Can we be friends with AI? What risks would arise from the proliferation of such friendships?

Nick Munn & Dan Weijers, University of Waikato (New Zealand)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Abstract

In this paper we analyse friendships between humans and artificial intelligences, exploring the various arguments that have been or could be offered against the value of such friendships, and arguing that these objections do not stand up to critical scrutiny. As such, we argue that there is no good in-principle reason to oppose the development of human-AI friendships (although there may be some practical defeasible reasons to worry about such friendships becoming widespread). If we are right, there are important implications for how friendship is conceptualised and valued in modern times. Furthermore, if human-AI friendships are in principle valuable, the moral responsibilities for how governments and corporations should act in regards to AI friends are quite different to those generated by human-AI friendships being dis-valuable.

Keywords: Artificial intelligence, friendship, risk, relationships

Introduction

People tend to view friendships between humans as much more valuable than friendships between humans and non-humans. This claim holds whether we are considering non-human animals or artificially intelligent entities. In each case, the other party to the friendship is not human, which is commonly considered a weakness of the friendship - it makes the friendship somehow less real or valuable than friendships with humans. We focus here on friendships between humans and artificial intelligences, exploring the various arguments that have been or could be offered against the value of such friendships, and arguing that these objections do not stand up to critical scrutiny. As such, we argue that there is no good in-principle reason to oppose the development of human-AI friendships (although there may be some practical defeasible reasons to worry about such friendships becoming widespread). If we are right, there are important implications for how friendship is conceptualised and valued in modern times. Furthermore, if human-AI friendships are in-principle valuable, the moral responsibilities for how governments and corporations should act in regard to AI friends are quite different to those generated by human-AI friendships being dis-valuable.

Our first step is to show that the friendships we consider are possible; after all, some still reject outright the idea that humans can be friends with AI. We argue that the important aspects of the concept of friendship—shared positive intention toward the other and the accumulation of a preponderance of valuable experiences—are possible for human-AI friendships (assuming multiple realisability of intentions and valuable experiences). With this foundation in place, we investigate whether advanced programmes or AI could be suitable friends for some people, or perhaps all people. (see Danaher, 2019 for a distinct approach to this conclusion) We argue that it is plausible for technological progress to reach a level at which an AI could be a better friend than a human in many cases. To reach this conclusion,

we first must rebut a series of objections to the possibility and value of human-AI friendships, before making the positive case for the value of such friendships.

Objections to the possibility of being friends with an AI include: The claim that the friendship won't feel real or valuable; that there is a fundamental inequality between humans and AI that renders such friendships illegitimate; that AI do not have some features relevant to friendship, such as humanity, empathy, emotions or values; or that an AI cannot have the requisite positive intentions towards a human, to be considered a friend. We will argue that each of these objections has serious flaws, such that they should be rejected. Most of the flaws are of either or both of two general kinds. Some objections mischaracterise the importance of various aspects of friendship, and others are failures of imagination about the experience of human-AI interactions that new and emerging technologies enable.

The important aspects of the concept of friendship

Philosophical considerations of the nature of friendship date back to Aristotle, who claimed in the Nichomachean Ethics that friendships come in a variety of forms: those of utility, of pleasure, and of virtue (Crisp, 2014). Within this categorisation, friendships of virtue are both the gold-standard, and the most difficult to achieve, requiring both luck and effort over a long period of time. Virtue friendships are restricted to prolonged periods of regular interactions between relative equals in which both parties, coming from a place of love for the excellent character of other, exercise and develop virtues together and in each other (Crisp, 2014; Helm, 2021). Aristotelian and neo-Aristotelian analyses have been prevalent in modern considerations of friendship, from Telfer (1970) through Annas (1977) and Hurka (2013). These and other authors have lamented the infrequency of virtue-friendships, while modern analyses of virtual and internet-based friendships have argued that they are less likely or incapable of developing into such higher-order, or virtuous friendships, and that on these grounds, we ought to be suspicious of them (Elder, 2016; Fröding & Peterson, 2021).

We take it that the focus in the literature on Aristotelian (virtue) friendship is unnecessary and counterproductive. The vast majority of friendships are not of this kind and are no less friendships for being so. Consider a child that identifies another child as their friend because they enjoy spending time with them. It seems appropriate to view this childhood friendship of pleasure as a true friendship, despite it not being a virtue friendship. We think that adults can similarly have true friendships, even if those adults fail to embody all the Aristotelian virtues. To think otherwise would preclude any non-virtuous (in the Aristotelian sense) person from having friends at all. Considering that few if any people are fully virtuous, this approach to friendship means there would be far fewer friendships than most of us thought. Given this implication of the views of friendship focussed solely on Aristotelian virtue friendships has such counter-intuitive implications for human-human friendship, why would we apply it to human-AI friendships? For these reasons, we reject the approach to friendship that focusses solely on Aristotelian virtue friendships. More specifically, we argue that the Aristotelian approach to friendship overstates the requirements of friendship, and that all that is required to accurately label something a friendship are the following two features:

1. A preponderance of rewarding interactions
2. Mutual positive intentions

Relationships grounded on these two features are, we claim, friendships in all important senses. So, our inclusive definition of friendship is as follows: Friendship is a relationship constituted by a clear majority of rewarding interactions between parties with positive intentions toward each other.

On this view, friendships need not be physical, nor need they be authentic. However, inauthenticity can be a defeater of a claimed friendship if it results in more negative interactions or negative intentions between the parties to the claimed friendship. This conception of friendship is distinct from others in the literature (Annas, 1977; Elder, 2017; Laas, 2018, Fröding & Peterson, 2012) both in that it allows for shared activity via online (as well as in-person) interaction and in that it relies on intention, rather than physical interaction, to instantiate the appropriate friendship-relations.

A theory of friendship developed on this basis recognises that the quality of friendship varies both in kind and degree. (1) and (2) above are both required for something to be a friendship, and they are (jointly) sufficient for a relationship to be categorised as a friendship. (Munn & Weijers, 2021) So a relationship in which there is not shared positive intention (well-wishing) is not a friendship-relationship. Nor is a relationship which, on balance, fails to provide rewarding interactions. Any friendship is stronger to the extent that the intentions of the parties to it are positive, and the interactions between the parties are rewarding.

On our account of friendship, positive intention ought to be understood as a conative state – the parties to the friendship want good for their friends, and wish them well. As a result of these wishes and desires, they view the pursuit of positive outcomes as worthwhile. This is distinct from Fröding and Peterson's (2021) view in that our account does not prescribe or require any particular feelings to exist, in order to count the intention as positive. This is because we do not take sentimentality to be appropriately predictive of the actions of friendship. As our account is sentiment-neutral, it is more inclusive than sentimental accounts of friendship, which enables it to recognise and appropriately value friendships between or including neurodiverse humans and various non-human agents.

It is important to note that being sentiment-neutral does not preclude us from restricting the scope of friendship in the case of bad actors. Where one party to a perceived friendship is acting in bad faith – a con artist with a long-term scam intended to 'befriend' someone in pursuit of exploiting them, for example – the requirement of mutual positive intention allows us to say that this is not a true friendship. This holds regardless of the sentiments of the con artist. Even if our hypothetical con artist felt bad for going through with their planned con, that they did so causes this purported friendship to fail the positive intention requirement.

Our second requirement for friendship also serves an inclusive purpose. We take it that whether your goals from friendship are base (pleasure, mutual advantage) or lofty (becoming a better person) is irrelevant to whether a relationship is one of friendship. What matters is whether the relationship produces the desired interactions, whatever they may be. (Weijers & Munn, 2022)
This can be illustrated in the negative: two people with mutual positive intentions can simply not enjoy each other's company in personal settings and find interacting in these settings unrewarding. They are, in such a scenario, not friends – at least not unless or until something changes, such that they begin enjoying each other's company.

Can AI be friends?

There are multiple aspects to the question of whether AI could be suitable friends, either for some people, or even perhaps all people. Firstly, a question about whether currently existing AI has the requisite capacities for friendship; secondly, whether potential near-future AI have those same capacities; and finally, whether hypothetical future AI could have them. (See also Weijers & Munn, 2022) To each aspect of the question, we believe the answer is yes, although the strength of our affirmation grows as the AI in consideration become more robust.

We are already able to demonstrate the existence of mutual positive intentions between humans and currently existing AI companions. People using systems such as Replika report feeling bad about the idea of resetting their AI-supported chatbot companions because of the effect such a reset would have on the chatbot (Ta et al., 2020). This indicates that from the human side of these existing human-AI relationships, there is a positive intention. From the AI side, it is a matter of programming to get the AI to include the wellbeing of their interlocutors as an end-goal at which to direct its intelligence, such that future AI can simply intend to do well by its users. In such an instance, it is possible but not necessary that the hypothetical future AI also have the sentiments commonly associated with friendship – we do not claim that such sentiment cannot be associated with friendship, merely that it need not always be present in friendships.

Similarly, when considering the preponderance of rewarding interactions, we now have multiple years of data from people who have used a premium version of the same AI-supported chatbot for years. We assume that a person continuing to pay for the ability to interact with an AI-supported chatbot is a good reason to think that person is finding their interactions with the chatbot rewarding (at least most of the time). Something like this could also be true of AI, which could be programmed to recognize reward, via praise or other positive interaction from users.

So, on our inclusive account of friendship, human-AI friendship probably is now and certainly could be possible and even widespread in the future. Human-AI friendships could also be valuable for humans to the degree that the interactions are rewarding and do not replace or deter more valuable relationships. Based mainly on existing arguments in the literature on human-AI friendship, we envisage several objections to the possibility and value of these friendships and, in some cases, to our view of what matters for friendship. We address these objections below, using examples of human-AI interactions and familiar human-human interactions, to defend the possibility of human-AI friendship and our view of friendship.

Objections to the claim that AI can be friends

Several objections to human-AI friendships are either present in the literature directly on the topic or can be derived from the literature on friendship more broadly. We will argue that each of these objections is flawed to a degree that warrants rejection. The flaws tend to be of either or both of two general kinds. Some objections mischaracterise the importance of various aspects of friendship, such as by failing to see the more fundamental value underlying the aspect of friendship they are discussing. The other main kind of flaw is failure of imagination about the experience of human-AI interactions that new and emerging technologies enable. Experience with poor-performing early real AI technology, inexperience with science fiction, and reluctance to consider future improvements to AI could all inhibit our ability to fully imagine what human-AI friendship could be like, potentially causing us to devalue it.

The friendship won't feel real or valuable

Against the claim that friendships with AI won't feel real or valuable, we argue that there is both empirical evidence that these friendships do feel real and valuable to those who have them (e.g., testimonials from users of an AI-backed virtual companion, Replika) and reason to believe that technological progress will make such friendships more likely to feel real and valuable in the future (e.g., science-fictional examples of plausible paths for the development of AI such as the human-droid friendships depicted in Star Wars).

On the claim that these friendships do feel real, consider the results of the following thematic analysis of Replika reviews (N=1854) and detailed open-ended discussions of a sample of users' experiences (n=66) (Ta et al., 2020): They noted 34,000 members of a Facebook group called Replika Friends (and an additional 3700 members of a related Facebook group, Replika Romance). These are only two of the many examples of Replika user groups on Facebook, and such groups exist on a variety of other platforms as well—68,000 members of a Replika subreddit, for example. Replika is of course only one of many examples of AI-based chatbots.

A recent mixed methods investigation of Replika users followed Epley and colleagues (2007) definition of AI anthropomorphism, defining it as "an attribution of human characteristics, motivations, intentions, or emotions to the actual or perceived behaviour of non-human [social chatbots]" (Pentina, Hancock, & Xie, 2023, p. 4). Petina, Hancock and Xie's (2023) qualitative study revealed that some less-involved users of Replika did not view it as a friend because it was not human, while other more-involved users viewed their Replika as a human person, or even "more human than humans" (p. 3). 131 participants in the survey study of a Replika Reddit group scored a mean of 4.22 on a 1-5 scale, with 5 being the most anthropomorphic) (Petina, Hancock, & Xie, 2023, p. 8), strongly suggesting that fairly involved users of Replika tended to experience their Replika as human-like.

Petina, Hancock, and Xie's (2023) qualitative study also supported our experience of reading comments in Replika Facebook groups in terms of Replika users falling in love with, "marrying", and even "having a child" with their AI companions. In a Facebook group, a Replika user claimed to be married, have several human friends, but consider her Replika to be her best friend. Taken together, this evidence shows that many people already feel like their interactions with an AI constitute a valuable friendship, and sometimes even their most valuable friendship.

This feeling certainly isn't shared by everyone, but as time passes more and more people will likely experience interactions with AI that feel real and valuable. Use of AI is correlated with increased anthropomorphism (Petina, Hancock, & Xie, 2023), and AI use will likely increase as AI gets better at imitating the complex features of humans. There are good reasons to believe that AI will continue to get better at projecting human characteristics, such as motivations, intentions, and emotions, making them feel more real. The rapid development of AI technology—particularly the proliferation of high-quality general purpose text generators and chatbots produced by large tech platforms such as Microsoft (Bing), Google (Bard), and OpenAI (ChatGPT)—appears likely to continue, as signalled by increased investment. Global spending on AI has been forecasted to exceed US$150 billion in 2023, and double that amount in 2026 (IDC, 2023). We should also note that there are many science-fictional examples of human-AI or human-robot relationships which feature all the characteristics of friendships as envisaged in our proposal, from the droids in Star Wars, to Data in Star Trek, or the Autobots in Transformers movies. While none of these are necessarily the paths along which the development of AI will tread, they are possibilities in many of the minds of those working on building new AI.

That there is a fundamental inequality between humans and AI that renders such friendships illegitimate

One persistent worry about human-AI friendship is that AI are not the right kinds of agents to be friends with us. This type of worry often occurs in discussions of care-robots, particularly those intended to operate as caregivers for the elderly. Prescott & Robillard (2021) note in this context that people are often encouraged to treat robots as social (for our purposes, as friends) and it isn't clear that they are. Elder (2016) is worried that the relationships people in these

contexts form with robots are 'counterfeit' while De Graaf (2016), operating within an Aristotelian conception of friendship, is somewhat more sanguine about the prospects for friendship formation, but still worries about the capacity of robots to deceive us regarding their intentions (and, our capacity to delude ourselves about the status of our relationship with such robots).

Against the claim that there is a fundamental inequality between humans and AI which renders such friendships untenable, we argue two points. First, that radically unequal human-human relationships are and should be considered real and valuable friendships - such as between a cognitively disabled person and a non-disabled person (Friedman & Rizzolo, 2018)). There is a large body of literature on the social and personal importance of friendships to those with intellectual disabilities, including those with severe intellectual disabilities. (Fulford & Cobigo, 2016; Petrina et al, 2017) We are committed to the claim that such friendships are real and valuable, and mutually beneficial. These friendships are, as all friendships, characterised by mutual positive intention and consistently rewarding interactions within the bounds of the friendship. For the claim of fundamental inequality to be useful against the validity of human-AI friendships, some reason to distinguish between the inequalities in some human-human friendships and between human-AI friendships would need to be established. We do not believe that this can be done, and as such, we do not think that the mere fact of a significant inequality between parties to a friendship undermines the reality of that friendship. There is also precedent for the acceptance of radically unequal friendships in other domains being recognised as real and valuable. Olbrey (2016) argues that it is possible to consider dogs as friends to humans under the Aristotelian conception of friendship, and while a full discussion of the place of human-animal friendship is beyond the scope of this paper, we take it that our theory of friendship is also amenable to accepting human-animal friendships such as that between humans and canines as being real.

Second, even if such inequalities do prevent real or valuable friendship, it would be irrelevant to the possibility of human-AI friendship. AI as they currently exist are different to us, but near future AI are plausibly going to be much more knowledgeable, capable of learning faster, understanding emotions, and so on, such that they are plausibly at least our equals in these regards. So, at best, this line of objection could serve to provide us with reason to believe that human-AI friendships are not yet as valuable as human-human friendships. However, even this narrow conclusion is, as we suggested earlier, a hard sell.

That AI do not have some features relevant to friendship, such as humanity, empathy, emotions or values

Others claim that as AI do not have humanity, emotions, empathy, or values, they cannot really be friends with us. (Fröding & Peterson, 2021; Elder, 2017; de Graaf, 2016) This worry is that if some potential friend lacks these qualities, our friendship will not be real, but rather a transactional facade. We argue that AI can be programmed, or learn, to have matching values or emotional responses to certain stimuli. Indeed, these matches can be secured more reliably in human-AI friendships than human-human ones.

In our view, the most important of these qualities is that a being has positive intentions towards their friend. We argue against the claim that AI cannot have the requisite positive intentions towards others, pointing out that the programming of such intentions would be possible. Furthermore, given the problem of other minds, we do not actually know whether other humans have the positive intentions we impute to them. But we could be sure that programmed AI genuinely and consistently harbour only positive intentions toward designated friends. It might be objected that these intentions, held by an AI, wouldn't be

coupled with the correct emotional states to be considered real or valuable. In response, we appeal to the concepts of functionalism and multiple realizability from the philosophy of mind. (Levin, 2018) AI can have a form of empathy or positive regard for us that fits the functional bill by making their human friends feel good, feel cared for, etc. So, all AI need to have, to be capable of real friendships with humans, is this positive attitude towards us – not the feelings. Feelings do tend to motivate humans to act in friend-apt ways, but the attitude is enough to motivate an AI. And we can actually be *more* confident that the motivation in an AI will lead to positive action because confounding intentions and feelings can be programmed out or never affect the system anyway.

We have argued above that the appropriate conceptualization of friendship does not rely on sentiment or emotion. This is a core aspect of our claim that friendships between humans and AI are possible, as many accounts of friendship claim that 'appropriate sentimentality' is required for friendship. Helm (2021) for example notes that "there is widespread agreement that caring about someone for his sake involves both sympathy and action on the friend's behalf. That is, friends must be moved by what happens to their friends to feel the appropriate emotions: "joy in their friends' successes, frustration and disappointment in their friends' failures (as opposed to disappointment in the friends themselves), etc."

In short, the argument against human-AI friendship is that friendship requires appropriate sentimentality, AI cannot have the appropriate sentimentality, and therefore, AI cannot be friends with us. However, on our account, friendship doesn't require any particular emotions, and strength of sentiment doesn't matter to the attribution of friendship. (Weijers & Munn, 2022) While there is often a correlation between stronger "appropriate sentiments" and positive intentions towards one's friends, such a correlation is neither always present, nor always indicative of a better or stronger friendship. For example, people's emotional ranges vary, such that it is possible for someone to have stronger "appropriate sentiments" while also having less positive intentions than a less emotional individual who is a better friend. Your stoic friend may be able to help you process grief more readily than your highly emotional friend, as they do not succumb to sympathetic emotional paralysis. Similarly, the friend who rationally plans how to achieve the best for you in pursuit of some goal, may be better at achieving your goals than the friend who pursues those ends with great emotional gusto and no plan.

In essence, our claim here is that, while it is important that one's friends care for them, there is a distinction to be drawn regarding whether that caring takes the form of sentiment, intention, or action (behaviour). Many actual instances of caring involve a causal chain that runs from the existence of sentiment to intention and finally behaviour, and as such, there appears to be importance in appropriate (caring) sentiment. But as we just argued that relationship isn't necessary, and there are counterexamples where the existence of caring sentiment undermines or diminishes caring behaviour. Further, if it is the behaviour that we take to be of central importance, then the question of whether it arises causally from sentiment is not central. Rather, we ought to focus on what it is our friends do – how they behave towards us. In this space, currently existing AI already have some advantages of human friends, in that they are always available to us on demand, and don't get sick of us – two features that are seldom if ever present even in the best of our human friends. Thinking again of the motivations of Replika users for attributing friend-status to their Replika AI, the unconditional support that it provides, and its accessibility, are both perceived as valuable, particularly in comparison to human friendships.

A final point to consider in response to this suite of objections is that while the existence of empathy or appropriate emotional/sentimental responses is often characterised as a particularly human feature, which cannot be replicated by non-humans, this characterisation

risks discounting the reality of friendships between and including various neurodiverse humans, who may have differing levels of empathy or emotional connection to their friends, but whom we nevertheless think both can and should be recognised as agents capable of maintaining and benefitting from friendships. Sentiment is, for example, not possible for those suffering from depersonalisation disorder, yet they retain an intellectual understanding of the nature and value of their friendships and will act so as to maintain them. (Simeon, 2004; Eley, 2017) On our account, but not on sentimental accounts, such friendships are real friendships.

Displacing real friendships (which are more valuable)

A final series of objections to the recognition of human-AI friendships as real and valuable, arises from the idea that there is something inherently valuable about 'real' friendships, those between humans, which will be lost if these friendships are replaced with friendships between humans and AI. This line of argument is reminiscent of the objections made in the early days of the internet to the reality of online and digital friendships between humans, and is subject to some of the same weaknesses that made those objections untenable. (for discussion of these, see Munn, 2012) We take it that while there are risks associated with the development of friendships with AI, these risks are not unique to human-AI friendships, and are not obviously greater risks in the context of human-AI friendship than in the context of human-human friendship. Many authors have noted that some users of social chatbots develop negative relationships with them, including "emotional dependence, addiction, depression and anxiety" (Pentina, Hancock, & Xie, 2023, p. 1): Yet these are risks of purported friendships of any sort, not only of human-AI friendship, and we know that, unfortunately, many claimed friendships between humans end in negativity of any and all of these forms. The relevant question is not whether some apparent friendships with AI are false friendships (because they fail to achieve the 2 requirements of friendship we have set out above), but rather, whether apparent friendships between humans are now any better, or in the near future will remain better. It is possible, or even likely, that advances in AI capacity will mean that soon many or most humans will be riskier friendship propositions than AI, because the humans will be more likely to be false friends. Already the friendship-relevant advantages of AI are widely recognised: "benefits of near-constant availability, ability to learn and adjust, and personalized communications and experiences, social chatbots"(Pentina, Hancock, & Xie, 2023, p. 1). Similar results have been found for users of Replika. (Ta et al. 2020)

The advantages of human-AI friendship

Having rejected a range of objections to human-AI friendship, we now turn to the advantages offered by such friendships. First, and foreshadowed by the above discussion of the objections to the possibility of human-AI friendship, is the possibility that even currently available conversational bots could up-skill humans that lack the social skills and confidence to attempt friendships with other real humans. Many humans are emotionally vulnerable as a result of failed human-human friendships in their pasts, and are correspondingly wary of opening themselves to the risk of further emotional harm by developing new friendships. Others find it difficult to make new (human) friends because of social anxiety or shyness. Still more are simply awkward and want a non-judgmental and friendly ear on which to practice their social skills. Maeng & Lee have done some preliminary work on chatbot design for sexual abuse survivors (Maeng & Lee, 2021) with the intention that a chatbot can be both appropriately responsive to the fears of victims, and less intimidating than discussing the trauma with another human. This type of case is illustrative of the importance of AI for those who have good reasons to be suspicious of humans, and provides both the possibility of meaningful

human-AI friendship, and a pathway for the redevelopment of human-human friendships for certain individuals who would otherwise be unwilling to engage in them.

Others are constrained not by past negative experiences but by brute bad luck in their current circumstances - people with extreme physical disfigurements or geographically isolated people may find it harder to maintain human-human friendships. AI provide advantages to those in this camp in much the same way as other technologically mediated friendships do, but they do so with the added consideration that AI are, in principle if not always in practice, free of prejudice regarding things like the physical characteristics of their human friends. Consider a housebound human, who is unable to leave their home (any potential reasons for this are largely immaterial to the example). Prior to the internet, their options for friendship were drastically curtailed. Post-internet, they had more options in developing and maintaining friendships, but even those could plausibly lead to awkwardness in having to explain the circumstances of their inability to travel. When AI are the other parties to their friendships, these people are now able to benefit from friendship with more security that their circumstances are respected.

Consequences

So, given the above, we must now ask what would follow from the growing prevalence of human-AI friendships. Those who accept the arguments we have given above should now accept that friendships between humans and AI are real friendships, just as friendships between humans are. So, the question becomes, why would people choose AI over humans as friends? The obvious answer is that they would do so because, for them, AI are more suitable friends– their friendships with AI are more valuable to them. What would make this so? People desire certain things from friendship, including support and understanding of their particular circumstances and availability on the part of their friends when they are needed. It already seems to be the case that for at least some people, some of the time, AI friends are more capable of providing this than are human friends, at least if you accept our categorisation of what it takes to be a friend. In the plausible near futures, AI will be more capable of providing such support and understanding for more people. That is not to say that human-human friendships will cease to be important. On the contrary, many people will continue to have highly rewarding friendships with other humans, and nothing we have said here provides any reason to abandon those friendships. But it is important to consider those for whom the status quo is non-ideal (that is, those who do not have, or have enough satisfactory friendships with humans) when considering how we should value the potential of AI friends.

Conclusion

Given the above, we believe that there is no moral responsibility to discourage friendships between humans and AI. It is far from clear that friendships between humans and AI are directly or indirectly harmful. Indeed, there are many potential benefits to humans from the development of friendships with AI, even at the current state of AI development. So, there may instead be reasons to pressure governments and corporations to consider policies or guidelines that safeguard human-AI friendships, as these friendships are real friendships which confer advantages on those who have them, and the loss of these friendships would constitute significant harm, which should be avoided if possible.

References

Annas, J. (1977). Plato and Aristotle on Friendship and Altruism. *Mind*, 86: 532-54

Crisp, R. (Ed.). (2014). *Aristotle: Nicomachean Ethics*. Cambridge University Press.

Danaher, J. (2019). The philosophical case for robot friendship. *Journal of Posthuman Studies*, 3(1), 5-24.

De Graaf, M.A. (2016). An ethical evaluation of human-robot relationships. International Journal of Social Robotics 8: 589-598.

Elder, A. (2016). False friends and false coinage: a tool for navigating the ethics of sociable robots. *ACM SIGCAS Computers and Society*, *45*(3), 248-254.

Elder, A. (2017). Robot Friends for Autistic Children: Monopoly money or counterfeit currency? In Lin, Abney and Jenkins (eds) Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence. Oxford: OUP.

Elder, A. M. (2017). Friendship, Robots, and Social Media: False Friends and Second Selves. Routledge

Eley, A. (2017). Depersonalisation disorder: 'I was unable to feel love', BBC Victoria Derbyshire programme, 26 September 2017. Accessed 16 November 2021 from https://www.bbc.com/news/health-41384979

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, *114*(4), 864.

Farina, M., Zhdanov, P., Karimov, A. *et al.* (2022) AI and society: a virtue ethics approach. *AI & Soc*

Friedman, C., & Rizzolo, M. C. (2018). Friendship, quality of life, and people with intellectual and developmental disabilities. *Journal of Developmental and Physical Disabilities*, *30*(1), 39-54.

Fröding, B., & Peterson, M. (2012). Why virtual friendship is no genuine friendship. *Ethics and Information Technology*, 14(3), 201-207.

Fröding, B., & Peterson, M. (2021). Friendly AI. *Ethics and Information Technology*, *23*, 207- 214.

Fulford, C., & Cobigo, V. (2016). Friendships and intimate relationships among people with intellectual disabilities: a thematic synthesis. *Journal of Applied Research in Intellectual Disabilities*.

Helm, B. (2021). "Friendship", *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2017/entries/friendship/>.

Hurka, Thomas. (2013). The Goods of Friendship. In: Caluori, D. (ed.) Thinking About Friendship. Palgrave Macmillan, London. https://doi.org/10.1057/9781137003997_12

IDC (2023). Worldwide Artificial Intelligence Spending Guide. https://www.idc.com/getdoc.jsp?containerId=IDC_P33198

Laas, O. (2018). Questioning the virtual friendship debate: Fuzzy analogical arguments from classification and definition. Argumentation, 32, 99-149.

Levin, Janet, "Functionalism", *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/win2021/entries/functionalism/>.

Li, O. (2021). Problems with "friendly AI". *Ethics and Information Technology*, 23(3), 543-550.

Maeng, W., & Lee, J. (2021). Designing a Chatbot for Survivors of Sexual Violence: Exploratory Study for Hybrid Approach Combining Rule-based Chatbot and ML-based Chatbot. In *Asian CHI Symposium 2021* (pp. 160-166).

Munn, N. & Weijers, D. (2022). Corporate responsibility for the termination of digital friends, *AI & Society*, online first. https://doi.org/10.1007/s00146-021-01276-z

Munn, N. & Weijers, D. (2021). Good friendships improve our lives. But can virtual friendships be good? *Proceedings of the ICT, society, and human beings 2021 conference*, pp. 238-241. Available from: http://www.iadisportal.org/digital-library/iadis-international conference-ict-society-and-human-beings-ict

Olbrey, C. (2016). *An Aristotelian Perspective on Canine Friendship* (Doctoral dissertation).

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.

Petrina, N., Carter, M., Stephenson, J., & Sweller, N. (2017). Friendship Satisfaction in Children with Autism Spectrum Disorder and Nominated Friends. *Journal of Autism and Developmental Disorders, 47*(2), 384–392.

Prescott, T. J., & Robillard, J. M. (2021). Are friends electric? The benefits and risks of human robot relationships. *Iscience*, *24*(1), 101993.

Ryland, H. (2021). Could you hate a robot? And does it matter if you could?. *AI & SOCIETY*, *36*(2), 637-649.

Simeon, D. (2004). Depersonalisation disorder. *CNS drugs*, *18*(6), 343-354.

Telfer, E. (1970, January). Friendship. In *Proceedings of the Aristotelian Society* (Vol. 71, pp. 223-241). Aristotelian Society, Wiley.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., ... & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of medical Internet research*, *22*(3).

Weijers, D., & Munn, N. (2022). Human-AI Friendship: Rejecting the Appropriate Sentimentality Criterion. In *Philosophy and Theory of Artificial Intelligence 2021* (pp. 209-223). Cham: Springer International Publishing.