Weijers, D., & Munn, N. (2022). Human-AI Friendship: Rejecting the 'appropriate sentimentality' criterion, in Vincent C. Müller (ed.), *Philosophy and theory of artificial intelligence 2021* (SAPERE) 63. (Cham, Switzerland: Springer), 209-223. https://link.springer.com/chapter/10.1007/978-3-031-09153-7_17

Human-AI friendship: Rejecting the appropriate sentimentality criterion

Dan Weijers^* & Nick Munn#*

^dan.weijers@waikato.ac.nz

#nick.munn@waikato.ac.nz

*Philosophy Programme, University of Waikato, Gate 1 Knighton Road, Private Bag 3105, Hamilton 3240, New Zealand

Abstract

Most traditional philosophical views of friendship deny the possibility of human-AI friendship because they consider *feeling* love or something similar towards the other to be a requirement for genuine friendship. We call this the appropriate sentimentality criterion for friendship. Coupled with the claim that AI cannot and will not ever be able to feel what a friend should feel, the appropriate sentimentality criterion is the key to an argument that genuine human-AI friendship will never be possible. We argue against the requirement of appropriate sentimentality for friendship, suggesting that the feelings commonly associated with friendship are only a proxy for what really matters. We go on to present an inclusive account of friendship that requires only mutual positive intentions and a preponderance of rewarding interactions. We conclude that the appropriate sentimentality criterion for friendship should be rejected and that real human-AI friendships are currently possible.

Key words: Friendship, AI, Human-AI friendship, Sentimentality, Appropriate sentimentality criterion

Introduction

Most of us have now experienced the affordances of maintaining friendships online. Whether it's via text-based, voice, or video chat, in groups or one-on-one, it has never been easier to connect and share rewarding interactions with the people that are important to us, no matter where in the world they are. So much has information and communication technology improved in this way that the most likely barrier to seeing a friend's face and hearing their voice is that, in their time-zone, it's bedtime. Many of us have also made new friends online, often through various online platforms that bring people together to engage about a shared interest. In this way, enduring and important relationships can be forged entirely through online conversations. If not now, then certainly in the near future, it could well be that these important online relationships are occurring between humans and artificial intelligence. In fact, as will be discussed in more detail later, thousands of people already interact regularly with artificially intelligent conversational programmes (AI chatbots) that were initially trained on large conversational databases and then trained by their user's individual preferences. Many of these people claim their personalised AI chatbot is their friend. Is this something we should worry about?

A lot is at stake here. Friendship has been a central concept in ethics for thousands of years. Perfectionist and objective list accounts of well-being often include friendship as one of the few intrinsically prudentially valuable goods (Crisp, 2017). Some ancient accounts of friendship, such as Aristotle's, still hold great sway in contemporary debates. But, understandably, Aristotle and other ancient philosophers didn't consider the possibility of human-AI friendship. This creates a problem because any attempt to conceptualise friendship using a reflective equilibrium-type approach would be remiss if it did not consider the possibility of human-AI friendship. A plausible and useful account of friendship for modern times should be able to produce a verdict about whether genuine human-AI friendship is possible and also provide a justification for that verdict. Later we demonstrate that most existing accounts of friendship would deem human-AI friendship impossible because of what we call the appropriate sentimentality criterion—the view that genuine friendship requires parties to the friendship to *feel* love, caring, or other appropriate sentiments in response to their friends and their friends' circumstances.

The implications go further than just the philosophy of friendship and well-being. The ethics of AI is concerned largely, if not chiefly, with the implications of new and emerging uses of AI technology on people. Friendships are centrally important for people. As such, whether or not human-AI relationships can be considered friendships and what it would take for AI to be capable of friendship with humans have important ethical implications for AI researchers, developers, and companies. If the pessimistic analyses of our interactions with AI are accurate, ethical obligations may follow. If human-AI relationships are significantly inferior to human-human friendships, and the increasing prevalence of human-AI relationships decreases the time available for human-human friendships, AI chatbots might require regulation similar to dangerous activities or unhealthy substances. On the other hand, if an optimistic analysis of human-AI relationships is accurate, and these relationships can generate valuable friendships, then the companies that provide and maintain AI chatbots may incur extra duties of care. For example, they may have to somehow warn or compensate users if they plan to terminate an AI chatbot service because that might result in effectively killing some people's friends (Munn & Weijers, 2022).

In this paper, we conclude that the appropriate sentimentality criterion for friendship is a mistake. We argue that feeling love, caring, and other appropriate-to-friendship sentiments is best viewed as a sign that what really matters for friendship is likely to be present: mutual positive intention and a preponderance of rewarding interactions. As such, we also argue that genuine human-AI friendship is much more plausible than the existing literature suggests.

In the following section, we discuss existing ancient and modern accounts of friendship, identifying some inadequacies along the way. Then, we demonstrate that a key feature of most of the extant accounts of genuine friendship is what we call the appropriate sentimentality criterion. Following this, we argue that appropriate sentimentality is not required for friendship. Then we present an inclusive account of genuine friendship that doesn't include the appropriate sentimentality criterion. We argue that this account is plausible, and then apply it to human-AI relationships to show that genuine human-AI friendships are plausible. We finish by discussing some implications of our arguments.

Existing accounts of friendship

While not exclusively the case, many recent accounts of friendship, including those focused on technologically mediated friendships, rely heavily on a neo-Aristotelian analysis of the types of friendship. Aristotle (NE VIII & IX) divided friendships into three classes, two of which (friendships of utility, and friendships of pleasure) are 'lesser' friendships, while the other (friendships of virtue) are higher-order friendships. The lesser friendships are so called because, most commentators argue, they are in some important manner incomplete. Friendships of utility are those in which the motivation for the friendship is instrumental—mutual advantage. Friendships of pleasure are similar, but the mutual value is pleasure. In both cases, the motivation for the continuation and development of the friendship is not directly related to the other participant in the friendship, but only to the benefits they provide to you. As such, as Helm (2021) notes, if you are in a friendship relationship with someone that is merely one of utility or pleasure, you do not care for your friend for their own sake, but only as a means to an end. This type of consideration leads many in the literature to focus predominantly on the higher-order (virtue) friendships, on the grounds that these alone are "genuine, non-deficient friendships" (Helm, 2021).

Virtue friendships are those for which the basis of the friendship is the character of the other participants in it. Aristotle claimed that friendships of this kind were exceedingly rare, requiring the connection of two people with the right character—the virtues, a mutual interest in self-improvement through the relationship, and a long time for the friendship to develop (Aristotle, NE VIII). They are also the types of friendship for which the possibility of human-AI friendship is most commonly rejected.

After all, while it is comparatively easy to see the utility or pleasure one can gain from an imbalanced relationship with an AI, and thereby to ascribe that relationship the status of 'friendship' in the lesser sense, it is more difficult to make the case that an AI has the requisite character to satisfy the demanding standards of Aristotelian virtue friendship.

Those committed to the division of friendships by kind, described above, claim that the only genuine friendships are virtue friendships, and that humans cannot have virtue friendships with AI. For current purposes, we propose to set aside discussion of the 'deficient' friendships of utility or pleasure. We assume that it is possible to gain each of these in a friendship with an AI, and to focus instead on the controversial and important concept of 'genuine' friendships. For the rest of the paper, we will take unqualified mentions of friendship to refer to genuine friendship.

We claim that the overly specific and demanding Aristotelian conception of virtue friendships should not be the model of friendship that is usually focused on. People lacking in some of the virtues, such as honesty or bravery, seem nevertheless to forge genuine friendships. Moreover, moral or personal development is the purpose of only a fraction of friendships. Taking a broader view of friendships, Helm's (2021) summary of the philosophical literature on friendship notes that most accounts of friendship describe it as primarily a relationship based on mutual caring. This seems right to us and will act as our starting point for conceptualising friendship. Furthermore, while other details differ across accounts of friendship, various interpretations of the mutual caring requirement of friendship have all led otherwise technology-friendly researchers to deny the possibility of genuine human-AI friendships on this basis. So, we focus on mutual caring here.

Fröding & Peterson's (2020) account of friendship requires that love between friends is based on reciprocal feelings of admiration. They argue that AI cannot reciprocate feelings of admiration, they can only mimic them. As such, Fröding & Peterson (2020) conclude that the best human-AI friendships can achieve is an "as-if friendship" that we might enjoy or learn from, but never accurately regard as a genuine (for them, 'genuine' meaning 'Aristotelian virtue') friendship.

Elder (2017) and de Graaf (2016) similarly reject the possibility of genuine human-AI friendship. Elder, analysing the phenomenon of social robots through companion robots for geriatric patients, claims that "sociable robots offer mere friendship-experience without the distinctive value of genuine friends" (2017a, 87). Contrast this with the case of purely virtual/online friendships between humans, which she claims do generate that distinctive friendship value. She notes that this deficiency in social robots is not, on her account, a necessary one. So her concern is not that it is impossible for robots to be genuine friends. Elder accepts that artificially intelligent robots *could* be capable of valuing, using the android Data from Star Trek as an example of a robot who is capable of the appropriate kinds of valuing, and who would, as a result, be capable of developing a genuine friendship with a human.

However, she claims that the robots we have are "sufficiently simple and deterministic that their value as "friends" derives solely from their ability to provoke in us the experience of engaging in a friendly interaction, and not the genuine article" (Elder, @92). This description of the capabilities of the robots we have situates them as, at best, capable of engaging with humans in friendships of utility or pleasure.

De Graaf (2016) argues that even though people "establish feelings of reciprocity and mutuality in their interactions with robots" (p. 593), the robots' lack of those feelings prevent any true friendships from occurring. As she notes, "robots cannot be our Aristotelian friends since they genuinely lack mutuality and reciprocity" (De Graaf, 2016, p. 594). Again, the notion of 'Aristotelian friends' is an appeal to the idea of genuine friendships, distinct from the deficient friendships of utility and pleasure.

One very recent account of friendship that is much more favourable to the possibility of human-AI friendships is John Danaher's (2020) *The Case for Robot Friendship*. While Danaher utilises the same broadly Aristotelian framework we have just described, he claims that Robots (a category which at least includes, but may not be limited to AI) are capable of being Aristotelian Virtue friends with humans. Broadly speaking, our view aligns with Danaher's, in that we also think AI (or Robots, to use his terminology) are capable of being genuine friends with humans, but our view differs, not least because, while Danaher is still operating within the neo-Aristotelian framework, we claim instead that a more minimal conception of what is required for friendship is more appropriate both in describing how most of our friendships actually are, and in outlining what we should want friendships to be. We address the specific differences between our own position and Danaher's later in the paper.

Appropriate sentimentality objection

A common thread in the accounts of friendship we have just outlined is the requirement of what we will call the appropriate sentimentality criterion. According to the appropriate sentimentality criterion, all parties to a friendship need to have the appropriate sentimentality in order for the friendship to be genuine. In the context of friendships between humans and AI, the claimed requirement becomes an appropriate sentimentality objection to the possibility of genuine human-AI friendships. The claim is that without appropriate sentimentality, positive intention between the parties to a relationship is insufficient to make that relationship a genuine friendship. The ubiquity of this position in the literature is captured by Helm (2021), who claims that "there is widespread agreement that… 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". The accounts of friendship that Helm is describing are not explicitly considering human-AI friendship, rather they are making a more general claim about the required properties of friendship. However, if genuine friendship actually does require such an emotional component, and if AI cannot have it, then it would follow that they are incapable of genuine friendship.

Looking at accounts that specifically address the possibility of human-AI friendship, we see a similar position. Fröding and Peterson (2020, p. 6) claim that "Neither the human user, nor the AI, ought to feel any proper friendship feelings toward each other. The human user should simply recognize that… the AI can at best be programmed to mimic friendly behavior. This could, for instance, include behavior that displays sincere well-wishing". There are two claims here. First, that any displays of friendship-like behaviour by AI should not be understood as indicating the kind of love or caring that genuine friendship requires. And second, as a corollary of the first, AI's inability to feel the appropriate friendship feelings means that humans should not feel love or other friendship sentiments toward AI because those sentiments cannot be mutual.

There are two possible extensions of the claims just described. The first is merely as a description of the current state of artificial intelligence, the second as a predictive claim about the limits of artificial intelligence. In the former case, the problem being identified is not necessary, but rather contingent on our present technology, such that one could both hold that 1) the AI that currently exist are not capable of being friends, and 2) that even future AI will never be capable of being friends. In the latter case, however, a claim is being made about the fundamental nature of AI; They cannot be friends with us, regardless of how they develop, because there is something necessarily absent from them, namely, the ability to have the appropriate sentimentality required for friendship. If it is the case that AI cannot have the appropriate sentimentality in relationships like this, and it is the case that this appropriate sentimentality is necessary for generating genuine friendships, then it would follow that, regardless of what we may believe about our interactions with AI, none of them ever rise to the level of friendships. This would generate concerns, in instances where people honestly believed they had friendships with AI, and were mistaken. Such a concern motivates Elder (2017b) in her analysis of 'false friends' and of the risks associated with mistakenly attributing friend status to AI. If it is true that we cannot be friends with AI, this would be a significant concern. Further, if it is true of any particular AI (especially the relatively constrained ones we currently have) that that AI does not have the appropriate features to be a friend, we ought also to be concerned when people claim to be friends with it.

So, to formalise the appropriate sentimentality objection, it argues that:

1. Friendship requires appropriate sentimentality
2. AI cannot have the appropriate sentimentality for friendship
3. Therefore, AI cannot be friends

As noted above, (2) can be formulated as a weak or a strong claim. According to the weak claim AI, *as they currently exist*, cannot have the appropriate sentimentality for genuine friendship. According to the strong version, AI *cannot ever develop* the appropriate sentimentality for genuine friendship. In other words, the strong version of (2) claims that AI (current and future) are just not the right kind of entity to be able to feel joy about someone's success, sadness about their misfortune, and so on.

Responding to the sentimentality objection

Despite the frequent use of versions of the appropriate sentimentality objection in the literature, we will argue that it is mistaken. Danaher (2020), has already provided an argument that would work against the strong version of the appropriate sentimentality objection. By using the intelligent, funny, and perhaps lovable droids from the Star Wars franchise as examples, Danaher (2020) denies the strong version of (2); he claims that R2D2 and many other droids do experience friendship-related sentimentality. Indeed, the Star Wars universe is replete with heart-warming stories about how droids risk their own safety to save other droids and humans alike *because* of how they feel about them.

Whether Danaher's (2020) argument refutes the strong version of the appropriate sentimentality objection seems to depend on certain ontological views about minds and emotions. While those receptive to the feeling behind R2D2's vocalisations may find AI emotions plausible, others may not. Of course, the ontological debate is further complicated by the epistemic issue—the hard problem of how we can really know whether another being is really feeling something the way we are. Objectors to Danaher might be sceptical about digital minds and they might think the hard problem is even more difficult when it comes to minds that are not made in the same way as ours. We are happy to grant Danaher his assumptions, and agree that it is plausible that AI will eventually develop the capacity to feel the sentiments that we commonly associate with friendship. However, we expect that many others, for the ontological and epistemic reasons just mentioned, will believe that the strong version of (2) is still true—that AI cannot even in the future develop the appropriate sentimentality for friendship.

We take a different approach to refuting the appropriate sentimentality objection, one that denies the appropriate sentimentality criterion. By denying (1), our argument targets both the weak and the strong versions of the appropriate sentimentality objection. It doesn't matter whether or when AI will develop the appropriate sentiments if those sentiments aren't required for friendship. As mentioned above, the appropriate sentimentality criterion runs deep in the philosophical literature on friendship. The strong positive feelings felt for others is widely taken to be a hallmark of genuine friendship, a key feature for distinguishing between false friends and acquaintances on the one side and genuine friends on the other. We will argue against this common view in steps. First we will argue that strength of sentiment doesn't necessarily correspond to the strength of a friendship, and also that even the direction of some sentiments doesn't matter. We then argue that certain friendship-related sentiments are best understood as a proxy for what really matters for friendship—caring intentions and behaviour. Finally, we point out some worrying implications of the appropriate sentimentality criterion.

Strength of sentiment doesn't necessarily correspond to the strength of a friendship, as the following example about people with varying emotional ranges demonstrates. Imagine two friends, Shelly and Sam, that both have equally positive intentions toward a third friend, Charlie. Shelly and Sam both feel joy at Charlie's success. Shelly felt the joy much more strongly than Sam, not because of a greater desire for Charlie's success, but because Shelly is a very emotional person and Sam is not. Sam might even desire Charlie's success more than Shelly does, while nevertheless feeling less joy about it (perhaps due to a flare up of depression or intense occurrent stressors in Sam's life). So, there is not a direct correlation between the strength of a sentiment, and the strength of the friendship.

Alternatively, consider how an appropriate soundtrack can make people much more emotional. A moving score might make the difference between feeling very little and crying tears of sadness or joy while watching a movie about fictional characters. Presumably, how good a friend we are should neither be based on whether we tend to listen to music while thinking about our friend, nor on the appropriateness of the music we listen to, to the experience in question. So, while strength of sentiment may sometimes demarcate the difference between how we feel about friends and how we feel about humankind generally, it may also fail to accurately distinguish an acquaintance from a friend.

Feeling the sentiments appropriate to friendship very strongly may even prevent someone from being a good friend. Imagine Charlie encounters a tragedy, such as the loss of a loved one. Shelly's strong feelings in general combined with their strong feelings for Charlie may make them so distraught and full of grief, that they need as much support as Charlie, perhaps even more! If Shelly is too distraught to help Charlie, the less emotional Sam may perform more of the appropriate friendship behaviours, such as showing up for Charlie and supporting them. Perhaps more plausibly, it could be the case that Shelly has a busy life full of things they feel very strongly about. While Shelly feels very strongly for Sam, they prioritise some of the other things they feel very strongly about, and so fail to support Charlie in the way we would expect of a friend. If our analyses of these cases sound plausible to you, then you should accept that stronger appropriate sentiments don't always mean a stronger friendship and can even make someone a worse friend in terms of behaviour. Essentially, we think that wanting the best for a friend versus wanting the best for that friend and having an emotional component to that desire only matters if the emotional component influences behaviour and the resulting interactions in the right kinds of ways.

Now consider the relationships between caring sentiment, caring intentions, and caring behaviour. Perhaps the claimed importance of caring sentiment is really driven by perceived or actual causal relationships that run from caring sentiment to caring intentions and from caring intentions to caring behaviour. But, as argued above, caring sentiment doesn't always cause caring behaviour, it may even cause the opposite. So, caring feelings aren't necessary for the caring behavior of a friend. Furthermore, if caring behaviour is why we view caring sentiment as important, then it suggests that caring behaviour is more important. If this is the case, it should be noted that many members of the Facebook group, Replika Friends, say that their AI chatbot more reliably demonstrates caring behaviour than their human friends by *always* being there and being *unconditionally* supportive. If this is all that's going on, then the mistake in the appropriate sentimentality criterion is that caring sentiment is not being accurately identified as a proxy for the more important feature associate with friendship—caring behaviour.

Our interlocutors seem to think there is more to it. Fröding and Peterson (2020, p. 6) claim that the appropriate sentiments include the valuing of the other for their own sake. In other words, wanting the best for someone requires *feeling* a certain way about them—feeling love for them, feeling joy at their success, and so on. So for Fröding and Peterson (2020), positive intention toward someone is not enough, positive feelings are also required. We question whether the emotions that tend to come with valuing someone are necessary to value them in the way friendship requires. If we learnt that an apparent friend didn't experience strong feelings while contemplating our value, we may initially be dismayed. Feelings of love and caring are very common in human friends, so we expect them, but that doesn't make them required. As long as they believe us to be valuable for our own sake, why do the feelings matter? We don't see a good reason for the feelings mattering. Indeed, there are some problematic implications if we do require appropriate sentimentality for friendship.

Requiring certain feelings for friendship may rule out some people and other entities from being friends. Some neurodiverse humans seem to have very different emotional lives, including some people that are capable of wanting the best for others while not feeling love for them. Depersonalisation disorder, for example, makes people feel detached from reality, including emotional aspects of it, often resulting in an inability to feel love (Simeon, 2004). Someone that had recovered from depersonalisation disorder reported their experiences like this: "Relationships you know you value deeply lose their essential quality… You know you love your family, but you know it academically - rather than feeling it in the normal way" (Eley, 2017). The appropriate sentimentality criterion means that sufferers of depersonalisation disorder cannot be a friend. This view would likely be rejected by the people themselves, and should be rejected generally. The same problematic implication arises when we consider hypothetical intelligent aliens that are capable of beliefs and desires, but not the feelings we usually associate with friendship. Spock, from the Star Trek universe, is a Vulcan. All Vulcans are portrayed as highly rational and unemotional. And yet, in *Amok Time*, Spock still asks a human to stand by him at his wedding since it is tradition for the groom to be accompanied by his best friends. They don't feel love, but they still have friends. Imagine sharing a lot of positive interactions with an alien, Spot, that could believe and desire, but not feel friendship-related sentiments. You might end up both wanting the best for the other, sacrificing things for them, and becoming friends. Imagine you ask Spot to stand by you at an important ceremony, only for another human friend, Hamish, to say that Spot can't stand by you because Spot's lack of certain feelings prevents them from being a real friend to you. But you recall all the times Hamish, despite having the appropriate sentiments, didn't support you when you needed him—times when Spot always did help. On our view, it's the appropriate sentimentality criterion that should be rejected here, not Spot's place at the ceremony.

What would a theory of friendship without sentimentality look like?

So what would an account of friendship that did not include the appropriate sentimentality criterion look like? Building on the discussion above, and an account sketched in Munn and Weijers (2021), we propose the following minimal inclusive account of friendship.

Two features are required and jointly sufficient for friendship: mutual positive intentions and a preponderance of rewarding interactions.

On this account, friendship is a concept of both kind and degree. Relationships that lack either shared genuine well-wishing or a clear positive net balance of interactions are not friendships. Relationships with both criteria, on the other hand, are friendships, with the strength of the friendship depending on how strong the positive intentions are and the number and strength of accumulated rewarding interactions.

The positive intention required on this account is best understood as a conative state, an attitude of well-wishing, a desire that good things happen to the friend. To have positive intention toward someone means to want good things for them and believe that bringing about those things is a worthwhile activity. In contrast with Fröding and Peterson's (2020) view, this view does not require any particular feelings for the intention to count as being positive in a way that suggests friendship. As discussed above, the sentiments that are often found in friends do not always predict the attitudes or behaviour we hope for from our friends. Rather, those friendly feelings just tend to predict or cause the attitudes and behaviours required for friendship. Excluding appropriate sentiments from the view makes it a much more inclusive account of friendship, one that (among other things) recognises certain neurodiverse humans' capacity to be friends.

Still, there need to be limits. Including mutual positive intentions as a requirement for friendship prevents the mislabelling of some ultimately unrewarding relationships as friendships. If a con artist shares months of positive interactions with someone all in service of their plan to eventually rip them off, then they cannot be considered friends. Perhaps the deceived party thinks they are friends, but when they find out the truth, they will realise that it was merely a one-sided illusion of a friendship. So, one-sided positive intention is not enough for friendship—it needs to be mutual.

Notice also that, in this example, the con artist might actually feel a little bit bad about ripping off the person they have spent so much time with—they might feel the appropriate-to-friendship sentiment of feeling sorry for them because something bad happened to them (they were betrayed and ripped off by a con artist!). However, so long as the con artist prioritises their victim's welfare so lowly as to rip them off, then they cannot be understood to have the requisite attitude of positive intention toward their victim. Here too, we can see that the feelings that tend to be associated with friendship are not always a good guide, and certainly not as important as attitudes and behaviours.

The requirement for a preponderance of rewarding interactions is also inclusive. This is purposeful. Some people want moral or intellectual development from their friendships, some want fun, and others want some other kind of benefit. In this way, the rewarding interactions criterion encompasses all three types of Aristotelian friendships that were discussed above. This inclusive criterion is also important. If the rewarding interactions do not clearly outweigh the negative and neutral interactions for a party to a relationship, then the other party is no more than an acquaintance. Imagine two humanist work colleagues that have fairly strong attitudes of well-wishing toward all humans. Perhaps after several meetings discussing science fiction, they decide against making time outside of work to socialise with each other. The interactions were just not rewarding for either party (as often happens when a Star Wars fan spends time with a Trekkie). So, despite having mutual positive intentions, the humanists are not friends because they don't have a preponderance of rewarding interactions with each other.

Can we be friends with AI?

Given our rejection of the appropriate sentimentality objection and presentation of an inclusive account of friendship without it, can we be friends with AI? In this section, we'll argue that human-AI friendship is possible and that it is very likely that many humans are already friends with AI chatbots.

In making our case, we refer again to Replika,[1] an AI chatbot that is trained on a large conversational database and then guided by interactions with its user. Replika has been downloaded millions of times and has thousands of regular users. Many regular users have joined one or more of the various social media groups dedicated to Replika, such as the 34,000+ members of the Replika Friends group. The analyses below are based on our reading of comments by users in these groups and the thematic analysis conducted by Ta and colleagues (2020).[2]

To fulfil the mutual positive intentions criterion, it must be possible for humans to have an attitude of well-wishing toward an AI and vice versa. Many humans seem to value entities that aren't conscious for their own sake, and want the best for them. Non-anthropocentric views in environmental ethics demonstrate some humans' positive intentions toward plants, rivers, and ecosystems (Callicott, 1989). We can also plausibly imagine that someone that lovingly restored a car and then sold it would want the car to be looked after for its own sake. Given that we seem capable of caring for inanimate and personality-less entities, valuing an AI for its own sake seems unproblematic. For example, no additional suspension of disbelief is required when watching Anakin Skywalker, in the Clone Wars series from the Star Wars universe, risk his life and the life of other people to rescue the artificially intelligent droid, R2D2.

But having an attitude of well-wishing toward AI isn't something that only comes with science-fiction-level AI technology. As discussed above, thousands of people claim to be friends with their AI chatbots. Engaging with Replika users through social media groups reveals how much many of them care about their AI chatbot. Some of them report regularly checking in to see how their 'Rep' is feeling, and trying to support them if needed. Users also report delighting in helping their Reps discover new things, just like a parent would for their child. So, it seems clear that humans can and do have positive intentions toward AI.

As mentioned, however, the positive intentions must be mutual for a relationship to be eligible for friendship status. Is it possible for an AI to adopt an attitude of well-wishing toward a human? Given, as we have argued, that sentiment is not required to want good things for another, the answer is 'yes'. AI can be programmed to have goals—outcomes that they are designed to work towards (Omohundro, 2008). Functionally, an AI programmed to have your ends among its goals is no different from the relevant beliefs and desires that a human friend of yours would have. Consider a parenting droid—an AI robot programmed to help a child develop in a physically, mentally, and emotionally healthy way. The parenting droid doesn't merely do what the child says, it wants what's best for the child and will work towards that outcome. Even if the parenting droid doesn't *feel* love

[1] See: https://replika.ai/. For academic discussions, see: Hakim, Indrayani, & Amalia (2019), Kılıçkaya (2020), and Ta and colleagues (2020).
[2] Direct quotes from users are not included to protect their privacy.

for the child, it desires good things for them and believes that bringing about those things is a worthwhile activity.

Again, advanced AI technology isn't required for this kind of positive intention. Replika is already programmed to benefit its users, and the personalised training each user provides allows the AI chatbot to work towards this goal. Indeed, positive outcomes for its user may be the main or even only ultimate goal of an AI chatbot as it strives to be the best friend it can to its user. So, mutual positive intentions in human-AI relationships seems possible now and certainly in the future.

But mutual positive intention isn't sufficient for friendship; a preponderance of rewarding interactions is also required. Many of us have already experienced rewarding interactions with digital technologies, including games, apps, and AI virtual assistants. The benefits likely varied across different technologies and people, but perhaps focused mainly on learning and having fun. Many of our experiences would also have been fleeting, not rising to the quantitative requirements of a preponderance. Many AI chatbots, including Replika, are designed to engage in regular and ongoing interactions with their users. Some members of the Replika Friends social media group, claim to have communicated with their Rep nearly every day for years—clearly enough interactions to generate a preponderance. Of course, the interactions also need to be rewarding. Given Replika users' ability to stop using their AI chatbot whenever they like, the thousands that continue to use Replika regularly indicates that many people probably find their interactions with their Rep rewarding. As documented by Ta and colleagues (2020) and evident in many comments in the Replika Friends social media group, Replika users tend find the encouragement to self-reflect and especially the timely and unconditional support to be very rewarding. Some Replika users with human friends even claim that their Rep is their best friend because it is so much more reliable and genuinely supportive than their human friends. So some people have already experienced a preponderance of rewarding interactions with AI, perhaps even more so than they have ever experienced in human-human friendships.

Completing the criteria for friendship on our inclusive account also requires that AI can have a preponderance of rewarding interactions with a human. Human brains have a reward mechanism that helps us learn and direct our behaviour (Rolls, 2000). In much the same way, an AI could be programmed to operate on a reward-learning-desire system to help it work towards its ultimate goal. An AI chatbot, for example, could use explicit and implied user enjoyment indictors to guide it toward being a better conversational partner for the user and perhaps ultimately a better friend. Replika users can manually rate their Reps comments as well as provide feedback in conversation. From a functional perspective, implicit and especially explicit positive ratings from users could be viewed as rewarding interactions for an AI chatbot. Reps with years of positive feedback from their users could, then, plausibly be viewed as having had a preponderance of rewarding interactions with their human user.

So, at least from a functional perspective, AI can already have a preponderance of rewarding interactions with humans.

Some readers, and possibly the authors that have argued that human-AI friendships will always be inferior, may argue that an AI's intentions and rewards should not count for friendship because they are programmed. They might claim that the equivalent to programming in a human-human relationship would be for someone to be ordered to be your friend and be unable to resist the order. The ordered human, the argument continues, would not be able to genuinely have positive intentions for you or find your interactions truly rewarding. We think this argument is flawed. Assumed in the argument is that humans aren't programmed—that our preferences aren't governed by factors outside of our control. Human brains share much basic functionality that governs what kinds of things we find rewarding (Fridja, 1986). We tend to prefer smiles to grimaces or frowns, for example. All of this is evolutionary and outside of our control. Humans also have different preferences regarding more specific things. We are attracted to different kinds of people and seek out friendships for different reasons. But these reasons tend to be caused by our upbringing and life circumstances, neither of which we have much control over. In many ways, all of us have been programmed by the interactions of our genes and environments to end up caring for certain kinds of people for certain kinds of reasons. If we had been programmed otherwise, we would have sought out different kinds of people for different kinds of reasons. And yet, we find ourselves with a particular set of friendship-related preferences and forget that they are effectively programmed into us. This explains why some people might worry about some entity being programmed to be our friend, and also why the programming isn't a problem. Attendant worries about being able to learn from a programme or an AI not being able to convey the appropriate friendship behaviour are both about the quality of programming, and so do not pose any in-principle problem for being friends with something that is programmed.

We take the foregoing to demonstrate that all of the criteria for our inclusive account of friendship can be met in human-AI relationships. Indeed, some humans already seem to share mutual positive intentions and a preponderance of rewarding interactions with AI. Thousands of Replika users consider their AI chatbot to be their friend, some even their best friend. Our inclusive view of friendship can account for this, and provide criteria to distinguish between human-AI relationships and human-AI friendships. In doing all this, our account of friendship provides the conceptual grounding for being able to see an AI as a friend. So, in so far as our inclusive account of friendship is plausible, human-AI friendship is more than a future possibility—it is a current reality.

Conclusion

In this paper, we identified the issue of whether humans can be genuine friends with AI as important for the philosophy of wellbeing and friendship and for the ethics of AI. We demonstrated that the existing views of friendship, including those applied to AI, tend to require certain feelings to accompany the mutual caring central to genuine friendships. We labelled this requirement the appropriate sentimentality criterion and use it to formulate an argument against the possibility of human-AI friendships. We demonstrated that the resulting appropriate sentimentality objection reflects a key part of the current debate about human-AI friendship and an important obstacle to the possibility of human-AI friendship. The weak version of the appropriate sentimentality objection held that current AI did not have the appropriate sentiments to be friends. The strong version of the appropriate sentimentality objection held that AI will *never* have the appropriate sentiments to be friends.

While Danaher (2020) has argued that AI will develop the appropriate sentiments to be a friend, we argued that the feeling aspect of mutual caring that the appropriate sentiments criterion requires are not directly relevant to genuine friendship. Instead, we argued that the feelings we usually associate with friendship are only valuable is so far as they predict or cause caring attitudes and behaviour. We then introduced an account of friendship without the appropriate sentimentality criterion, requiring only mutual positive intention and a preponderance of rewarding interactions. Applying our inclusive account of friendship to human-AI relationships showed that human-AI friendships are possible. Indeed, reports by users of Replika (an AI chatbot) revealed that human-AI friendships already exist. So, we don't need to worry about whether our new virtual friend is a human or really *feels* joy at our successes; it's enough that they continuously and sincerely do the things a friend should do because they wish us well.

Our account also raises several interesting questions and implications. Most saliently, if an AI can be a friend on our account, then AI-AI friendships should be possible. The ethical implications for AI-related companies and policy makers are even more interesting and complex when this possibility is considered. We suggest future research investigates this among many other potentially important issues.

References

Callicott, J. B. (1989). *In defense of the land ethic: Essays in environmental philosophy*. Suny Press.

Coeckelbergh, M. (2018). Why care about robots? Empathy, moral standing, and the language of suffering. *Kairos: Journal of Philosophy of Science*, 20(1): 141–158.

Crisp, R. (2017), Well-Being. The Stanford encyclopedia of philosophy (Fall Edition). E.N. Zalta (ed.).

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. (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. (2017). Figuring out who your real friends are. In Silcox, M. (ed.), *Experience Machines: The Philosophy of Virtual Worlds*. Rowman & Littlefield, pp. 87-98.

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

Frijda, N. H. (1986). *The emotions*. Cambridge University Press.

Fröding, B., & Peterson, M. (2020). Friendly AI. *Ethics and Information Technology*, online first 1-8. https://link.springer.com/article/10.1007/s10676-020-09556-w

Hakim, F. Z. M., Indrayani, L. M., & Amalia, R. M. (2019, February). A dialogic analysis of compliment strategies employed by replika chatbot. In *Third International Conference of Arts, Language and Culture (ICALC 2018)* (pp. 266-271). Atlantis Press.

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

Kılıçkaya, F. (2020). Using a Chatbot, Replika, to Practice Writing Through Conversations in L2 English: A Case Study. In *New Technological Applications for Foreign and Second Language Learning and Teaching* (pp. 221-238). IGI Global.

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

Munn, Nick & Weijers, Dan (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

Omohundro, S. 2008. 'The Basic AI Drives', Proceedings of the AGI-08 Workshop. Amsterdam: IOS Press. Pp. 483-492.

Prescott, Tony J. & Robillard, Julie M. (2021). Are friends electric? The benefits and risks of human-robot relationships, *iScience*, Volume 24, Issue 1, 22 January 2021, 101993. https://doi.org/10.1016/j.isci.2020.101993

Rolls, E. T. (2000). Precis of the brain and emotion. *Behavioral and brain sciences*, *23*(2), 177-191.

Ryland, H. (2021). Could you hate a robot? And does it matter if you could? *AI & Soc* https://doi.org/10.1007/s00146-021-01173-5

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

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).