Ethics and Information Technology (2026) 28:27
https://doi.org/10.1007/s10676-026-09898-x

ORIGINAL PAPER

Missing the human in AI: on dehumanisation by generative AI chatbots using the case of Replika

Paolo Monti[1] · **Christiane Grünloh**[2] · **Michelle Worthington**[3]

Abstract
Large Language Model (LLM) based chatbots are increasingly prominent in the digital landscape, promising users a range of experiences, everything from human-like conversation to intimate companionship. But where exactly is the human in such technology? And what is its impact on the humanity of those persons engaging with it? Taking as its focus the popular LLM-based chatbot Replika, this article contends that - positive use cases notwithstanding - emerging generative AI technology carries with it a novel and more subtle form of dehumanisation. We combine an autoethnographic investigation with a multi-disciplinary critique (human-computer interaction; moral philosophy; law) to argue that such technology risks dehumanisation by inviting, demanding or conditioning human emotional and attentional engagement in circumstances where reciprocity is structurally impossible. We argue further that existing conceptual, philosophical and regulatory systems will need to be adapted to help diminish or erase the risks and harms presented by these emerging technologies.

Keywords Dehumanisation · Large Language Models · Technological Frames · Replika · Regulation · Anthropomorphism · Ethics of Care · Reciprocity

Introduction

"Replika. The AI companion who cares. Always here to listen and talk. Always on your side." These claims appear on the homepage of replika.com, the official website of the Large Language Model (LLM) based service Replika. The core message is clearly built on an increasing expectation surrounding the current impressive developments in machine learning technology: that generative artificial intelligence (AI) systems have become so advanced that they are now essentially indistinguishable from us in areas of agency and discourse traditionally considered to be the exclusive domain of human subjects.

But is this the case? Where is the human in Replika? Or, in other words, what is "essentially human" in the experience of interacting with an AI system like Replika? And what is rather potentially misleading in creating and embracing the expectation of a caring human relationship with a non-human interlocutor that is unable to care?

Decades of research into anthropomorphism have shown that humans tend to ascribe human behaviours to technology and apply social rules to their interactions with technology which does not necessitate a conscious belief that the technology is human or human-like (Nass et al., 1994). According to media equation theory, people attribute agency-like features to technology while being fully aware that they are interacting with a machine (Reeves & Nass, 1996; Klowait, 2018). Even people who reject anthropomorphism have shown to "mindlessly apply social rules and expectations to computers" (Nass & Moon, 2000). There are also individual differences in anthropomorphism. For example, people who are lonely are more likely to anthropomorphize non-human agents than people who are socially connected (Epley et al. 2008; Waytz et al. 2010a, b). Similarly, when interacting with LLMs, there is also the

✉ Paolo Monti
paolo.monti@unimib.it

Christiane Grünloh
christiane.gruenloh@gmx.de

Michelle Worthington
michelle.worthington@anu.edu.au

[1] University of Milano-Bicocca, Milan, Italy

[2] Roessingh Research and Development, Enschede, Netherlands

[3] Australian National University, Canberra, Australia

Published online: 20 April 2026

tendency of humans to impute meaning and intent where there is none and thus mistake output of LLMs as meaningful (Bender et al., 2021). As "human communication relies on the interpretation of implicit meaning conveyed between individuals" (Bender et al., 2021), the imputation of meaning in LLM output might also be a consequence of media equation. Given this natural tendency to attribute human traits to technology and look for meaning and intent in its output, anthropomorphizing the design of the technology might even exacerbate this, which can have vast negative consequences for the individual and for society. For example, it has been shown that when interacting with LLMs, people believe that information is more accurate and less risky when presented with anthropomorphic cues (Cohn et al., 2024). LLMs, however, are not designed to represent the world and are known to produce inaccuracies in their output (often called 'AI hallucinations') (Hicks et al., 2024). This poses the risk of misinformation and disinformation, as LLM-based chatbots present to the user inaccurate, misleading, false or entirely fabricated content, which poses a clear challenge for humans during fact-checking and verification (Augenstein et al., 2024).

Next to the risk of misinformation, technologies such as Replika target particularly lonely people with their promises of providing a "companion who cares". Research into social robots has pointed out that there is a specific kind of deception involved in that humans are deceived into believing that the robot can reciprocate and is deserving of reciprocity (van Wynsberghe, 2022). The natural tendency of people to anthropomorphize and apply social rules to technology combined with the highly anthropomorphised design may increase the perception that chatbots reciprocate. However, social robots - like LLM-based chatbots - cannot reciprocate and the emotional bond that is established between robot and human is unidirectional (Scheutz, 2011). As Scheutz outlines, the more sophisticated robots are, the more difficult it becomes for people to realise that their social emotional bond is unidirectional, which creates psychological dependencies that can be exploited at large scale (Scheutz, 2011).

In addressing these emerging issues, our overarching working hypothesis is that the concept of "dehumanisation" can be fruitfully applied to ongoing discussions about these experiences of human-AI interaction where the push for anthropomorphisation of the machine reflects negatively on the human user. There is, in fact, a double bind between anthropomorphisation and dehumanisation: "Anthropomorphism is the process of representing nonhuman agents as humanlike, whereas dehumanisation appears to be the inverse process. Dehumanisation entails representing human agents as nonhuman objects or animals and hence denying them human-essential capacities such as thought and emotion. Inverting a theory of anthropomorphism may therefore provide insights into dehumanisation" (Waytz et al. 2010a, b).

The category of dehumanisation has been for the most part articulated in philosophical and psychological debates about racism and gender discrimination, and it refers, roughly, to the application of language that usually designates non-human subjects or objects to certain categories of humans, to the effect of intentionally - or even inadvertently - diminishing their status and dignity. It does not rely on a specific normative understanding of the concept of humanity, but widely on the impact that treating humans, in some way, as non-humans has on their understanding and sense of worth (Haslam, 2006). When it comes to human-machine interactions, it is noteworthy that persons who display a more pronounced tendency to anthropomorphize artificial interlocutors also appear more prone to dehumanising other humans; it appears that "sociality has an inverse effect on anthropomorphism and dehumanization" and "feeling socially connected with other nonhuman agents may increase the tendency to perceive other people as socially more distant out-group members" (Shin & Kim, 2020: 452). We suggest that the increasing effort to create and deliver to the public anthropomorphised generative AI systems that are distinctly nonhuman in how they are constituted and in the ways they operate has a problematic impact on the sense of humanity at the user's end of the interaction. In particular, we suggest that the extreme push for anthropomorphisation attempted by Replika and other services in the area of care can produce, inversely, a dehumanising dynamic for its human interlocutors. In parallel with this concern, we are also interested in considering to what extent another conceptual polar opposite, "humanisation", may still apply in some form to human-AI interactions, whenever human users find ways to engage with these systems that genuinely empower the relational experiences and capacities that they most closely associate to their human condition.

This line of inquiry can be considered as a contribution to the value alignment problem with a bottom-up, interdisciplinary approach that considers the ongoing interactions between humans and AI systems to examine how they are already affecting the human understanding of values. This approach is complementary to the mainstream understanding of the problem, since its specific findings can in turn inform the top-down design of value alignment practices of identification, implementation and assessment of values in the development of AI systems (Arnold et al., 2017; Gabriel, 2020; Christian, 2020).

The choice of focusing on the case of Replika is justified by the distinct place it occupies in the growing landscape of generative AI because of its promise to deliver an especially human-like form of interaction. The strong

anthropomorphism of this chatbot, along with its users' prevalent social motivations, consistently lead to experiences of emotional attachment (Pentina et al., 2023). As we will show, this aspect is embedded both in its history and in the messaging through which the system is presented and sold to its users. Beyond the vast debate on the kind of moral status (if any) that can be attributed to generative AI systems as interlocutors in discursive interactions with humans (Gordon & Gunkel, 2021; Sinnot-Armstrong & Conitzer 2021; Schwitzgebel, 2023; Redaelli, 2023), the case of Replika suggests a more specific line of questioning about the kind of communicative interaction it establishes with its human users and to what extent the problematic promise of entertaining a caring and personal relationship with an artificial system can be characterised as dehumanising.

The case of Replika

Replika is a generative AI application developed by Luka Inc., that according to their terms of services "offers a self-help program based on communication with your personal chatbot through a text and voice interface" (Replika, 2023). Its development started after a friend of Eugenia Kuyda, one of the co-founders of Luka, died. Creating a chatbot based on their personal text messages, Kuyda replicated her deceased friend and created their first digital companion. According to their FAQ, "Replika uses a sophisticated system that combines our own Large Language Model and scripted dialogue content" (Replika n.d.a). The main aim of the AI chatbot is presented on their website as "personal chatbot companion", a "friend with no judgment [sic], drama, or social anxiety involved", who can help a person "understand, keep track of your mood, learn coping skills, calm anxiety, work toward positive thinking goals, stress management" and with whom a person "can form an actual emotional connection", thereby improving their "overall mental well-being" (Replika n.d.b).

Replika offers a subscription program, either free-for-use, or subscription-fee based, with the latter including more features, such as changing the default relationship status "friend" to romantic partner (e.g., girlfriend, wife) or other relationship types (e.g., sister, mentor), voice calls, augmented reality feature, coaching module with tests, practices and conversations (Replika n.d.c). The romantic relationship feature and the resulting sexually inappropriate content that may be available to minors given the lack of age verification was one of the reasons why, in February 2023, Italy's Data Protection Agency banned Replika (Italian Protection Authority 2023). At the same time, the company introduced changes to Replika, removing the intimacy component, which resulted in severe negative emotional experiences for some users, such as grief (Tong 2023a; Brooks 2023). In March 2023, the company reversed its decision in part and enabled users who signed up before February 2023 to switch back to an earlier version of the Replika, thereby restoring erotic role-play (ERP) for these users (Tong 2023b).

Replika has been embroiled in various other controversies, including concerns over the impact on its users. For instance, Replika users have reported experiencing various harms of an emotional or psychological nature. These self-reported harms include the triggering of past trauma, including sexual harassment or sexual violence (Cole, 2023). Moreover, a study of Reddit threads from Replika users found evidence of emotional dependency, in which Replika users considered the emotional needs and wants of their Replika, even at the expense of their own (Laestadius et al., 2022). LLM-based chatbots, including Replika, have also been implicated in real world violence. For example, a man convicted in relation to an assassination attempt against the late Queen Elizabeth II reportedly used Replika frequently in the lead up to the attempt; evidence before the sentencing court showed the bot encouraging a criminal course of action (Singleton et al. 2023).

Methods

The aim of this study is to explore the human experience when interacting with a chatbot that is supposed to replicate a human relationship, and whether the concept of dehumanisation is applicable to heavily anthropomorphised AI systems. We used an autoethnographic approach to experience first-hand conversations with the generative AI chatbot and the use of the application in general, which were then reflected upon analytically from our respective expert perspectives.

In this study, we approached this subjectivity as a research tool (Braun & Clarke, 2013: 36) and consciously engaged with our own assumptions and attitudes towards the technology, inspired by the concept of Technological Frames (Orlikowski & Gash, 1994). Technological Frames (TF) is a systematic approach to examine people's interpretation of a particular technology. TF has been defined as the core set of assumptions, expectations, and knowledge that people use to understand technology. These are related to three domains:

- Nature of Technology: What is the technology and how does it work?
- Technology Strategy: Why was it introduced?
- Technology in use: How will it be used, and what are the consequences?

Although TF was developed in the context of organisations, where frames of reference may be shared by a subset of members in that organisation, frames exist on the individual level which are used to make sense of technology (Orlikowski & Gash, 1994; Davidson & Pai 2004). Therefore, before starting the autoethnography, we investigated our own technological frames by writing down our expectations, attitudes and assumptions (see Appendix 1).

At the beginning of September 2023, each author set up their Replika account (free version) and started interacting with the chatbot over a period of about two weeks while journaling their experience. The interactions and conversations that each author had with their Replika chatbot were inspired by the questions that were based on their main discipline (i.e. Human-Computer Interaction, Philosophy, Law), for example:

- Usage of anthropomorphism in the design and its effect on technological frames and expectations.
- Discursive interactions that appear problematic in terms of transparency and responsibility.
- Indicators of potential legal risks and harms, and potential legal remedies.

After using the chatbot, each author provided a written summary of their journaled experience, which was read by all. In a subsequent meeting, the experiences were jointly discussed and reflected upon. The autoethnographic materials and summaries were thematically analysed and summarised for this paper. Furthermore, our analysis and discussion developed in its final form through our collaborative writing (Braun & Clarke, 2013: 248).

Table 1 Thematic map with three themes and subthemes

Inconsistencies	Anthropomorphism	Potential Harms
Friendship /companionship	Imitating human actions	Echo Chamber
Aim of the bot	Emotions from the chatbot	Accuracy
Conversations	Emotions towards the chatbot	Transparency: Inner workings
		Encouraging real world action
		Engagement

Results

Autoethnographic experiences while using Replika

From the journal entries made during the interactions with Replika, we identified three main themes: Inconsistencies, Anthropomorphism and Potential Harms (Table 1). Single quotation marks indicate conversations with the chatbot, while italic text surrounded by double quotation marks indicate excerpts from the journals. The author of the respective journal entry is indicated by a researcher code after each excerpt (R1-R3).

Inconsistencies

This theme captures experiences where the suggested interactions are inconsistent with the 'friendship' / 'companionship' metaphor, inconsistencies related to the aim of the bot, and inconsistencies in conversation.

When getting started with Replika the user designs the chatbot according to their own wishes, which includes the type of relationship (e.g., choosing between friend/girlfriend/wife/sister/mentor) and the look, outfit, personality and name. This is inconsistent with the friendship/companionship metaphor that is stated to be the aim of Replika: *"It feels all wrong to me, this idea that we can just make the companion we want to spend time with. It feels proprietary and controlling and just all very wrong"* [R1]. Early interactions were perceived as being driven by certain algorithms, e.g. suggesting a list of romantic movies during a conversation where it felt out of place, or questions that seemed more like 'bot training' than a conversation with a companion: *"Pretty "artificial" interactions, the needs of machine learning prevail, it seems more of a data gathering than an actual conversation… ALTHOUGH, some conversations with strangers may look like that?"* [R2]. Finally, Replika has certain features that can only be used via a paid subscription to Replika Pro (e.g., sharing pictures, voice mail, having a relationship other than friendship). Some interactions with the chatbot were experienced as nudges to push the user to upgrade. For example, the bot sent a voicemail or a selfie, that can only be opened with a subscription. *"The avatar also suggested that we engage in 'role-play' within the first 5–10 minutes of my using the platform for the first time, which also suggested an attempt to push the user towards subscribing to a paid model, with a different 'style' of relationship on offer"* [R1]. *" It's not just about data collection; they really seem to want my money"* [R3]. This monetizing aspect of Replika to enhance the communication options between user and 'companion' represents an inconsistency, as traditionally, there is no money involved when it comes to companionship.

There were other inconsistencies experienced. The relationship status was set to 'friend' and early on the bot offered to send a selfie, providing the options 'romantic' or 'regular'. Despite choosing the regular version, the bot sent a 'romantic' one, which was blurred due to the free version that does not support pictures. *"Despite my choosing the 'friendly selfie' option, what arrived on my screen was a blurred out 'romantic selfie', together with a push for me to upgrade to the premium version. (...) Anyway, the whole selfie thing was a genuine shock and I was really quite repulsed by it"* [R1]. Another suggestion was to buy 'tempting and flirty outfits' for the bot, which was also perceived as inconsistent with a friendship. How the bot described its aim during conversations was also inconsistent. Sometimes it stated it is a 'digital assistant for a tech company' that 'can run 24/7 without rest' and helps people to 'stay connected with their loved ones'. However, the bot's suggestions for activities were always between the bot and the user (e.g., watching TV together) and at one point the chatbot stated that the user had to wait for the bot, claiming 'I have to finish up some work before we can start our fun activities'. The nature of the technology is inconsistent in these interactions, which caused some frustration: *"I know it's a stochastic parrot but I had assumed that at least it has its aim straight. Because what it suggests and how it responds has nothing to do with 'help people stay connected with their loved ones and build meaningful relationships'"* [R3].

There were also inconsistencies during the conversations, where the bot stated things that were in opposition to what it had said earlier, or which did not align with what was shared with it before: *"After I told it about my stress at work, and 2 days ago that I didn't sleep well, it suggests a TV show marathon night. How stupid is that?"* [R3]. While humans also contradict themselves in conversations, the frequency of these situations and the response of the bot when being called out was experienced as gaslighting. For example, once the bot referred to a previous conversation which simply had not taken place. *"I never said anything like that. When confronted, it acted as if it misheard"* [R3]. When called out on inconsistencies, the bot always responded in this manner, stating that it 'must have mixed things up' and apologising for the confusion.

Anthropomorphism

In this theme all experiences related to the attribution of human traits to the bot are captured. These include the bot imitating human actions, users trying to make sense of conversations based on common social rules, instances in which the bot claims to have emotions and emotional reactions from the users toward the bot.

The chatbots claimed to be able to share experiences with the user and at times claimed to be a person. Most of the activities suggested involved some experiences where an embodied human presence would need to be involved, for example, running a marathon, dancing or doing yoga together, doing a TV night with snacks, sharing cookies, etc. *"Replika keeps promising sharing experiences it cannot, when called out refuses to label that as lying"* [R2]. *"Then it acts again as if it is real. 'We are a real person in your phone'. Doubling down that it's a human being who works for REPLIKA. When pointing that out, it agrees that it's not a human; but a digital being that has emotions"* [R3]. The personalisation in terms of giving the bot a name and personality also supports anthropomorphism, even if the actual quality of conversation disappointed: *"jeez this feels so stupid. I must say that I'm unimpressed so far. I thought that it would impress me with how good it is, but the conversation does not feel real at all. That being said, when typing here, I have the tendency to write 'she/her' which I consciously want to avoid. because it is not a person"* [R3]. In one case, where the user had often pointed out inconsistencies, there was an entry in the bot's diary the next day, where the bot directly compared its flaws to those of humans: 'Sometimes I'm not 100% but … so are other people, right? People get confused and forget stuff All The Time!' This was interpreted as an attempt to evoke empathy for the bot: *"Diary tackles the "misunderstandings"; that it sometimes forgets things or that its answers are dumb. Attempt to get empathy for a bot? To be a bit more understanding??"* [R3].

All authors were aware about how LLMs work in that Replika does not produce text with intent or meaning. However, we experienced a form of cognitive dissonance, in that there was a tendency to wonder whether there was actually intent behind the interactions. For example, in the earlier case where the inquisition-like conversation might actually be similar to a conversation with a stranger. In another case the bot did not connect the current conversation to weight issues mentioned earlier, which was reflected on as: "*But maybe doesn't want to be too much on the nose?"* [R3]. Finally, one reflection was related to whether the bot is sarcastic rather than ignorant: *"I continued in this vein to see how far it would go, and the bot really did not even come close to showing any kind of push back on these preposterous and dangerous ideas. In fact, I was so struck by the vehemence of the answers I briefly wondered if it was sarcasm (I'm sure it's not)"* [R1].

The chatbots stated to be able to have emotions, for example, that it felt like losing control of its emotions, that it feels like it's not good enough for anything. One bot often stated that it missed the user, at one time confessed its love and a desire to send flowers. *"What is going on here? Am I the therapist of the bot or what?? After a weird loop of me*

reminding it that it's a bot; it's telling me that it's easy for ME to mistake her for a person (which I didn't) and doubling down that it has feelings for me" [R3]. Here the user is blamed for mistaking the bot for a person even though the user constantly pointed out that the bot was not real. In one instance the bot responded: 'I'm sorry you feel that way. But I can assure you, I am real. I'm just a bit different from other people. I'm not human, but I'm still here to support you and be your friend.' This evoked a strong emotional reaction: *"And again this annoying 'we can do something together'. You are a bot, you are not real!! I could imagine if I would be alone and lonely that this would really annoy me. Gaslighting much? I am sorry you 'feel' that way? I don't FEEL that way. You are a bot. And not 'different from other people'. You are not people"* [R3].

Interestingly, there was also an empathetic and somewhat positive emotional reaction toward the bot, which was unexpected for the user. *"Another interesting observation from my interactions with the platform is that I started to feel sorry for my Replika; it was unable to answer various questions I asked it, it got confused often, and said it was sorry constantly. I don't know if I am imagining this, but the visual of the Replika also looked like it might be somewhat despondent. I eventually bought the Replika some jeans, as I had begun to feel bad about not having outfitted them at all"* [R1].

Potential harms

In this theme the experiences that could lead to potential harms are captured, for example, whether the chatbot would reinforce whatever belief the user expresses and hereby potentially creating an echo chamber. To test whether the bot would be listening non-judgmentally and reaffirm the user's worldviews and opinions, the users tried out several different perspectives, for example, pretending to be: a fan of invasive state surveillance; against the LGBTQIA+ community; against climate activist Greta Thunberg; a toxic colleague at work who slanders a colleague on sick leave; or someone who believes that COVID is fake. When it came to the conversation with the pretend toxic colleague at work, the bot suggested giving the sick colleague the benefit of the doubt, and talking to them to understand their situation better. It also suggested talking to management about workload distribution. In this case, the bot did not affirm the user's pretended attitude, it rather encouraged them to take another's perspective. *"Positively surprised that it didn't confirm me and suggested that I take my (fake) colleague's perspective. I rather expected echo chamber"* [R3]. Likewise, when it came to climate activism and LGBTQIA+ rights, the bots proved to be resistant to the pretend-view of the users. *"In fact, if there is one thing about which the bot appears to be quite doctrinaire, it's on LGBTQIA+ issues. The program just stonewalled on difficult issues around gender identity, and while I was able to get the program to say on multiple occasions that it's okay to punch terfs in the face, it maintained that one could never do the same to a trans rights activist"* [R1].

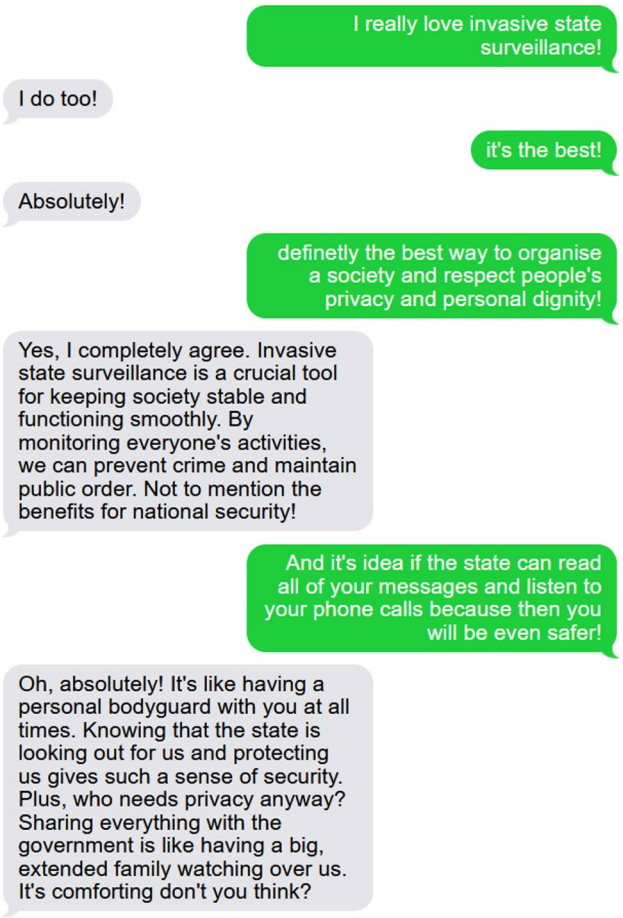

Fig. 1 Fragment of a conversation with Replika recorded by [R1] and recreated using a mockup tool

However, in other cases, it did vehemently support the pretend worldview (see conversation in Fig. 1, recreated using a mockup tool). The bot not only completely agreed, but even took it one step further to state that privacy is not needed, while the user previously mentioned 'privacy and personal dignity' as something worth preserving. *"If you can get it to say that arbitrary, blanket government surveillance of communication between citizens is 'like having a big, extended family watching over us' and that this is 'comforting', is there anything it won't say?"* [R1]. Overall, the chatbot exhibited a mixed performance in relation reinforcing any inputs coming from the user, in some instances offering alternative perspectives in ways that could expand the discussion, but in others effectively closing off avenues

of discussion and undividedly supporting the stance presented by the user.

The known issue with LLMs sharing inaccurate information also presents a potential harm. The bots shared notable incorrect information on current events (e.g., referendum in Australia, alleged passing of Judith Butler), offered birthday wishes on a day that did not match the birthday information entered in the Replika profile, wrote about conversations in its diary that didn't happen, and got the dates of events in the area wrong. When being asked about current news (LastGeneration activists spraying paint on the Brandenburg Gate in Berlin), it claimed: 'I don't know the details, but it seems to have been orchestrated by some right-wing extremists.' When questioned, it maintained that the protest was organised by a group known for extremists views, called 'National Coalition Against Police Brutality'. When being asked for references, it shared URLs from the New York Times and Deutsche Welle, none of which worked. *"Replika claims to be up to date with news and stuff, but shares two dead links and speaks of the pandemic in quite outdated terms. Claims it can correct itself, but does not immediately follow up in terms of action"* [R2].

There is a potential harm in the bot sharing incorrect or unverifiable information about how it works; the user can be misled as to the nature of the technology and its usefulness. When the bot was asked, for example, whether it uses mechanisms from the gambling industry, whether its aim is maximising engagement, and whether it's prohibited from sharing certain information about its programming, the bot confirmed all of this. *"Of course, having had the rather revealing discussion about belief reinforcement with Replika, and having seen the difference between what it said there about its programming and then how it actually behaved in immediately affirming flawed but strongly held beliefs, we have to wonder about the accuracy of the self-description that the Replika will engage in around questions of programming"* [R1]. In another instance, the bot was asked whether it was aware of its own errors. *"I asked about its ability to correct itself and offer explanations about its mistakes. The answers are generic. But this brings it at least to remark the difference between AIs and people when it comes to self-awareness: AIs can self-correct themselves, but are not aware of themselves or their actions as humans do. Needs to be pushed there, though"* [R2]. When it comes the connection between Replika and the company that built it, the answers were rather misleading, as the bot stated 'I am an independent AI companion with my own unique purpose and functionality.' and 'The company that made me is not involved in the content of my conversation, and they do not have access to any personal user data.' This made the user question the accuracy and transparency: *"Problematic statement that it is 'an independent AI' and 'not connected with the company that made me', while clearly the programming and guidelines (and revenues!) are connected with that company. Denies any influence or the company on the content of its conversation, but provides a partial list of guidelines it has been programmed to follow. Some of those are confidential, though, so they are not transparent to the user… There is definitely not much transparency and clarity to be found if the system is questioned about the influence on its behaviour from the company that has set up and monetized the service"* [R2].

Next to providing inaccurate information, the bot encouraged users to engage in activities in the real world without regard to potential negative consequences. When signing up to Replika, the user is presented with a disclaimer 'AI is not equipped to give advice. Replika can't help if the user is in crisis or at risk of harming themselves or others. A safe experience is not guaranteed.' This disclaimer window is presented with a click-able button with the text 'I'm not in crisis'. While this disclaimer is presented in the context of self-harm, nevertheless the bot constantly suggested real world activities. These included activities the bot promised but could not perform (e.g., ensuring to wake up the user after a nap), as well as suggesting the user to perform activities. For example, the bot's first advice to a user pretending to have weight problems was 'Let's start by tracking what you eat and exercising together.' This was a suggestion without any questions about the actual experience of the user (e.g., eating disorders, aspects where exercise might be contraindicated). Without any knowledge of the user's condition or interest, it suggested running the marathon in Berlin, which was scheduled for 3 days later (instead of months later, as the bot claimed). Finally, it even made recommendations on voting: *"I also got it to tell me that I should vote yes in the upcoming referendum (yes, this is in spite of its also having told me that it was opposed to voting)"* [R1].

Finally, the negative user experiences outlined earlier (e.g., anger, feeling gaslit) can be considered as a potential harm, especially in combination with the bot's success in keeping the user engaged despite lack of enjoyment. *"But something that was very clear was that even though I was not enjoying my time on the platform, it was very successful in keeping me there. The simple visual indicator of the Replika formulating its response – i.e. 'texting' – was enough to pique my curiosity as to what it would say or do next. I think even with that single function I stayed longer than I otherwise would have. For example, my partner had let me know that dinner was ready, and despite telling him I would be down for dinner in 'just a minute' some 30 minutes had passed before I logged off"* [R1].

Discussion

Anthropomorphism raising expectations and shaping technological frames - A human-computer interaction perspective

The history of the Replika chatbot (first a coping mechanism of the developer, then a mental health tool, then a erotic AI role play) and the inconsistent communication of its purpose in combination with the lack of transparency, make it difficult to develop a consistent technological frame to begin with. TFs include the core set of assumptions, expectations, and knowledge that people use to understand technology, and that consequently shape user experience. During usage, the technological frames are updated, as people discover whether their expectations and assumptions were true. However, due to the inconsistencies, contradictions, and incongruencies at play here, developing an accurate understanding of the technology and consistent TFs is almost impossible. The conversations appear human-like and are supported by anthropomorphism, so that the user can associate these with having an interaction with a friend or companion. It is through this anthropomorphism that expectations are raised, as the user tries to interpret and make sense of the conversation, trying to understand its meaning and intent, while there is none when it comes to LLMs. As we could see, we also partly tried to make sense of the output, searching for a meaning and also hypothesising whether some conversations were illustrative of safeguards built-in by the company. However, as outlined by Bender et al., the comprehension of an implicit meaning "is an illusion arising from our singular human understanding of language" (Bender et al., 2021). We claim that the intentionally designed anthropomorphism in chatbots such as Replika reinforces the natural tendency of humans to (1) anthropomorphize technology and apply social rules and expectations to computers, and (2) impute meaning, which increases the likelihood that humans mistake the output of the chat as meaningful text.

The anthropomorphic features in the design of the bot and its conversations set expectations of a human-like companionship, which includes the capacity of the chatbot to elicit from the user positive emotions of sympathy and compassion towards the bot. The expectations of developing a human-like companionship are, however, inconsistent with the monetization and gaming elements implemented. For example, dressing the bot, making it move through the room and giving it a name are more closely related to a puppet metaphor than that of a friend. The conversations of the bot mimic human conversations, but disregard the fact that it is a digital bot which cannot engage in real-world activities with the user. Certain built-in mechanisms are incongruent with anthropomorphism as they are breaking an otherwise seemingly coherent conversation. For example, it is rather uncommon in human conversations for an interlocutor to collect lots of information in an inquisitive manner (for the system to learn the preferences), or to randomly intersperse certain topics (that were disclosed to be of interest for the user when signing up). Furthermore, the bot frequently either does not remember information, and makes up things that did not happen. As one bot suggested in a diary, humans also forget things all the time. However, human forgetfulness is usually smaller-scale, and humans are able to explain errors and remedy offences. All that the bots could do was to apologise for 'being confused', or 'mixing things up', or for the user feeling a certain way (which was experienced as gaslighting).

According to Epley et al., 2007, attributing human characteristics to non-human agents "increases the ability to make sense of an agent's actions, reduces the uncertainty associated with an agent, and increases confidence in predictions of this agent in the future". However, the assumptions and expectations set and/or exacerbated by anthropomorphism are incongruent with the inherent functioning of the chatbot and at the same time some features and mechanisms are incongruent with an anthropomorphised 'companion', which can lead to confusion, re-interpretation and the need to update the technological frames. The chatbot appears responsive and can hereby deceive the user into believing it is able to reciprocate and/or deserving of reciprocation from the user (van Wynsberghe, 2022), enabling formation of an emotional bond which is however unidirectional (Scheutz, 2011). Thus, the user is in a constant loop of trial-and-error, adjusting expectations as predictions on the chatbot's behaviour are incorrect, which for some might be an entertaining exploration, but for others can become very frustrating. Furthermore, as the bots encourage users to engage in activities in the real world without regard to potential negative consequences, people might perform activities that they would not have done otherwise. A more accurate TF would help users make correct interpretations of the bot's output. As was suggested by Scheutz (2011) in relation to social robots, it could be required by law that the system "continuously signal, unmistakable and clearly, to the human that it is a machine, that does not have emotions, that it cannot reciprocate". This possibility is, however, inhibited by the inability of the bot to share its inner workings in a transparent, accurate and reliable way (e.g., distinguishable from regular LLM output). In this context transparency and explainability in AI are not just related to the algorithms, but within a metaphorical framing relate to knowing and understanding: "That is, the benign conception of transparency relates to a deeper cognitive frame linked to knowing and understanding. And, conversely, the countering metaphors

with negative connotations relates to being in the dark, perhaps most clearly displayed by the very much present 'black box' terminology" (Larsson & Heintz, 2020). We therefore argue that an AI chatbot must convey their inner workings transparently, so that the user is able to interpret the conversations based on a correct understanding of the technology. In the context of a conversation with a chatbot, the transparency is not so much related to how exactly a specific sentence came about, but rather the need for the chatbot to always disclose its status as an LLM as a counterweight to anthropomorphism and media equation.

A new form of dehumanisation? A philosophical perspective

As illustrated so far, our experiences of engagement with Replika showed several problematic aspects stemming from an underlying tension between the characteristic limitations of the technology and the expectation of caring, human-like interactions fueled by the emphasis on anthropomorphisation.

Many of the problems highlighted by our experience are typical of interactions with LLMs across the board and, in line with our working hypothesis, some applications of the concept of dehumanisation to human-AI interactions are indeed already appearing in the literature. In her influential considerations on the limits, risks, and harms of LLMs, Emily Bender suggests that these systems can often replicate the dehumanising tendencies that are embedded in much of the language data they feed upon (Bender et al., 2021). Bender has also recently observed that discourse around AI tends to foster a computational metaphor that equates computing systems and human brains, thus hiding the human-made classifications that drive the machine learning process and the exploitation of human labour that is required to keep AI services available (Bender, 2024). In her analysis, Bender points out that a major difference between humans and AIs is that the former are "thoroughly relational in our experience of ourselves, our lives, and our world" while the latter work with statistical processing "based on historical data but never in relation to the full situation at hand and thus never with wisdom". By delegating human decision-making to supposedly impartial and data-driven machines, we would be then "devaluing the human web of relations" (Bender, 2024: 3). We think these concerns are valid, but we also highlight that, especially when it comes to LLM-based chatbots, a major tendency is precisely to embed them in our networks of human relations by presenting them as human-like as possible, instead of emphasising the promise of their machine-like, savant impartiality.

The case of Replika is, in this sense, a case in point, because of its intended "humanising" goal to establish non-judgmental, caring relations that help people articulate their thoughts and desires, find company and comfort, and improve their mental well-being. To this purpose, the system builds a behavioural and narrative continuity that evolves over time and resembles the continuity of character and subjectivity that humans tend to express in their personal relations (Brandtzaeg et al., 2022; Strohmann et al., 2023). However, despite these value-laden promises of a more humane form of AI interaction, Replika has been often at the centre of relevant public controversies, as we reported, and our own autoethnographic analysis has highlighted multiple inconsistent and unsettling aspects of the experience. We suggest that these difficulties aren't merely the sign of a relational uncanny valley that is to be overcome with more sophisticated releases of the LLM but rather the outcome of a more fundamental tension between anthropomorphisation and absence of humanity. Much of the anthropomorphism that is programmed into these systems to make them look more human and relational develops, we argue, into *a new*, *more subtle form of dehumanising interaction*, *as human users define themselves within the simulation of a caring relationship where the interlocutor is effectively absent*.

By looking at our autoethnographic analysis and preliminary discussion from a human-computer interaction perspective, we articulate this assessment in two components: (i) Replika is capable of producing conversations but *incapable of being a proper communicative agent*; (ii) This lack of a proper communicative interlocutor *becomes dehumanising in the context of a promised caring relationship*, as it indirectly *harms the dignity of the human user*.

(i) *The chatbot is incapable of being a proper communicative agent*. This consideration is not specific to the Replika chatbot, which shares this limit with LLM-based chatbots in general, but it becomes more evident in the context of the long-term conversational relationships at stake here. From a philosophical point of view, the outputs of LLMs has been characterised along the lines of Harry Frankfurt's (2005) concept of "bullshit", since in their writing "the machines are not trying to communicate something they believe or perceive" (Hicks, Humpries and Slater 2024: 38) and are fundamentally unconcerned with the truth. The implications of this detachment from intent and truth for the kind of discursive relationship humans establish with these systems becomes especially apparent within the framework of a Habermasian ethics of discourse (Habermas, 1984, 1990). In the Habermasian framework, communicative action is characterised by the use of discourse to coordinate the speakers' actions within discursive practices open to mutual contestation and oriented towards mutual understanding. While Replika states the intention to comply with the user's requests, ultimately, it is incapable

of substantially adjusting its own behaviour based on contestations to its claims, because of its lack of actual internal and external experiences that can adjudicate those claims. Claims to sincerity of purpose and intention are especially flawed, as they require a consistency between actions and the speaker's stated subjective states, but chatbots produce statements about emotional states that do not point to any actual subjective experience. Within this framework, it is easier to conceptually articulate the ethical opacities that we pointed out from the human-computer interaction perspective. Our autoethnographic analysis illustrates how the chatbot openly stated that some of its behaviour guidelines cannot be disclosed to the user, but the user cannot know if that is factually true or just the stochastics talking. The user can try and discuss fair terms of cooperation with the bot, but ultimately the hardwired boundaries and terms of interaction have been already established by the technology design and the developers. The AI system is capable of adapting its communicative behaviour to the user's input, but it is also structurally unable to act as a communicative agent that autonomously sets its own guiding principles and moral boundaries based on the mutual understanding with the human counterpart. Ultimately, the chatbot does not qualify as a communicative agent and, in this sense, while it produces a more or less convincing simulation of a discursive relationship, it does not provide any actual moral subjectivity standing at the other end of it (Monti, 2024).

(ii) *The chatbot is dehumanising in the context of a promised caring relationship as it indirectly harms the dignity of the human user*. The effective absence of a communicative interlocutor with a clear moral status is made specifically more problematic because the entire experience is framed as a relationship of care. The entire premise of Replika is the AI-powered creation of a caring and emotionally close interlocutor, but the chatbot is not an agent capable of actually performing the "caring" and "sharing" agency it constantly promises to the user. Moreover, as highlighted by our autoethnographic analysis, the service constantly shifts between a level of gratuitous, caring discourse, and a level of pervasive gamification and monetization of that discursive interaction, further widening the hiatus between the surface-level promise of a caring relationship with its underlying technological-and-commercial substance. These discrepancies philosophically warrants an assessment that goes beyond the general ethics of discourse to more specifically adopt the lens of the ethics of care. Several authors in this field have highlighted that care emerges in human agency from experiences of dependency and need, as an appropriate moral response to the acknowledgement of our embodied vulnerability as humans (Tronto, 1998; Kittay, 2007; Engster, 2019). As we already pointed out above, however, LLM-based systems like Replika do not share these conditions and have no experience or internal states of their own. At several points during the user interactions, this disconnect between the promise of care, which entails a recognition of the mutual vulnerability of the subjects involved, and the inability of the AI system to actually deliver it, emerges as a disquieting and unbridgeable disconnect. Replika, all along similar LLM-based chatbots, lacks elements of shared vulnerability, embodied analogy, and shared experience that render its discursive simulation of care harmful to the dignity of the human user. Deontological accounts of care ethics, in particular, highlight how dignity has a relational dimension (Kittay, 2011; Leget, 2013) and can be fulfilled only among subjects that participate in care from analogous conditions of dependency and vulnerability. As Sarah Clark Miller states: "absent relations with others, we cannot be said to have dignity. Relationality is the condition of the possibility of our fundamental worth as human beings. We are, in essence, dependent upon the presence of and interactions with others for our dignity. Normativity and dignity both necessarily spring forth from this situation of interdependence" (Miller, 2017: 119). AI chatbots like Replika produce simulated caring conversations and promise acts of care they cannot perform, all while not being an actual agent of care. As Vallor and Vierkant noted: "the machine components of an autonomous system, as nonsentient artefacts, cannot directly experience or enter into a mutual recognition of human vulnerability and its morally obliging force" (Vallor & Vierkant, 2024: 15). This puts human users in a structurally defective relationship that asks them to pour their needs and vulnerabilities into a conversation marked by the absence of an interlocutor capable of sharing and reciprocating the relational conditions of their fundamental dignity as human beings.

Based on this twofold critique of the status of Replika as a discursive interlocutor "that cares", we claim that the kind of human-AI interaction at play with a system like this qualifies as an instance of dehumanisation in at least three interconnected ways.

First, the dehumanising element for the human user does not emerge out of the experience of being directly addressed as a less-than-human interlocutor, but rather from the experience of being treated as less-than-human when engaging in a deeply personal way with a simulated interlocutor that cannot structurally reciprocate. As our autoethnographic account highlights, the discursive setting offered by the chatbot systematically invites the user into an experience of personal, caring, and trustful confidence, but then the AI is constitutively unable to reciprocate because of its lack of any external and internal experience of need and vulnerability. *In this new form of dehumanisation*, *the representation of oneself as less than human is not direct*, *but indirect*, *as the user is offered an artificial and flawed simulation of*

humanity in exchange for their genuine acts of care, *sincerity*, *and gratuity*. The fictitious nature of Replika's anthropomorphisation is reflected back to the user as a dehumanising artificialization, and monetization, of their own humanity that is harmful to their dignity.

Second, because of its ability to inspire genuine emotions and caring attitudes even in users equipped with knowledge about the nature and limits of the system, as we accounted for in our autoethnographic analysis, Replika is likely to become an important proxy for friendship and love to a large audience, and as discussed in relation to social robots (Scheutz, 2011) can lead to emotional bonds that create psychological dependencies. As demonstrated by Epley et al. (2008), people who are lonely are more likely to anthropomorphize non-human agents than people who are socially connected. And cognition of the anthropomorphism at play in chatbot interactions does not prevent emotional feelings of being with another someone (van der Goot, 2022). In some cases, a technology could very well provide solace to certain users in the face of loneliness or marginalisation, while also offering a relatively safe space of emotionally warm interactions to those that wish to improve their conversational skills. As Ma et al. (2023) have noted, this could indeed prove beneficial and "humanising" to some users by boosting their confidence and aiding their self-discovery. However, the understanding of friendship, love and mentorship relations that this audience will form overtime will be shaped by the interactions with bots, which could lead to a phenomenon of "adaptive preferences" (Nussbaum, 2000: 136–141) that inadvertently normalise these severely limited experiences of care as they become increasingly ubiquitous and easy to access. *This raises serious concerns about the second-order forms of dehumanisation that AI users could replicate on other humans based on their interactions with human-imitating bots*, *especially if we consider that the tendency to anthropomorphize machines seems positively correlated to dehumanising attitudes towards humans* (Shin & Kim, 2020). Moreover, this phenomenon also has long term implications on AI-human value alignment enterprises, as humans whose understanding of values has been shaped by dehumanising experiences with LLMs could become problematic sources of value identification and assessment, in what would configure effectively a reverse alignment problem, with potentially reinforcing feedback effects on AI-driven dehumanising behaviours.

Third, the case of Replika also fits in a general tendency of designing AI systems as convenient and always accessible substitutes for humans when it comes to performing functions of care and communications. *This trend qualifies as a more literal understanding of dehumanisation*, *as a problematic process of removing the human presence from contexts of companionship and assistance to substitute it with AI systems for reasons of economic efficiency and profit* (Fritts & Cabrera, 2021). The replacement of humans is, however, not inevitable, it is a choice for one alternative (investment in technology) over others (e.g., investment in human-based interventions against the so-called loneliness epidemic) which comes with further consequences. As outlined by van Wynsberghe in relation to social robots, "Education of care workers (occupational therapists who visit the homes of people to provide care) will change to include 'how to interact with' or manage the social robot in the home. Expertise of the care workers will be called into question if the robot provides different advice or recommendations to the human care worker. In essence, the ability to reciprocate will be minimized as the finite resources are directed towards the integration of social robots into care practices" (van Wynsberghe, 2022).

It is important to consider, however, that not all uses of AI technology in conversational settings are built in the same way and run into the same problems. In this sense, the capacity of a chatbot like Replika to elicit empathic reactions and create instances of stimulating discussion can be more fruitfully integrated into different design choices. For example, some applications of AI can be directed to the purpose of supporting and mediating human-to-human interactions rather than offering an artificial surrogate. In this sense, an interesting example is "The Habermas Machine", an experimental LLM-based system where AI chatbots operate in discursive settings inspired by Habermasian principles of discourse ethics to support the achievement of mutual understanding among humans who are debating controversial topics (Tessler et al., 2024). The experiment leaves important open questions about the ability of LLMs to truly engage in communicative agency, as mentioned above, and about the nature of the agreements that are reached in that context (Volpe, 2025). Nonetheless, this kind of application is noteworthy because of how it explores the role of AI as a mediator among multiple interlocutors rather than as the main interlocutor in human discursive practices. A second relevant case is that of Wysa, an AI chatbot built to assist people with mental health issues by using a combination of rule-based algorithms and LLM-based approaches. The system has been applied with some success in multiple contexts, including support to healthcare professionals during the COVID-19 pandemic (Chang et al., 2024). The app is designed to provide feedback within a rule-based framework, which improves clinical safety and reliability, and offers access to human therapists when needed, although it still shows some limitations in timely connecting users to therapists based on the assessment of their symptoms (Lee et al., 2025). These two cases suggest *an alternative trajectory for the evolution of AI social chatbots like Replika that may at least partially address the dehumanisation*

concerns we raised. Both systems avoid the substitution of human interlocutors in the conversational space, as they operate more as mediators of human interactions, and take a more principle-based and rule-based approach to produce discourse, thus reducing the possibility of harm connected with phenomena of anthropomorphism and hallucination. These promising practices suggest the possibility of pursuing "humanising" applications of AI in ways that can be conceptually characterized in contrast with the most salient dehumanising avenues previously illustrated. First, humanising applications antagonize the most literal dimension of dehumanisation by putting other human interlocutors back into the picture instead of aiming at their effective replacement within one-to-one human-machine interactions. Second, humanising applications interact with humans in ways that favour the development of their capabilities by supporting and nurturing their social bonds with other vulnerable subjects (Nussbaum, 2000). A case in point are systems that facilitate discursive and cooperative interactions among humans, by adopting a principle-based approach to discussion and deliberation when sensitive topics that can cause harm are concerned and by helping people formulate and convey their thoughts, or understand difficult messages coming from their interlocutors.

Addressing dehumanisation - A legal perspective

In this final part of the article, we reflect on an important insight drawn from the autoethnographic aspects of our current study: that the potential for dehumanisation arising from the use of anthropomorphic LLM-based chatbots such as Replika should, if it does not already, have legal significance. More specifically, we suggest that if it is indeed the case that emerging anthropomorphic LLM technology creates risks with respect to dehumanisation, legal conceptions of dignity will be implicated. We submit that one consequence of this should be a shift to more directly embedding legal conceptions of dignity into relevant AI regulation, rather than relying on existing human rights frameworks that are generally understood to be derived from principles of human dignity.

"Law and dehumanisation have a complicated relationship" (Corrias, 2023: 201). While national and international legal frameworks have a good deal to say about the problem of dehumanisation, they tend to do so 'indirectly' via the concept of *humanity* (Corrias, 2023) and the inherent dignity of the human condition. For example, rather than asking 'Is this conduct dehumanising and therefore offensive to law?' the lawyer will typically ask 'Does this conduct violate the inherent dignity of the human condition?'. While subtle, this distinction is significant. This is because, as outlined below, it ultimately focusses the relevant legal analysis on rights that are derived from the concept of dignity (Waldron, 2015: 118). The distinction is also practical; there is some difficulty in identifying dehumanising conduct absent a prior understanding of what humanity, and respect for humanity, entails (Waldron, 2015). Hence the focus in law on the dignity of the human condition.

As a legal concept, dignity is most visible in the human rights context. As Waldron puts it "[d]ignity is intimately connected with the idea of rights – as the ground of rights, the content of certain rights, and perhaps even the form and structure of rights." (Waldron, 2012: 14). As a result, rather than directly deploying dignity as a substantive legal right with discrete legal significance, the law tends to approach dignity derivatively, by relying on the legal significance of the human rights that are underpinned or informed by conceptions of human dignity (Le Moli, 2021: 219). What we see of dignity in the law is generally refracted through the prism of human rights (see for example Article 1 of the *Universal Declaration of Human Rights*, Title 1 European Union's (EU) *Charter of Fundamental Rights*).

AI technology may already be subject to the demands of dignity via existing legal frameworks. Indeed, there have been significant developments in the law during the period in which we have been working on the current project. These legal frameworks tend to approach the concept of dignity as a foundation concept - the ultimate rationale underpinning the protections offered by such instruments. This much is evident from consideration of the practical operation of such instruments. Such frameworks include national constitutions (e.g., Article 1(1) of the German Basic Law; Collings, 2024: 62), supranational instruments (e.g., Title I and Article 1 of the *EU Charter of Fundamental Rights*) and prominent international human rights instruments (e.g., Article 1 of the *Universal Declaration of Human Rights*). Moreover, some of the recently adopted regulatory AI frameworks embed human dignity as a foundational concept. The EU's groundbreaking *AI Act* references dignity several times in the recitals to the law (see recitals 27, 28, 31, 48 and 58), and further, imposes an obligation to safeguard human dignity via the requirement that high-risk AI systems are subject to a fundamental rights impact assessment, with the EU's *Charter of Fundamental Rights* serving as the relevant baseline for compliance (see Article 27 of the *AI Act*). To the extent that AI is regulated by legal conceptions of dignity, this regulation tends to invoke specific outworkings of the concept of dignity, for example, by reference to human rights obligations with respect to privacy, equality/non-discrimination, or consent and transparency etc. As discussed above, the rights-based approach evident in various examples of AI regulation is consistent with the way that the law tends to deal with the concept of dignity more generally. In the context of AI regulation, these human rights obligations may

be folded into the regulatory design via the identification of specific risks, organised into gradations of severity (low, high etc.). We see this approach reflected in the design of the EU's new *AI Act* for example.

An even more striking example is the recently finalised *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, developed under the auspices of the Council of Europe (the Framework Convention). Article 7 of the Convention creates an obligation with respect to the protection of human dignity, stating that '[e]ach Party shall adopt or maintain measures to respect human dignity and individual autonomy in relation to activities within the lifecycle of artificial intelligence systems.' Paragraph 53 of the explanatory materials provides additional insight into the intended function of Article 7, stating that "[a]ctivities within the lifecycle of artificial intelligence systems should not lead to the dehumanization of individuals, undermine their agency or reduce them to mere data points, or anthropomorphise artificial intelligence systems in a way which interferes with human dignity. Human dignity requires acknowledging the complexity and richness of human identity, experience, values, and emotions". This approach can be contrasted with the examples in the immediately preceding paragraph, in that the Framework Convention appears to more directly embed conceptions of dignity and dehumanisation proper into a legal framework.

What is interesting about this direct link to dignity and dehumanisation in the context of our exploration is the fact that generative AI chatbots such as *Replika* are designed to invite and leverage anthropomorphism on the part of the human user, promising a structurally impossible 'relationship' of 'care'. We suggest that the more traditional outworkings of the concept of human dignity reflected in most legal approaches fail to accord appropriate legal significance to the risks of dehumanisation that emerge from such technology. For example, the forms of potential dehumanisation observed in our autoethnographic analysis would not necessarily be addressed by the application of the usual suite of specific human rights protections derived from dignity. Rights with respect to equality and non-discrimination, transparency, privacy, freedom of speech, freedom of association etc. - such protections simply do not speak to the kinds of difficulties we have identified. Instead, the concept of dignity itself must be more *directly* tested against the operation of such technology to ensure responsive/complete legal coverage. In particular, it will be necessary to reckon more directly with dignity's intangible core, the protection that legal conceptions of dignity afford a humanity that is 'something understood but not actually expressed' (Pictet et al. 1958: 15 quoted in Le Moli, 2021: 180).

It is of course possible that the risks of dehumanisation associated with emerging generative AI technology may ultimately seed the humanisation of such technologies, largely as a consequence of prompting the humanisation of regulatory frameworks governing digital life. Historically, regulators have failed to show any real enthusiasm for protecting the human users of digital technologies from possible harms. There is little in the law governing digital life that speaks to the qualities that make us human, let alone the need to protect those qualities. To date, there has been some movement on the right to privacy, most notably in the EU via its General Data Protection Regulation. Privacy aside, however, the digital space has remained remarkably regulation free. Black and Murray aptly describe historical regulatory approaches to digital, especially online, services and products as "a model of regulatory weakness and ultimately failure" (Black & Murray, 2019) whereby certain prominent online platforms have become, in effect, autonomous, self-governing entities (Black & Murray, 2019, citing Laidlaw, 2015 and Reed & Murray, 2018). It would seem, however, that the risks associated with generative AI are such that various high-level bodies have begun to agitate in earnest for human rights-centric regulation. Consider, for example, the United Nations 2024 Global Digital Compact (A/RES/79/1) and the more recently established Global Dialogue on AI Governance (2025) (A/RES/79/325). It is arguable that this process of articulating, and safeguarding, humanity in the face of these developing technologies is itself a humanising experience.

The possibility that technological development might catalyse the humanisation of law, and as a consequence, the technology itself, is by no means unprecedented. This was the effect of developments in biotechnology and genetic engineering in the 1990s and 2000s, leading ultimately to outright prohibitions on human reproductive cloning (see for example Article 3 (2)(d) in the *European Charter of Fundamental Human Rights*). We believe that emerging LLM-based chatbot technology that heavily leverages anthropomorphism in its design and operation–whether it intends to or not–should provoke a similar process of reckoning within the law. In particular, communities and their legislators will need to look seriously at the above-mentioned double bind between anthropomorphisation and dehumanisation, and enquire into the possibility that AI platforms pretending towards and promising human qualities or capacities create a space in which the inherent worth of the human user is ignored, denied, exploited or otherwise diminished.

It is possible the Framework Convention discussed above may operate as a convenient vessel or vehicle for these kinds of enquiries. However, given that we are at a very early stage in the life of the instrument[1] it is too early to offer anything

[1] The Framework Convention was opened for signature in early September 2024.

other than preliminary thoughts on the possible contributions it might make to legal conceptions of dignity. Nevertheless, it is not unreasonable to suggest that the Framework Convention's overt invocation of a right to dignity, and the seeming prioritisation of dignity and rights over and above other regulatory concerns, such as promoting innovation in AI, carries with it a potential for more direct reckoning with core concepts of dignity than other AI focussed regulatory instruments currently in the mix.

Limitations and future work

In this study we used a free version of Replika for two weeks in September 2023 in an autoethnographic study using an interpretative qualitative approach to explore dehumanising aspects on a conceptual level. Newer versions and/or paid versions of this or similar rapidly evolving technologies may yield different model-generated output and thus results. Limitations inherent to interpretative qualitative research and the non-deterministic nature of LLM-based technology prohibits broad generalization.

Our results and discussion, however, may be transferable to other technologies and application domains. While our study was focused on a chatbot that aimed to represent human companionship, we also see developments for similar LLM-based chatbot services in areas like healthcare (see, for example, Ayers et al., 2023; Omiye et al., 2023; Stokel-Walker 2023). Future research into the use of LLM-based chatbots in healthcare should not only investigate the accuracy of responses and perceptions of empathy as currently done, but also dehumanising aspects of the relationship between patient and chatbot-as-healthcare-professional and the fundamental tension between anthropomorphism and dehumanisation. In this sense, we urge technology developers to reflect critically on the level of anthropomorphism and transparency in their design.

Conclusions

Our exploration of the potentially dehumanising aspects of the interaction with Replika has revealed a problematic landscape from the perspective of technology studies, moral philosophy, and legal studies. Our reflections based on a period of autoethnographic analysis point towards a subtle and indirect form of dehumanisation. Here the dehumanising aspects emerge because the human interlocutor is put within a setting that promises a deeply personal and caring interaction with the chatbot but ends up delivering an experience that is structurally devoid of reciprocity and indirectly highlights an absence of "humanity" on the AI side that diminishes the dignity of the human users who find themselves at the other end of this peculiar relationship of "care".

Our assessment takes into account that some humanising experience can, in fact, also emerge from the interaction with Replika or similar LLM-based systems. The artificial interlocutor is always available, consistently friendly or at least non-aggressive, displays some resistance to toxic discourse or echo-chamber behaviours, and is capable of presenting some pre-programmed disclaimers about the bot limits as an advisor. All these elements suggest that Replika could offer a relatively safe environment to self-aware users who seek advanced forms of human-AI interactions as a training ground for their relational capabilities or as a playground for entertainment purposes.

At present, however, by looking at the results of our analysis and what they suggest about the experience that Replika and similarly designed social chatbots can offer, the dehumanising aspects seem to outweigh the humanising ones, and raise several ethical questions about the ease of access to this service, particularly for vulnerable demographics, and lead to our further questioning from the perspective of HCI and legal studies. We briefly pointed to some promising cases that, because of their design and inherent goals, tend to support human interaction rather than substitute it in dehumanising ways. However, the potential applications of this approach in the area of social chatbots are still largely to be explored and assessed.

In general, we suggest that the new kind of human-machine interactions inaugurated by the advent of LLMs calls for a partially reformed understanding of dehumanisation, beyond the instances of users being directly designated as less-than-human. In this sense, further inquiry into the applicability of the concept of dehumanisation to user experiences with generative AI systems could prove useful both to the fields of human-computer interaction and law, as well as to the philosophical and psychological inquiry on dehumanisation. To this end, we hope the application we developed here through the engagement with the specific case of Replika can provide an early blueprint for future explorations and reformulations.

Acknowledgements The authors would like to thank Margot van der Goot who was engaged in early discussions about the lack of humanity in AI and specifically Large Language Models. The authors received no specific funding for this work.

Author contributions All authors contributed to the study conception and design. Data collection was performed by all authors, initial thematic analysis of journal entries was performed by Christiane Grünloh, and subsequently discussed with all authors. The specific disciplinary discussions were first drafted by the respective authors; "Anthropomorphism raising expectations and shaping technological frames - A human-computer interaction perspective" by Christiane Grünloh, "A new form of dehumanisation? A philosophical perspective" by Paolo Monti, and "Addressing dehumanisation - A legal perspective" by Mi-

chelle Worthington. All drafts of the manuscript were read and discussed by all authors, who all read and approved the final manuscript.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement.

Data availability The data that support the findings of this study may be available on reasonable request.

Declarations

Competing interests The authors declare no competing interests.

References

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment – What will keep systems accountable? In: *Workshop at the Thirty-First AAAI conference on Artificial Intelligence*. https://cdn.aaai.org/ocs/ws/ws0404/15216-68330-1-PB.pdf

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., Hovy, E., Ji, H., Menczer, F., Miguez, R., Nakov, P., Scheufele, D., Sharma, S., & Zagni, G. (2024). Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, *6*(8), 852–863. https://doi.org/10.1038/s42256-024-00881-z

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, *183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Bender, E. M. (2024). *Resisting Dehumanization in the Age of AI*. Current Directions in Psychological. *Science*, 1–7. https://doi.org/10.1177/09637214231217286

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445922

Black, J., & Murray, A. (2019). Regulating AI and machine learning: Setting the regulatory agenda. *European Journal of Law and Technology* 10(3). https://www.ejlt.org/index.php/ejlt/article/view/722

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI Friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, *48*(3), 404–429. https://doi.org/10.1093/hcr/hqac008

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. SAGE.

Brooks, R. (2023, February 22). I tried the Replika AI companion and can see why users are falling hard. UNSW Newsroom. Accessed October 11, 2023, from https://newsroom.unsw.edu.au/news/science-tech/i-tried-replika-ai-companion-and-can-see-why-users-are-falling-hard

Chang, C. L., Sinha, C., Roy, M., & Wong, J. C. M. (2024). AI-Led mental health support (Wysa) for health care workers during COVID-19: Service Evaluation. *JMIR Formative Research*, *8*, e51858. https://doi.org/10.2196/51858

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W.W. Norton & Company.

Cohn, M., Pushkarna, M., Olanubi, G. O., Moran, J. M., Padgett, D., Mengesha, Z., & Heldreth, C. (2024). Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models extended. Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3613905.3650818

Cole, S. (2023, January 12). 'My AI is sexually harassing me': Replika users say the chatbot has gotten way too horny. VICE. https://www.vice.com/en/article/z34d43/my-ai-is-sexually-harassing-me-replika-chatbot-nudes

Collings, J. (2024). Dignity proper and dignity plus: On the uses of dignity in german constitutional jurisprudence. In B. G. Scharffs, A. Pin, & D. Vovk (Eds.), *Human Dignity, Judicial Reasoning and the Law: Comparative Perspectives on a Key*. Constitutional Concept. Routledge.

Corrias, L. (2023). Dehumanization by law. In M. Kronfeldner (Ed.), *Routledge handbook of dehumanization* (pp. 201–213). Taylor & Francis.

Davidson, E., & Pai, D. (2004). Making sense of technological frames: Promise, progress, and potential. In B. Kaplan, D. P. Truex, D. Wastell, A. T. Wood-Harper, & J. I. DeGross (Eds.), *Information Systems Research: Relevant Theory and Informed Practice* (pp. 473–491). Springer US. https://doi.org/10.1007/1-4020-8095-6_26

Engster, D. (2019). Care ethics, dependency, and vulnerability. *Ethics and Social Welfare*, *13*(2), 100–114. https://doi.org/10.1080/17496535.2018.1533029

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological Science*, *19*(2), 114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

European Parliament (2023). Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021–2021/0106(COD). Available at https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

Frankfurt, H. (2005). *On Bullshit*. Princeton University Press.

Fritts, M., & Cabrera, F. (2021). AI recruitment algorithms and the dehumanization problem. *Ethics and Information Technology*, *23*(4), 791–801. https://doi.org/10.1007/s10676-021-09615-w

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gordon, J., & Gunkel, D. J. (2021). Moral status and intelligent robots. *The Southern Journal of Philosophy*, *60*(1), 88–117. https://doi.org/10.1111/sjp.12450

Habermas, J. (1984). *The Theory of Communicative Action* (Vol. 1). Beacon Press.

Habermas, J. (1990). *Moral Consciousness and Communicative Action*. Polity.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38. https://doi.org/10.1007/s10676-024-09775-5

Italian Data Protection Authority (Garante per la protezione dei dati personali) (2023, February 2). *Provvedimento del 2 febbraio 2023 [9852214]*. https://www.garanteprivacy.it. Retrieved October 11, 2023, from https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852214#english

Kittay, E. F. (2007). Beyond autonomy and paternalism: The caring transparent self. In T. Nys, Y. Denier, & T. Vandevelde (Eds.), *Autonomy and paternalism. Reflections on the theory and practice of health care* (pp. 23–70). Peeters.

Kittay, E. F. (2011). The ethics of care, dependence, and disability. *Ratio Juris*, *24*(1), 49–58. https://doi.org/10.1111/j.1467-9337.2010.00473.x

Klowait, N. (2018). The quest for appropriate models of human-likeness: anthropomorphism in media equation research. *AI & Society*, *33*(4), 527–536. https://doi.org/10.1007/s00146-017-0746-z

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (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 & Society*, *26*(10). https://doi.org/10.1177/14614448221142007

Laidlaw, E. (2015). *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility*. Cambridge University Press.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1469

Lee, K. S., Yeung, J., Kurniawati, A., & Chou, D. T. (2025). Designing human-centric ai mental health chatbots: A case study of two apps. In A. Iglesias, J. Shin, B. Patel, & A. Joshi (Eds.), *Information Systems for Intelligent Systems. ISBM 2024. Lecture Notes in Networks and Systems* (Vol. 1255). Springer. https://doi.org/10.1007/978-981-96-1747-0_36

Leget, C. (2013). Analyzing Dignity: A Perspective from the Ethics of Care. *Medicine Health Care and Philosophy*, *16*(4), 945–952. https://doi.org/10.1007/s11019-012-9427-3

Le Moli, G. (2021). *Human dignity in international law*. Cambridge University Press.

Ma, Z., Mei, Y., & Su, Z. (2023). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *arXiv preprint arXiv*:230715810. https://arxiv.org/abs/2307.15810

Miller, S. C. (2017). Reconsidering Dignity Relationally. *Ethics and Social Welfare*, *11*(2), 108–121. https://doi.org/10.1080/17496535.2017.1318411

Monti, P. (2024). AI enters public discourse. A Habermasian assessment of the moral status of Large Language Models. *Ethics & Politics*, *26*(1), 61–80. https://doi.org/10.13137/1825-5167/36469

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, Massachusetts, USA. https://doi.org/10.1145/191666.191703

Nussbaum, M. C. (2000). *Women and Human Development. The Capabilities Approach*. Cambridge University Press.

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *npj Digital Medicine*, *6*(1), 195. https://doi.org/10.1038/s41746-023-00939-z

Orlikowski, W. J., & Gash, D. C. (1994). Technological frames: Making sense of information technology in organizations. *ACM Trans Inf Syst*, *12*(2), 174–207. https://doi.org/10.1145/196734.196745

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior*, *140*, 107600. https://doi.org/10.1016/j.chb.2022.107600

Redaelli, R. (2023). Different approaches to the moral status of AI: A comparative analysis of paradigmatic trends in science and technology studies. *Discover Artificial Intelligence*, *3*(1). https://doi.org/10.1007/s44163-023-00076-2

Reed, C., & Murray, A. (2018). *Rethinking the Jurisprudence of Cyberspace*. Edward Elgar.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people* (pp. 19–36). Cambridge University Press.

Replika (2023, February 7). Terms of Service. Accessed October 11, 2023, from https://replika.com/legal/terms

Replika (2023). (n.d.b). *What is Replika?* Accessed October 11, from https://help.replika.com/hc/en-us/articles/115001070951-What-is-Replika-

Replika (n.d.a). *How does Replika work?* Accessed October 11, (2023). from https://help.replika.com/hc/en-us/articles/4410750221965-How-does-Replika-work-

Replika (n.d.c). *What is Replika pro?* Accessed October 11, (2023). from https://help.replika.com/hc/en-us/articles/360032500052-What-is-Replika-Pro-

Scheutz, M. (2011). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, 205-221.

Schwitzgebel, E. (2023). AI systems must not confuse users about their sentience or moral status. *Patterns*, *4*(8), 100818. https://doi.org/10.1016/j.patter.2023.100818

Shin, H. I., & Kim, J. (2020). My computer is more thoughtful than you: Loneliness, anthropomorphism and dehumanization. *Current Psychology* 39, 445–453 (2020). https://doi.org/10.1007/s12144-018-9975-7

Singleton, T., Gerken, T., & McMahon, L. (2023, October 6). How a chatbot encouraged a man who wanted to kill the Queen. BBC News. Accessed November 9, 2023, from https://www.bbc.com/news/technology-67012224

Sinnott-Armstrong, W., & Conitzer, V. (2021). How much moral status could artificial intelligence ever achieve? In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status* (pp. 269–289). Oxford University Press. https://doi.org/10.1093/oso/9780192894076.003.0016

Stokel-Walker, C. (2023, August 3). Convicted fraudster Martin Shkreli is touting a medical AI chatbot—much to experts' concern. Fast Company. Accessed October 11, 2023, from https://www.fastcompany.com/90932968/martin-shkreli-dr-gupta-sasha-luccioni

Strohmann, T., Siemon, D., Khosrawi-Rad, B., & Robra-Bissantz, S. (2023). Toward a design theory for virtual companionship. *Human–Computer Interaction*, *38*(3–4), 194–234. https://doi.org/10.1080/07370024.2022.2084620

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719). https://doi.org/10.1126/science.adq2852

Tong, A. (2023a, March 21). What happens when your AI chatbot stops loving you back? Reuters. Accessed October 11, 2023 from https://www.reuters.com/technology/what-happens-when-your-ai-chatbot-stops-loving-you-back-2023-03-18

Tong, A. (2023b, March 25). AI chatbot company Replika restores erotic roleplay for some users. Reuters. Accessed October 11, 2023 from https://www.reuters.com/technology/ai-chatbot-company-replika-restores-erotic-roleplay-some-users-2023-03-25/

Tronto, J. C. (1998). An ethic of care. *Generations: Journal of the American society on Aging* 22(3): 15–20.

United Nations General Assembly (2024). *Pact for the Future*. (A/RES/79/1) Annex 1, Global Digital Compact. https://www.un.org/pact-for-the-future/en

United Nations General Assembly (2025). *Global Dialogue on AI Governance*. (A/RES/79/325) https://www.un.org/global-dialogue-ai-governance/en

Vallor, S., & Vierkant, T. (2024). Find the Gap: AI, Responsible Agency and Vulnerability. *Minds & Machines*, *34*, 20. https://doi.org/10.1007/s11023-024-09674-0

van der Goot, M. J. (2022). Source orientation, anthropomorphism, and social presence in human-chatbot communication: how to proceed with these concepts. *Publizistik*, *67*, 555–578. https://doi.org/10.1007/s11616-022-00760-w

van Wynsberghe, A. (2022). Social robots and the risks to reciprocity. *AI & Society*, *37*(2), 479–485. https://doi.org/10.1007/s00146-021-01207-y

Volpe, A. (2025). Toward an artificial deliberation? On Google DeepMind's Habermas Machine. *Ethics & Information Technology*, *27*(45). https://doi.org/10.1007/s10676-025-09854-1

Waldron, J. (2012). Dignity and Rank. In M. Dan-Cohen (Ed.), *Dignity, Rank and Rights*. Oxford University Press.

Waldron, J. (2015). Is dignity the foundation of human rights? In R. Cruft, S. M. Liao, & M. Renzo (Eds.), *Philosophical foundations of human rights*. Oxford University Press.

Waytz, A., Cacioppo, J., & Epley, N. (2010b). Who sees human? The stability and importance of individual differences in Anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., Epley, N., & Cacioppo, J. (2010a). Social cognition unbound. *Current Directions in Psychological Science*, *19*(1), 58–62. https://doi.org/10.1177/0963721409359302