Friendly AI

Abstract

In this paper we discuss what we believe to be one of the most important features of near-future AIs, namely their capacity to behave in a friendly manner to humans. Our analysis of what it means for an AI to behave in a friendly manner does not presuppose that proper friendships between humans and AI systems could exist. That would require reciprocity, which is beyond the reach of near-future AI systems. Rather, we defend the claim that social AIs should be programmed to behave in a manner that *mimics* a sufficient number of aspects of proper friendship. We call this "as-if friendship". The main reason for why we believe that as-if friendship is an improvement on the current, highly submissive behavior displayed by AIs is the negative effects the latter can have on humans. We defend this view partly on virtue ethical grounds and we argue that the virtue-based approach to AI ethics outlined in this paper, which we call "virtue alignment", is an improvement on the traditional value alignment approach.

Key words: AI, friend, friendly, value alignment, virtue ethics

1. Introduction

In December 2019 the second generation of the Crew Interactive Mobile Companion robot, known as CIMON-2, arrived at the International Space Station. Designed by the German branch of Airbus, it uses artificial intelligence powered by IBM's Watson technology. One of CIMON-2's tasks is to serve as a conversational companion for lonely astronauts. According to Matthias Biniok, Lead Watson Architect at IBM, "studies show that demanding tasks are less stressful if they're done in cooperation with a colleague".[1] CIMON-2 is programmed to behave like an artificial colleague by answering questions and engaging in conversation. This enables astronauts to perform better and thereby make space missions less stressful and more successful.

CIMON-2 is a significant improvement of its predecessor, CIMON, which was tested at the International Space Station in 2018. One of the problems with CIMON was that it was perceived as mean and unfriendly by crew members:

[1] Ashley Strickland, C. (2020). "Robot arrives on the space station to keep astronauts company." Retrieved 29 January 2020, from https://www.cnn.com/2019/12/06/world/cimon-space-station-scn-trnd/index.html

> In an early demonstration in 2018, it was CIMON — not Gerst [a German astronaut] — that needed a morale boost. After Gerst asked CIMON to play his favorite song, the 11-pound bot refused to let the music cease, defying Gerst's commands. And, rather than acknowledging it had jumped rank, CIMON accused Gerst of being mean and finished with a guilt-trip flourish by asking Gerst, "Don't you like it here with me?[2]

Sophisticated AI technologies currently reserved for space missions are likely to become more widely available in the future. Some of these AIs will be designed to fulfil social functions in our daily lives and in the healthcare sector.[3] Already now there is demand for social companions (robot dogs and robot cats) for senior citizens residing in assisted living facilities. It is, arguably, time to discuss what type of behavior these AIs should be programmed to display.

In this paper we focus on what we believe to be one of the most important features of near-future AIs, namely their ability to (appear to) be *friendly* to humans, as opposed to mean, hostile or unfriendly in other ways.[4] Presumably, the claim that AIs designed to fulfill social functions ought to behave in a friendly manner is an intuitive claim that is hard to reject. The challenge that lies ahead of us is primarily to construct a plausible interpretation of what it *means* for an AI to behave in a friendly manner, not to show that friendliness is desirable.

Our analysis does not presuppose that proper friendships between humans and AI systems like CIMON-2 could exist. That would require reciprocity, which is beyond the reach of the Watson technology and other current, or near-future, social AI systems discussed here. We merely claim that social AIs should behave in a manner that *mimics* a sufficient number of aspects of proper friendship. We will call this "as-if friendship". Although AIs could perhaps mimic reciprocity ("as-if reciprocity"), we take proper friendship to require proper reciprocity, meaning that as-if reciprocity can at most generate as-if friendship.[5]

To start with, we note that being friendly to someone is not the same as being that person's friend. Consider, for example, the staff in your favorite coffee shop, bar, grocery store or gym. Presumably they are friendly to you, but that does not mean that they are your friends in a deeper sense.

---

[2] Ibid.
[3] When we use the term "AIs" or "AI systems" in this paper we include assistants, companions, facilitators etc. which may or may not be presented as robot technology.
[4] In this paper we limit our discussion to *non-antagonistic* AI systems. If it is ever morally permissible to use AI systems in warfare, which is a controversial claim we will not discuss here, then such military AIs should arguably be allowed to be hostile to the enemy.
[5] Note that we do not take all forms of as-if friendship to require as-if reciprocity. Some lesser forms of as-if friendship require no type of reciprocity at all; see section 3.

A way to think about the notion of human friendliness is the following: an individual is friendly in case she behaves as a proper friend *would* behave in a similar situation. Because not all friends behave in the same way, this criterion entails that people can be friendly in numerous ways, which we take to be plausible. However, the above suggestion needs to be qualified as there are situations where friendliness does not require the same type of behavior as proper friendship. Consider, for example, how a proper friend will sometimes be required to behave in a way that is not friendly in the everyday sense. Imagine that you come to know that your friend has some very negative habits (e.g. over-consumption of alcohol). In this case true friendship could require that you intervene, and this involves confrontation, truth-telling and accountability. However, this is not expected from someone who is merely friendly to you. In addition, we also note that a proper friend can, in other cases, be legitimately required to sacrifice herself to some extent; she sometimes has to inconvenience herself and put the good of the friend before her own good. Again, this is typically not required of someone who is merely friendly.

Yet another difference, which bears specifically on the discussion of friendly AI, is that an AI can, and sometimes should, be friendly to you in situations that you would never encounter together with a human friend. Consider, for instance, AIs that help users keep track of and analyze millions of business transactions. A friend of a human user would never be asked to perform such a monotone task. Therefore, this task cannot be performed by a human in a friendly manner. But as monotonicity would not bother AIs, we can imagine AIs that perform such tasks in friendly as well as unfriendly ways.

These and other examples indicate that it is far from trivial to base a theory of friendly AI on an analysis of how a proper friend of the user would perform this task. However, as we will explain shortly, our suggestion for how AI systems should be programmed to behave can take care of the objections outlined here. Our proposal is that for a human person to qualify as friendly, she should mimic *a sufficient number* of aspects, but not all, of proper friendship. For example, behaving friendly could plausibly require sincere well-wishing, the intrinsic valuing of the other, helpfulness, and empathy. Proper friendship, on the contrary, requires more, for example a commitment to honesty which in examples of self-destructive behavior would require the proper friend to confront the other. Hence, a friendly AI system should, in the case of the addict, not be required to confront the user (although a future AI capable of proper friendship would), and in the case concerning monotone work tasks, the friendly AI would not complain (unlike the first version of CIMON). Our strategy for analyzing AI friendliness can thus be summarized as follows: We use the notion of proper human friendship as a point of

departure; we will return to what we mean by proper human friendship later in the paper. We then define human friendliness as behavior that mimics sufficiently many aspects of proper friendship. Although AIs cannot be proper friends, we claim that AIs can be friendly if they mimic sufficiently many aspects of proper friendship, but as noted we maintain that the ways in which humans and AIs mimic friendship relations need not always be the same.

As we see it, the most troubling ethical challenge with designing friendly AIs is that we may end up treating the AIs as mere slaves rather than as-if friends. This would be undesirable because, as will be elaborated on in the following sections, this is likely to have a negative impact on the development of the human users' character virtues. For the sake of clarity, we will articulate and discuss three types of as-if relationships that can hold between AIs and humans, which we believe it is important to keep apart:

1. *Slave AIs:* These AIs are programmed to behave as-if they were slaves controlled by human masters.
2. *Utility AIs:* These AIs are programmed to behave as-if they were facilitators of the common good in human societies.
3. *Social AIs:* These AIs programmed to behave as-if they were friends of individual humans.

We reserve the terms "friendly AI" for AIs of the second and third types, that is, utility AIs and social AIs. Whether slave AIs could be friendly is debatable, but irrelevant in the present discussion as we claim that slave AIs are morally problematic and should be transformed into friendly utility AIs or friendly social AIs. We present our argument for this claim in Section 2 and then contrast the notion of friendly utility AIs and friendly social AIs with Asimov's well-known laws of robot ethics. We point out that if the notion of friendly AI is construed in terms of as-if friendship, then two of Asimov's three laws of robot ethics have to be rejected for what we take to be convincing reasons. In Section 3 we summarize some of the key aspects of Aristotle's theory of friendship, which is put to work in Section 4 for articulating our preferred notion of friendly AI in terms of as-if friendship. Finally, in Section 5, we conclude the paper by pointing out some differences between our approach and the traditional value alignment project.

2. Why treating AIs as slaves is morally wrong

Nearly all AIs developed so far, including CIMON and CIMON-2, are premised on the idea that the purpose of the AI is to improve, facilitate, or simplify human activities. This is, to a certain extent, unproblematic from a moral point of view. The near-future AIs we discuss in this paper have no moral standing, so it would make little sense to raise the Kantian objection that the AI is being used as mere means to an end. It is indeed true that AIs *are* being used as mere means to an end, but that is permissible because they have no moral agency.

However, we would like to highlight another, more problematic attitude toward the different versions of CIMON and near-future AIs, namely, our tendency to think of AIs as artificial slaves. Some of the best-known AIs we use in our daily lives have ordinary human names such as Siri or Alexa.[6] These AIs are frequently ascribed, explicitly or implicitly, a range of human-like properties. Notably, we think nothing of treating such AIs as we would any other machine, i.e. as a facilitator of human needs and desires. Our concern is not that Siri, Alexa or CIMON can suffer, feel pain, or have moral rights. We do not claim that we can wrong AIs by treating them as artificial slaves. Our concern is rather the negative effects this way of thinking of AIs may have on *us*.

At first look this might sound counterintuitive. Why should we not have the AI technology at our disposal to do our bidding and facilitate our lives as much as possible? To this we have two answers, the first is virtue ethical and the second consequentialist. The virtue ethics response goes as follows: In order to flourish and lead the good life we need to develop a set of moral and epistemic virtues that inform our behavior. Much of this instilling of virtues is about practicing and mimicking good behavior. If we are surrounded by artificial intelligent entities – AIs – programmed to behave like slaves, then that is unlikely to facilitate the development of our virtues. Indeed, it seems to allow, perhaps even encourage, us to behave viciously. Were the AIs rather to behave in a friendly manner, which includes setting boundaries, then they would regulate our behavior and at the very least not actively undermine the development of virtue. More speculatively, but not impossibly so, given that AIs will be more advanced in the future, utility AIs and social AIs could become role models and inspire virtue in humans.

The consequentialist objection to tolerating slave AIs is the following: If we get used to having our AI slaves doing our bidding that might spill over on how we behave towards human

[6] Cimon is the name of an influential Athenian statesman who lived c. 500 BC.

beings. While this claim is of course based on empirical assumptions, it seems likely that a certain behavior that we get used to in one setting (i.e. at home) may influence how we behave in other settings, say work, society and with friends. This holds true especially if the AI is advanced and perceived to be similar (capacity wise and/or look wise) to a human. Consider, for example, how over-consumption of violent films, video games etc. can have a normalizing effect and stimulate certain individuals to export their fantasies to real life.[7]

Our concerns about treating AIs as slaves differ in at least two ways from more mainstream approaches to AI ethics such as Asimov's laws of robot ethics. Firstly, the prime concern of Asimovs's principles is that humans might be harmed by robots. Secondly, the laws of robot ethics do not seem to allow for the type of as-if friendship we advocate. Consider Asimov's first law.[8]

First law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.

This law is incompatible with our account of friendly AI. It is possible that a friendly social AI behaving in a manner that mimics a sufficient number of aspects of proper friendship will *sometimes* injure humans, or allow that to happen, so doing so is not *always* wrong. The problem is that Asimov's first law is overly absolutistic. It is sometimes permissible to injure a friend physically or cause her psychological harm, or even allow a friend to injure herself. Avoiding harm is not the *only* moral value friends care about; other important values include truth and overall well-being. Imagine, for instance, as a person under the influence of alcohol is trying to start their car with the intention to drive off. A friendly social AI would intervene even if doing so would result in (light) physical harm to come to the driver.

Second law: A robot must obey the orders given it by human beings except where such orders would conflict with the first law.

[7] Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and social psychology bulletin*, *40*(5), 578-589. Adachi, P. J., & Willoughby, T. (2013). Demolishing the competition: The longitudinal link between competitive video games, competitive gambling, and aggression. *Journal of youth and adolescence*, *42*(7), 1090-1104.
[8] Asimov I. (1950), *I Robot*, Gnome Press, New York.

Asimov's second law is, just like the first, incompatible with the notion of friendly AI developed in this paper. An AI behaving in a friendly way should not *always* obey orders issued by a higher ranked authority. Unethical orders should typically be disobeyed. Imagine, for instance, that an authoritarian regime uses AI systems for erasing valuable historical records cataloging human thoughts, or for manipulating elections in foreign countries. Erasing historical records does not always under all circumstances, harm anyone (even if that may of course sometimes be the case), and it is theoretically possible that the manipulation of an election in a foreign country has no consequences at all for anyone. If, say, a single vote is counted twice that may not affect the distribution of seats in parliament. However, even if no one is harmed it would nevertheless be unethical to use AIs for erasing historical records or manipulating elections. We may, for example, believe that the collected thoughts of past generations have intrinsic value in an impersonal sense, just as fair elections. The problem for Asimov is that according to his second law, no AI would be allowed to disobey the orders described here even when they are unethical. A friendly social AI, on the other hand, would not participate in the destruction of valuable historical documents or manipulate elections.

Another reason for rejecting the second law is that it permits us to treat AIs as electronic slaves. As explained above we think that the electronic slave itself (the AI) is not harmed by being enslaved, but if slavery is permitted that is likely to have negative effects on us. If we get accustomed to the idea that we somehow own and control another intelligent (electronic) being, then that is likely to make us less sensitive to other moral issues. The wrongness of AI slavery is thus not supervenient on the wrongness of ordinary slavery, but rather on the negative effects AI slavery has on us. If we get accustomed to treating AIs as slaves this is likely to affect our dispositions toward each other in harmful ways, which will eventually harm ordinary humans. Consider, for instance, a Turing-like scenario in which you struggle to determine if the airline representative you are chatting with on the airline's website is a real person or an AI. If you become accustomed to treating AIs as slaves and believe that the 'agent' chatting with you is an AI you are more likely to not treat the unknown entity you are chatting with dignity and respect. If the airline representative is a real person she will be harmed by your behavior, at the same times as you may also be harming yourself indirectly by becoming less generous, moderate and respectful to others.

Third law: A robot must protect its own existence as long as such protection does not conflict with the first or second law

Unlike the first two laws, we believe that the third law is fairly uncontroversial. Indeed, we agree that friendly AIs should, within reason, protect their own existence. However, the discussion in this section shows that the notion of friendly AIs introduced above differs in fundamental ways from Asimov's theory of AI ethics. In the next couple of sections, we will render our notion of friendly AIs more precise.

3. Aristotle on Friendship

Before we critically discuss the suggestion that we ought to transform slave AIs into friendly utility AIs or friendly social AIs, it is helpful to provide an account of friendship.[9] A natural point of departure is Aristotle's seminal account in the *Nichomachean Ethics* (NE). Aristotle argues that there are three main qualities for which someone is cherished. The first is usefulness, the second is pleasantness, and the third is excellence. According to Aristotle these qualities translate into three types of friendship.[10]

1. Friendships based on mutual admiration
2. Friendships based on mutual pleasure
3. Friendships based on mutual advantage

According to Aristotle the first type of friendship is more valuable than the other two because it is based on excellence. In this type of friendship, what the two friends admire is the virtue of the other; it deals with the inner qualities of a person. It has been observed that this is a highly moralized idea of friendship which does not capture all the good of such relationships (see Cocking and Kennett, 2000). While that may well be true, the 'mutual admiration' type of friendship is not a description of the as-if friendship that we believe that AIs should be programmed to display. As made clear earlier we do not believe that human beings ought to have friendship feelings towards a machine. One reason for this is that there can be no reciprocity since the AI will merely be programmed to display a behavior that mimics, but is not proper, friendship feelings.

In what follows we will focus on the other two forms of friendship described by Aristotle. While they are lesser forms of friendship (as they are less complete), it has been

[9] This account draws on [Author 2008].
[10] See Book 8.3, /NE1156a6-8/. We discuss this account in [Author's work 2012]

argued that they might still qualify as genuine forms of friendship as they are similar enough.[11] This resembles our idea of 'as-if relationships' mimicking a sufficient number of the aspects that constitute proper friendship. Notably, on the Aristotelian account it is possible to have minor ends, e.g. utility and pleasure, and at the same time be committed to *eudaimonia* as the most *teleios* end.

Friendships of the second and third type are less excellent forms of friendship but as argued by Aristotle in /NE1157a30-33/ such relationships can still merit to be called friendships as they can also contain elements of genuine love and goodwill.[12] A possible reading is that while friendships based on mutual pleasure or mutual advantage may not contain all the features of friendships based on mutual admiration, they contain enough such features to qualify as friendships. This, however, is a contested reading. Critics have argued that friendships based on mutual pleasure or mutual advantage do not qualify at all, while others have suggested various intermediate positions.[13] Part of the debate regards how much goodwill friendships based on mutual pleasure or mutual advantage have to contain in order to 'count'.

To further complicate matters, Aristotle does not make it entirely clear what is meant by 'goodwill'. In some parts of the NE the concept of goodwill is used in a very broad sense, seemingly covering everything from the very low-level *philia* one might feel towards all of humanity to strong feelings of love and affection. In other parts, however, the notion of goodwill seems to refer to the general benevolence and respect we ought to feel towards other humans [14] whereas both 'liking' and 'loving' are narrower.

To make this more applicable to the subject discussed here, consider the following example. Ana loves her friend Dr. Mona for her willingness to let Ana jump the que at the emergency room and attend to medical issues whenever Ana or her children need medical care. Ana, who is a piano teacher, returns the favor to the best of her ability by offering Dr. Mona free piano lessons. Ana does not love Dr. Mona *in her own right* as she would if they had had a friendship based on mutual admiration, nor does Dr. Mona love Ana in her own right. A relationship like that between Ana and Dr. Mona does not meet Aristotle's friendship criteria of goodness. Ana does not love Dr. Mona for her own sake, or vice versa - quite to the contrary

---

[11] See e.g. Cooper J. M, (1977). Aristotle on Friendship. In, *Essays On Aristotle's Ethics*, ed. A O. Rorty, 1980, University of California Press, p. 305.
[12] See also /NE1157a25-35/ and /NE1158b5-11/).
[13] For a debate see e.g. Price, Walker and Cooper in Price A. W. (1989). *Love and Friendship in Plato and Aristotle*. Oxford: Clarendon. Cooper J. M, (1977). Aristotle on Friendship. In, *Essays On Aristotle's Ethics*, ed. A O. Rorty, 1980, University of California Press, pp. 301-340. Walker A. D. M. (1979). Aristotle's account of friendship in the Nicomachean Ethics, *Phronesis*, Volume 24, Number 2, 1979 , pp. 180-196(17).
[14] This is what we take Aristotle to mean when he talks about of *philia* for all others.

they seem to appreciate very different things in one another. What Ana finds good and useful in Dr. Mona might be her medical skills and her willingness to help Ana's children. What Dr. Mona finds good, pleasant and useful in Ana might be her music skills and her willingness to offer piano lessons.

What would be the problem in this situation? Very broadly it seems it could be rejected on two grounds. Firstly, this is not love; to the extent that there is any genuine goodwill and love here it is drowned out by the (mutual) exploitation, and secondly, there is goodwill and even a little love but the problem is that the feelings are primarily based on how useful the other is. If either party ceases to deliver it seems plausible that the relationship will lose its appeal to either one or both of the parties. But if this type of relationships are rejected, some scholars (see, for instance, Cooper 1977) have expressed concerns that a first-class friendship might become an unattainable ideal.[15] He, and others, argue that a better reading of Aristotle is that the 'perfect friendship' is intended as a paradigm case, not as a description of the only acceptable form of friendship we can have. It would then follow that the lesser forms of friendship can also be worthwhile and contain genuine well-wishing and goodwill for the sake of the other. Such a reading appears to be in line with Aristotle's general use of paradigm cases, i.e. the examples he uses for illustrating perfect instances of virtues like courage, generosity and indeed friendship. But while such instances undoubtedly are excellent, perfection is not the only way. There can also be friendships of a lesser, but 'good enough' kind. Exactly which elements such relationships would have to contain is an open question but this idea is in line with the argument in this article where we talk of AIs being programmed to mimic a sufficient number as aspects of proper friendship and act in line with them.

In the NE (e.g. in Book 8.9-8.12) Aristotle also talks about another form of friendship, namely that of civic friendship. Civic friendship is the type of general affection people in the same city or country could feel for each other, and such a feeling is central to the good functioning of the political society as it maintains the whole social project.[16] According to Aristotle both justice and friendship are central to the well-functioning society: the friendship binds the citizens together and enable them to go after a common goal but, equally, without justice the whole cooperative system just collapses. So while justice is key to maintaining political order, friendship also plays an important role in keeping society together:

---

[15] Cooper J. M, (1977). Aristotle on Friendship. In, *Essays On Aristotle's Ethics*, ed. A O. Rorty, 1980, University of California Press, p. 305.

[16] Aristotle compares the structure of the family to that of the structure of political society see e.g. /NE1160b24-27/ /NE1160b33-1161a3/.

> …friendship would seem to hold cities together, and legislators would seem to be more concerned about it than justice. For concord would seem to be similar to friendship, and they aim at concord among all, while they try above all to expel civil conflict, which is enmity. Further, if people are friends, they have no need of justice, but if they are just they need friendship in addition; and the justice that is most just seems to belong to friendship. (/NE1241a16-18/.)

While friends might not agree on everything they would certainly have a joint basic understanding of how life should be lived and what overall goals are worthy of pursuit.[17] Notably, however, this would not require the friends to lead very similar lives – they might certainly pursuit the good in different ways but they would, as Aristotle put it, share ”conversation and thought”.[18]

To reconnect the tripartite Aristotelian account of friendship with our account of AI friendliness recall how we reserved the term friendly AI for ‘utility AI’ and ‘social AI’. While our two categories do not map on to any single type of friendship as defined by Aristotle they both share features from pleasure and advantage but not from mutual admiration as that would require reciprocity and, possibly, complete virtue. These relationships will be further elaborated on below.

4. Two types of friendly AIs

Against this Aristotelian account of friendship, we are now in a position to flesh out our suggestion that slave AIs ought to be transformed into to friendly utility AIs or friendly social AIs. Recall that there is no reciprocity between AIs and humans. Neither the human user, nor the AI, ought to feel any proper friendship feelings toward each other.[19] The human user should simply recognize that the AI is a machine and the AI can at best be programmed to mimic

[17] /NE1241a16-18/.
[18] /NE1170b12/.
[19] Recall that we in section 1 defined proper friendship as containing sincere well-wishing, the intrinsic valuing of the other, helpfulness, empathy and a commitment to honesty as well as an obligation to sometimes put the good of the friend above your own. Can an AI behave *as if* it has such feelings? We believe that it could be programed to behave in such a way. After all it is the actual behaviour of the AI that will trigger the virtuous response and enable the cultivation of virtue in the human and it does not require the human to be under the illusion that the AI is something else or something more than a machine.

friendly behavior. This could, for instance, include behavior that displays sincere well-wishing, the intrinsic valuing of the other, the commitment to honesty, loyalty and other shared values.

The reason why the AI should be transformed into friendly utility AIs or friendly social AIs is that such interaction will facilitate the development of human virtues required for eudemonian lives. In order to maintain the virtues we need to exercise them, and as AI technologies get increasingly advanced they will be embedded in the fabric of society. Plausibly, then, many human-to-human contacts will be replaced by interactions with machines and, if those machines are not programmed to be friendly utility AIs or friendly social AIs, humans would be deprived of many opportunities to practice the virtues. This would be bad for us as it reasonably could be expected to result in less virtuous behavior, as we would get less skilled at identifying and being sensitive to relevant moral features of situations. Notably, our only concern is what is good for humans. We are not arguing that the AI itself could benefit, be harmed, or be affected in any other morally relevant way.

For an example of a friendly utility AI, consider a scenario in which an AI functions as a decision-support system nudging towards environmentally friendly transport choices: On a rainy morning it is more likely that car owners will drive to work than they are to use public transport. Pollution from cars has a negative impact on the environment so were more people to use public transport that would be positive. Therefore, a friendly utility AI could function as a decision-support system designed to nudge behavior promoting the collective good by for example providing information about local timetables, transport time to destination (it could well be faster than the car option), cost savings, remind the person of the positive effects on the environment their behavior would have and so on. Suggestions and prompters of a similar kind could equally well have come from a friendly human e.g. one's partner or a neighbor. Admittedly someone in a position of authority (instead of a friend) could make a similar suggestion but a difference then would be that they could force you to comply, which is not the case with a friend or the friendly AI: it merely makes a suggestion. Compare this with the notion that the AIs around us are here simply to facilitate our lives and make them as pleasurable and easy as possible. That would be the behavior we would expect from a slave AI, which could per definition not be friendly. Plausibly, such AIs could fuel a sense of indifference and stifling inertia with regards to decisions that impact others in a tangible way. As many decisions which are in line with the common good collide with short term personal convenience – e.g. taking the bus – an AI which incentivizes and prompts a less selfish and more socially and environmentally sustainable behavior would be significantly better than an AI programmed to accept being treated as a slave.

For an example of a friendly social AI, consider a scenario in which an AI system functions as a decision-support system as well as a partner or coach in the habituation of the virtues. Imagine, for instance, something like a future generation of CIMON. Such an AI could plausibly function as a partner facilitating emotional regulation and development of the moral and intellectual virtues of the human user. In addition to prompting positive behavior friendly AIs like future versions of CIMON could also assist in reducing behavior that would be negative for the user. Imagine for instance a person who frequently lies in their social media communication. On no plausible account of virtuous behavior could this qualify as acceptable. A friendly social AI should therefore assist not only in flagging the lie and possibly stop the posting but also play an important role in the development of the user's virtues. For example, a friendly social AI could explain why the behavior is bad, offer a set of training scenarios, and be a partner in practicing. The friendly social AI can be can be expected to contradict, criticize and reprimand when required and this would be required for facilitation of the development of the virtues. Such behavior will not be displayed by a friendly utility AI or a slave AI.

Understandably, these scenarios might raise concerns pertaining to lack of privacy, autonomy, informed consent and a reduction of the individual's freedom. We agree that these are important concerns. The development and use of friendly social AIs (and perhaps also of some friendly utility AIs) should be held to high ethical standards regarding transparency, safety, responsibility, justice, social sustainability and the promotion of wellbeing. However, the fact that friendly social AIs may raise ethical concerns of this sort is not a reason for not developing them. Nearly every new technology can trigger reasonable moral concerns if it is used in the wrong way. Consider for example, gene editing technologies or various forms of surveillance technologies. Both have great potential for good and for bad. What this shows is that we have to be careful when we develop and use friendly social AIs, not that it is unconditionally wrong to do so.

5. Concluding remarks

We believe the virtue-based approach to AI ethics outlined in this paper has several advantages over the traditional "value alignment" approach advocated by Stuart Russell and others.[20] There is much disagreement on how such a value alignment process is supposed to be conducted.[21] For example, how can AI systems be prevented from inadvertently acting in a way that is

[20] See, for instance, Russell (2016).
[21] For an overview, see [author's paper].

harmful to humans or incompatible with human values? And what specific list of values should AIs be designed to respect? Even assuming that the relevant values could be identified, individuated, agreed upon, and value conflicts mitigated, it would still be a complex task to identify the relevant utility function to be programmed into the AI.

The suggestion of the present article, namely that AIs should be programmed to *behave* in a manner that *mimics* a sufficient number of aspects of proper friendship, avoids at least some of these problems. Computers are good at analyzing similarities and differences in large data sets. Therefore, if the aim is to mimic certain examples of human behavior, which are identified ex ante by humans as virtuous, and then use those examples as prototypes in a training process, this seems more likely to result in stable AI systems that can be called friendly and promote human virtues in society. There will be no need to identify a complete set of moral values ex ante, nor to somehow balance conflicting values against each other with the aim of constructing a suitable utility function. By instead focusing the human input on identifying behavior that exemplifies prototypical virtues, the question about values becomes redundant, or at least secondary. So instead of a value alignment process, we recommend a virtue alignment process.

References

Adachi, P. J., & Willoughby, T. (2013). Demolishing the competition: The longitudinal link between competitive video games, competitive gambling, and aggression. *Journal of youth and adolescence*, *42*(7), 1090-1104.

Aristotle (1980). *The Nicomachean Ethics*, (translated by W. D. Ross), OUP.

Asimov I. (1950), *I Robot*, Gnome Press, New York.

Cocking, Dean, and Jeanette Kennett. "Friendship and moral danger." *The Journal of Philosophy* 97.5 (2000): 278-296.

Cooper J. M, (1977). Aristotle on Friendship. In, *Essays On Aristotle's Ethics*, ed. A O. Rorty, 1980, University of California Press

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and social psychology bulletin*, *40*(5), 578-589.

Price A. W. (1989). *Love and Friendship in Plato and Aristotle*. Oxford: Clarendon.

Russell, S. (2016). Should we fear supersmart robots. *Scientific American*, *314*(6), 58–59.

Strickland, Ashley C. (2020). “Robot arrives on the space station to keep astronauts company.” Retrieved 29 January 2020, from https://www.cnn.com/2019/12/06/world/cimon-space-station-scn-trnd/index.html

Walker A. D. M. (1979). Aristotle's account of friendship in the Nicomachean Ethics, *Phronesis*, Volume 24, Number 2, 1979 , pp. 180-196(17).