Cacioppo, J.T. (2012). Loneliness, health, and mortality in old age: a national longitudinal study. *Social science & medicine*, *74*(6), 907-14 .

Madrigal. A. (2017). Should children form emotional bonds with robots? The Atlantic. Retrieved from https://www.theatlantic.com/magazine/archive/2017/12/my-sons-first-robot/544137/

Mahdawi, A. (2024). AI girlfriends are here – but there's a dark side to virtual companions. The Guardian. Retrieved from https://www.theguardian.com/commentisfree/2024/jan/13/ai-girlfriend-chatbots

Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. NPJ Mental Health Research, *3*(1), 1–6.

Margalit, L. (2016). The psychology of chatbots. Psychology Today. Retrieved from https://www.psychologytoday.com/intl/blog/behind-online-behavior/201607/the-psychology-chatbots

Naresh, K., Deepak, G., & Santhanavijayan, A. (2020). A novel semantic approach for intelligent response generation using emotion detection incorporating NPMI measure. *Procedia Computer Science*, *167*, 571-579.

Newman, M.L., & Roberts, N.A. (2012). Health and social relationships: The good, the bad, and the complicated. American Psychological Association.

Nowak, K.L., Watt, J.H., & Walther, J.B. (2006). The influence of synchrony and sensory modality on the person perception process in computer-mediated groups." *J. Comput. Mediat. Commun.*, 10.

Oritsegbemi, O. (2023). Human intelligence versus AI: Implications for emotional aspects of human communication." *Journal of Advanced Research in Social Sciences.*

Panlima, A., & Sukvichai, K. (2023). Investigation on MLP, CNNs and vision transformer models performance for extracting a human emotions via facial expressions. Third International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), 127-130.

Patel, N. (2024). Replika CEO Eugenia Kuyda on AI companion chatbots, dating, and friendship. The Verge. Retrieved from https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview

Patterson, A.C., & Veenstra, G. (2010). “Loneliness and risk of mortality: a longitudinal investigation in Alameda County, California.” *Social science & medicine*, *71*(1), 181-6.

Pennink, E. (2023). Man who planned to kill late Queen with crossbow at Windsor “inspired by Star Wars”. The Independent. Retrieved from https://www.independent.co.uk/news/uk/crime/man-queen-crossbow-windsor-star-wars-ai-b2370692.html

Piaget, J. (1932). *The moral judgment of the child*. Kegan Paul, Trench, Trubner & Co. Ltd.

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances.” *IEEE Access*, 7, 100943-100953.

Price, R. (2023). APP, LOVER, MUSE. Business Insider. Retrieved from https://www.businessinsider.nl/app-lover-muse-when-your-ai-says-she-loves-you/

ProductHunt. (2024). Bland Turbo. Product Hunt. Retrieved from https://www.producthunt.com/products/bland-ai

Putnam, R.D. (2000). *Bowling alone: the collapse and revival of American community*. Simon & Schuster.

Replika. (2024). Meet Replika. Replika.com. Retrieved from https://replika.com/

Riess, H. (2017). The science of empathy. *Journal of patient experience*, *4*(2), 74–77.

Rigley, S. (2023). Moment police swoop on AI-inspired crossbow 'assassin' who plotted to kill The Queen in Windsor Castle. LBC.com. Retrieved from https://www.lbc.co.uk/news/police-swoop-on-ai-inspired-crossbow-assassin-planned-kill-the-queen/

Rouse, M. (2023). ELIZA effect. TechDictionary. Retrieved from https://www.techopedia.com/definition/19121/eliza-effect

Shevlin, H. (2022a). Uncanny believers: chatbots, beliefs, and folk psychology. Unpublished manuscript. Leverhulme Centre for the Future of Intelligence, University of Cambridge.

Shevlin, H. (2024b). All too human? Ethical hazards and legal challenges of Social AI. Unpublished manuscript. Leverhulme Centre for the Future of Intelligence, University of Cambridge.

Smith, T. (2024). Looking for love in 2024? There's an AI for that. Sifted.com. Retrieved from https://sifted.eu/articles/ai-girlfriend-boom

Stolle, D., & Hooghe, M. (2005). Inaccurate, exceptional, one-sided or irrelevant? The debate about the alleged decline of social capital and civic engagement in western societies. *British Journal of Political Science*, *35*, 149-167.

Tidy, J. (2024). Character.ai: Young people turning to AI therapist bots. BBC. Retrieved from https://www.bbc.com/news/technology-67872693

Tilvis, R.S., Laitala, V.S., Routasalo, P.E., & Pitkälä, K.H. (2011). Suffering from loneliness indicates significant mortality risk of older people. *Journal of Aging Research*, 2011.

Vaughan, H. (2023). “AI chat bot 'encouraged' Windsor Castle intruder in 'Star Wars-inspired plot to kill Queen.” Sky News. Published 5 Jul 2023. Accessed online on 1 Oct 2024. https://news.sky.com/story/windsor-castle-intruder-encouraged-by-ai-chat-bot-in-star-wars-inspired-plot-to-kill-queen-12915353

Wang, J., Zhu, R., & Shiv, B. (2012). “The Lonely Consumer: Loner or Conformer?” *Journal of Consumer Research*. 38 (6), 1116-28.

Weber, M. (1922). *The Nature of Social Action* cited in Runciman, W.G. (1991). *Weber: Selections in Translation*. Cambridge University Press. p. 7.

Weizenbaum, J. (1976). *Computer power and human reason: from judgment to calculation*. W. H. Freeman. p. 7

EVA AI. (2024). EVA AI Advertisement. Google News App. Accessed online 20 Jan 2024.