Ethics and Information Technology
https://doi.org/10.1007/s10676-021-09595-x

ORIGINAL PAPER

Problems with "Friendly AI"

Oliver Li[1]

Accepted: 7 April 2021

Abstract
On virtue ethical grounds, Barbro Fröding and Martin Peterson recently recommended that near-future AIs should be developed as 'Friendly AI'. AI in social interaction with humans should be programmed such that they mimic aspects of human friendship. While it is a reasonable goal to implement AI systems interacting with humans as Friendly AI, I identify four issues that need to be addressed concerning Friendly AI with Fröding's and Peterson's understanding of Friendly AI as a starting point. In a first step, I briefly recapitulate Fröding's and Peterson's arguments for Friendly AI. I then highlight some issues with Fröding's and Peterson's approach and line of reasoning and identify four problems related to the notion of Friendly AI, which all pertain to the role and need for humans' moral development. These are that (1) one should consider the moral tendencies and preferences of the humans interacting with a friendly AI, (2) it needs to be considered whether the humans interacting with a Friendly AI are still developing their virtues and character traits, (3) the indirect effects of replacing humans with Friendly AI should be considered with respect to the possibilities for humans to develop their moral virtues and that (4) the question whether the AI is perceived as some form of Artificial General Intelligence cannot be neglected. In conclusion, I argue that all of these four problems are related to humans moral development and that this observation strongly emphasizes the role and need for humans moral development in correlation to the accelerating development of AI-systems.

Keywords Artificial Intelligence · Friendly AI · Human interaction with AI · Virtue ethics

Introduction

Questions concerning AI ethics have often been framed from the point of view of how to implement ethical frameworks in AI systems. These systems function in some sense as 'agents', make decisions, or interact with humans. One obvious suggestion has been that any AI which achieves or approaches human-level intelligence should be programmed as 'Friendly AI' (Coeckelbergh, 2020, 52; Yudkowsky, 2008). In the case of AI with human-level intelligence or artificial general intelligence—AGI[1] the reasons for programming such AI as Friendly AI are quite clear. An AI with cognitive abilities similar to or exceeding human cognitive abilities would need to be friendly to ensure that it does not pose a threat to humanity. However, in AI as it is implemented at present, with limited abilities and lacking consciousness, there may be other reasons to suggest that an AI interacting with humans should be a friendly AI. Barbro Fröding and Martin Peterson have recently argued in this direction from the perspective of virtue ethics (Fröding & Peterson, 2020).

Virtue ethical approaches in the field of AI ethics have explicitly been called for by, for example, Thilo Hagendorff in an evaluation of the current status of AI ethics: "I argue that the prevalent approach of deontological AI ethics should be augmented with an approach oriented towards virtue ethics aiming at values and character dispositions" (Hagendorff, 2020, 112 my emphasis). However, even if reports like Ethically aligned Design published by the IEEE discuss virtue ethics and other 'classical' ethical approaches in philosophy and recommend implementing virtue ethics-based approaches in autonomous systems (The IEEE

✉ Oliver Li
oliver.li@teol.uu.se; oliver.li@crs.uu.se

[1] Department of Theology, Center for Multidisciplinary Research On Religion and Society (CRS Uppsala), Uppsala University, Box 511, 75120 Uppsala, Sweden

[1] For definitions of AGI or artificial superintelligence—ASI, see, for example, Boström, Nick *Superintelligence,* McCarthy, John. "From Here to Human-Level AI." in *Artificial Intelligence* or Kaplan, Andreas, and Michael Haenlein. "Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence." in *Business Horizons,* (Bostrom, 2014, 26, Kaplan and Haenlein, 2019; McCarthy 2007).

Published online: 28 April 2021

Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019), they do not discuss the effect of artificial systems on the development of virtues in humans (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). Likewise, Mark Coeckelbergh, for example, has suggested a relational approach to the question of whether AI-robots should be granted rights and discussed this approach in relation to virtue ethics (Coeckelbergh, 2010). Still, neither discusses the effect of artificial systems on the development of virtues in humans.

Fröding's and Peterson's work, therefore, is an example of a well-needed contribution to the discussion of virtue ethics as anticipated by Hagendorff or the report mentioned earlier. Their work also explicitly raises the question of possible adverse effects from our interaction with AI on us as humans (Fröding & Peterson, 2020, 4). They focus not merely on the ethics of an AI in itself but also on the effects of AI on human's ethical development, a perspective which, to the best of my knowledge, is seldom discussed.

In this paper, I wish to further investigate issues related to Friendly AI's implementation with Fröding's and Peterson's novel virtue ethical approach as a starting point. In a first step, I briefly recapitulate Fröding's and Peterson's arguments for Friendly AI. I then highlight some issues with Fröding's and Peterson's approach and line of reasoning. Furthermore, I identify four problems related to the notion of Friendly AI from a virtue ethical perspective, which all pertain to the role and need for humans' moral development. These are that (1) one should consider the moral tendencies and preferences of the humans interacting with a friendly AI, (2) it needs to be considered whether the humans interacting with a Friendly AI are still developing their virtues and character traits, (3) the indirect effects of replacing humans with Friendly AI should be considered with respect to the possibilities for humans to develop their moral virtues, and that (4) the question whether the AI is perceived as some form of artificial general intelligence cannot be neglected. In conclusion, I argue that all of these four problems are related to humans' moral development and that this observation strongly emphasizes the role and need for humans' moral development in correlation to the accelerating development of AI-systems.

A summary of Fröding's and Peterson's arguments for 'Friendly AI'

Initially, Fröding and Peterson give an example of an AI system, CIMON-2, which was intended to be a companion to lonely astronauts but was sometimes perceived as mean and unfriendly. They pose the question of which type of behavior should be programmed into AI-systems that fulfill social functions and interact with humans (Fröding & Peterson, 2020, 1). Fröding and Peterson importantly point out that being a friend and being friendly cannot be equated in general. They introduce the distinction between Slave AI, Utility AI, and Social AI (Fröding & Peterson, 2020, 2–3). According to them, the latter two forms of AI can and—as they subsequently argue for—should be implemented as Friendly AI. They also make the reasonable presupposition that the AI in question is not artificial general intelligence—AGI. Their Friendly AI based on present technology would lack human traits such as emotions or consciousness. They subsequently argue that treating AIs as slaves is morally wrong.

Central to Fröding's and Peterson's reasoning are the following two claims: (i) "If we are surrounded by artificial intelligent entities 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" and (ii) "If we get used to having our AI slaves doing our bidding that might spill over on how we behave toward human beings" (Fröding & Peterson, 2020, 3). They conclude negatively that "[…] 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" (Fröding & Peterson, 2020, 4) and positively that Friendly AI "[…] would regulate our behavior and at the very least not actively undermine the development of virtue" (Fröding & Peterson, 2020, 4).

Based on Aristotle's work on ethics, Fröding and Peterson identify three types of friendship: the first is based on mutual admiration, the second on mutual pleasure, and the third on mutual advantage (Fröding & Peterson, 2020, 4; Aristotle NE1156 a6-87). While they emphasize that Friendly AI cannot be based on mutual admiration, since this would presuppose reciprocity in the relationship, which is not realized in technology at present, they argue that as-if friendship based on features related to the second and third types of friendship is possible and that AIs interacting with humans should be transformed into Friendly Utility AI and Friendly Social AI which implement as-if friendship. However, even in these cases, they emphasize that there is no reciprocity between humans and the proposed Friendly AI (Fröding & Peterson, 2020, 6).

In Fröding's and Peterson's examples for Friendly Utility AI and Friendly Social AI, they underline that if AIs with which humans interact are not programmed to be friendly, that is, if they are Non-Friendly AI (observe that non-friendly does not entail unfriendly), then the humans "[…] would be deprived of many opportunities to practice (the) virtues" (Fröding & Peterson, 2020, 6).

In summary, at least the following four central thoughts in relation to Fröding's and Peterson's virtue-ethical perspective on Friendly AI can be identified.

(A1) Non-Friendly AI system would explicitly behave non-friendly, and would possibly be treated as a Slave AI. Interactions with the Non-Friendly AI and Slave AI would negatively affect the interacting humans. The interacting humans would not develop virtues and would possibly become less sensitive to moral issues.
(A2) Friendly AI would not actively undermine the development of virtues.
(A3) Friendly AI would play an important role in the development of human virtues.
(B) They assume that the Friendly AI is not some form of AGI.

The first three claims (A1-A3) are related to each other, and all suggest that Friendly AI would have positive consequences for the development of human virtues. While I do not question the implementation of Friendly AI as such (why should we develop non-friendly or even hostile systems?), I believe that some essential qualifications related to the implementation and use of Friendly AI need to be made.

A critical assessment of Fröding's and Peterson's understanding of Friendly AI

In the following section, these four central thoughts will serve as a starting point for assessing how humans should understand, implement, and relate to Friendly AI. I will also base my arguments on the virtue ethical approach applied by Fröding and Peterson. Here it can be noted that virtue ethics generally is understood as an agent-oriented form of ethics and that virtue ethics focuses on the virtues of humans. It thus is anthropocentric (Gunkel, 2012, 88, 159).

Furthermore, Aristotle described virtue as a mean between the extremes of excess and deficiency (Aristotle NE II 1106a25-1107a10). The development of virtues and character traits in a more common-sense meaning is a natural part of human development. It is clear that becoming an entirely virtuous person and a final perfected state of moral development cannot be achieved. It is likewise clear that there different stages in the moral development of humans. Typically, children are in, what I wish to denote as, a significant process of moral development in which significant learning takes place and habits are developed. Also, at least some humans have a relatively stable state of moral development for the better or worse.

To reveal and explicate some qualifications that need to be made in relation to Friendly AI, consider two different cases related to the above mentioned central thoughts (A1)–(A3). The first case would be (1) that the interacting human is *not* in a significant process of development, and the second (2) that the interacting human is developing his/her virtues. The first case can be further divided into two sub-cases: (1a) the interacting human is *not* in a significant process of development but instead has a relatively stable character with some well-defined virtues, and (1b) the interacting human is *not* in a significant process of development. Instead, s/he is considered to be a mean and vicious person.

Now consider the first claim (A1) that humans would not develop virtues and would possibly become less sensitive to moral issues if they were to interact with Non-Friendly AI systems or Slave AI (Fröding & Peterson, 2020, 3). In other words, as Fröding and Peterson tentatively claim: Getting "[…] used to having our AI slaves doing our bidding […] *might* spill over on how we behave toward human beings." (Fröding & Peterson, 2020, 3 my emphasis). I take it that they believe that this is a serious danger and that Friendly AI can avert this negative effect. At first glance, this seems reasonable.

However, if, as in (1a), the human is *not* in a process of development, then the behavior of the AI should not matter as much in the interaction with Slave AI. If the interacting human does *not* already have the disposition to behave viciously but instead has a stable character with some well-defined virtues, why would s/he act maliciously? In, for example, the CIMON-2 case, why would an astronaut, who has presumably been trained to have a stable character and react calmly and rationally, react maliciously even if the bot CIMON-2 responds inadequately and perhaps causes irritation? To better specify, even if it were true that the behavior of the AI can be seen as an indirect encouragement for bad behavior, it is hard to see why a person who does not have a tendency to behave badly or viciously, who has developed the *virtue* of acting friendly, for example, would actually behave viciously. A morally stable person would presumably be less affected by an AI behaving like a Slave.

In a second case, (1b) consider a mean and vicious person who is *not* in a process of development. Would it matter for this person that the AI behaves friendly? I believe not. The fact that the interacting person is already vicious and mean would most probably, in many cases, lead to an adverse reaction by this person. Again, if the astronaut in the above example were mean and vicious (although this is an unlikely scenario in reality due to the careful training of astronauts), an inadequate response by CIMON-2 would possibly cause irritation which in turn in a mean and vicious person most probably would lead to malicious behaviour. It is even conceivable that given the knowledge that the AI is merely a machine, a mean person might exhibit even meaner behavior since the person might believe that their behavior—mean or not—does not matter to a machine.

In these first two cases (1a) and (1b), humans are not in a significant process of development; the humans have developed their character and are psychologically stable. To be sure, as mentioned above, humans in some sense are always *in* development and can be affected by frequently encountering harmful behavior, for example, in brainwashing.

However, I take it that the Friendly AI suggested here is intended to interact with humans in situations that are not extreme. So in the case that the interacting humans are not in the significant process of developing their character, the implementation of Friendly AI does not seem to make a significant difference.

However, what if, as in case (2), the interacting human is developing his/her virtues? Although the claim (A1) that humans would not develop virtues and possibly become less sensitive to moral issues if they were to interact with non-Friendly AI systems or Slave AI seems correct further consideration reveals that there is no given answer.

What if the interacting human is a child? Imagine that a Friendly AI is taking care of a child. Firstly, if the child realizes that the caretaking AI is a machine—which a child may undoubtedly do—then it is unclear whether the child would *care* about behaving in any specific way. The lack of emotional connection to the machine could be harmful to the development of virtues in a child, as suggested in claim (A1), but that could importantly *also* be the case if the AI were Friendly AI. The child may learn that misbehaving towards the machine while interacting with the machine is not actually harmful for the machine since the machine has no emotions. The child may thus develop the habit of misbehaving simply because the AI is a machine and independent of the fact whether the machine is a Friendly AI. This would even be the case if the child merely believed that machine has no emotions.[2] Similarly, an argument based on the adverse effects on humans, resulting in the claim that humans should take a "moral stance" towards forms of artificial agents, has been put forward by Joel Parthemore and Blay Whitby (2014).

Furthermore, it is well-known that in the development of children, being confronted with a certain amount of unfriendly behaviors in their surroundings is essential for developing a child's character. By experiencing the consequences of bad behavior in other humans, the child realizes that it may be better *not* to behave in a bad manner and subsequently develops virtues (see, for example, von Tetzchner, 2005, chaps. 16, 17, 20). In other words, the child unconsciously might think, "I do not want to behave in this or that way." Of course, this only holds to a certain, very limited extent; a child should obviously not grow up with an excessive amount of unfriendliness or morally harmful behavior in his/her surroundings. This example suggests that it would be essential to make distinctions between stages of the moral development of humans interacting with an AI.

Thus, while it is correct in this case that interaction with Non-Friendly AI or Slave AI could result in, for example, children becoming less morally sensitive, this does not necessarily depend on the use of Slave AI or Non-Friendly AI. Possible habitual effects of interaction with AI cannot necessarily be averted by introducing Friendly AI instead of, for example, 'neutral' AI.

Similar issues occur concerning the second case (A2), in which Friendly AI is claimed not to actively undermine the development of virtue. Indeed, if, as in (1a) and (1b), the interacting humans are not in a significant process of developing their virtues or characters, then (A2) would obviously be true. However, then the alleged positive effects of Friendly AI are trivial. In the more interesting and presumably more relevant case (2) in which the interacting humans are in a process of development with respect to their virtues, such as a child or an adolescent, the situation is again not clear. Although it may matter that Friendly AI is friendly, in parallel to the first case, the mere fact and realization that the AI is a machine may lead to unexpected negative behavior. Consider the caretaking Friendly AI from the previous example. Again a child taken are of may learn and develop the habit to misbehave while interacting with the AI simply because the AI is a machine and independent of the fact whether the machine is a Friendly AI or not. To be sure, one could argue that this does not actively undermine the development of virtue. However, suppose it is correct that the mere fact and realization that the AI is a machine may lead to unexpected negative behavior. In that case, it is reasonable to think that the AI nevertheless is actively undermining by 'being' an AI, although not necessarily by its proposed friendliness.

Moreover, there is another case related to (A2) in which Friendly AI, although not actively, would undermine the development of virtue. Imagine the following situation. It is a good habit to open the door for an elderly or disabled person. This habit may eventually develop into a virtue. Nonetheless, in Swedish society, for example, the opportunity for this ethically virtuous behavior has almost disappeared. Older people, disabled people, or any person for that matter can easily open doors in public buildings by merely pressing a button. Here, I wish to emphasize that I have no objections to the general introduction of these types of aid or assistance. Still, since the opportunities for acting virtuously, in this case by opening the door for someone in need, are gone, I firmly believe that people who never have experienced the need to open doors for others in need possibly often do not even realize that there may be that kind of need.

The central thought behind the situation described transfers to other similar cases. What would happen if Friendly AI would replace healthcare or childcare professionals?

[2] In a scene in his recent novel, *Klara and the Sun,* Kazuo Ishiguro describes how a group of children bullies and intends to mistreat an 'AF—Artificial Friend'. Presumably, they believe that the AF is nothing but a machine without emotions and a consciousness of its own which, at least in this fictive case, allows for the possibility of misbehaving (Ishiguro, 2021,74–79).

Apart from the important ethical problems, how the caretakers would receive Friendly AI, the number of opportunities for humans to develop the qualities and ultimately the virtues involved in and needed for taking care of other children, older people, or sick people would decrease. If, as in the example above, childcare to a greater extent would be performed by Friendly AI, humans would to a lesser extent engage in childcare. However, humans, in general, evidently learn extensively by engaging in caretaking. Thus a broad or broader introduction of Friendly AI in society as Social AI or Utility AI, types of AI which Fröding and Peterson suggest should be friendly (Fröding & Peterson, 2020), would have an impact on the possibilities for humans to develop important and valuable social abilities and ethical virtues. Even Friendly AI would indirectly undermine the development of virtues. Using Friendly AI as a companion for lonely astronauts, as in the CIMON-2 example, is one thing; using Friendly AI on a broader scale, as Fröding and Peterson believe will be the case (Fröding & Peterson, 2020,1), would be another. Importantly, Fröding and Peterson do not consider the consequences for the humans the Friendly AI would be interacting with and the humans it would possibly replace. However, relationships with other humans have an intrinsic value both for the interacting and the replaced humans. One obvious consequence would be that there would be fewer opportunities for humans to develop in caretaking situations. The lack of such situations for human development would then have to be replaced by other training.

Furthermore, Sherry Turkle observes that it is a social choice whether we decide to introduce Friendly AI systems on a broader scale in health care or society in general (Turkle, 2011, 108). However, in relation to elderly care, she suspects "[…] that it has already been decided, irrevocably, that we have few resources to offer the elderly" (Turkle, 2011, 123). Turkle indirectly hints at the above conclusion that Friendly AI's introduction could have consequences for the moral development of those who are replaced by Friendly AI. She writes: "But in the long run, do we really want to make it easier for children to leave their parents? Does the 'feel-good moment' provided by the robot deceive people into feeling less need to visit?"(Turkle, 2011, 125). If Turkle is correct in her observations, the above indirect consequence of Friendly AI's introduction could become a reality even in the near future.

How about the third slightly stronger claim (A3) that Friendly AI would play an important role in developing human virtues? Would it be *better* to use Friendly AI? Firstly, in parallel to cases (A1) and (A2), persons who are not in a significant process of development regarding their virtues, as in (1a) and (1b), would presumably be more or less unaffected by the fact that the interacting AI is a Friendly AI. Also, since in (1a) and (1b), the interacting humans are not in a process of development, a Friendly AI obviously would have no part in the development anyway.

The case (2) in which the interacting humans are in a significant process of development is again more interesting. Here, I think, one can consider the following scenario. Imagine a Friendly AI, which, even when it does not agree with the human's behavior, kindly objects, explains, and attempts to persuade him/her to do the right thing. I take it that this is what Fröding and Peterson have in mind in one of their examples for a Friendly Social AI and how they might imagine a replacement or development of CIMON-2-like AIs. It seems that the Friendly AI could play an important role in developing human virtues in this case. Yet again, I believe that further consideration shows that this is not necessarily the case.

Importantly, by the assumption (B) that the Friendly AI is not an AGI, an interacting human would know that he/she is interacting with a machine. Consider furthermore that the interacting human is a child. If virtues are understood as complex rational, emotional, and social skills (Kraut, 2018), then the development of them would require the involvement of emotions, for example. However, it has been presupposed that the Friendly AI lacks emotions and consciousness, so how could the Friendly AI, which is presupposed not to be another sentient being, play an important role in the development of, for example, emotional skills? Indeed, it is a well-known fact from developmental psychology that children need to encounter and experience both positive and negative emotions, and that emotional attachment to other beings is essential in their development (see, for example, von Tetzchner, 2005, chaps. 16, 17, 20). Also, from an Aristotelian perspective, the importance of developing habits and dispositions in childhood has been emphasized (Hartman, 2013; Aristotle NE II 1103 b19-30).

Furthermore, the project of involving digital technology in childcare has been questioned from the point of view of psychology and research on the brain (Spitzer, 2014). In other words, why would a Friendly AI, which kindly objects, explains, and attempts to persuade a child to do the right thing, but which nevertheless, since it is not an AGI, lacks the necessary emotional connection to a child, have positive effects on the development of the child's virtues? I think, at least in the case of childcare, the positive impact of Friendly AI is questionable.

More generally and independent of the interacting human's age, consider that the moral persuasions by the Friendly AI are mainly based on the words and the content of the persuasions by the AI. However, more likely, in the case of human interaction, the persuasive power is also dependent on the emotional connection to the person who wishes to persuade the human. A loving and caring human partner is more likely to be successful in persuading the human than a person with whom the human does not have any emotional

connection, or even worse a person he/she might hate or have strong negative feelings towards, even if the actual content of the persuasion is the same. Even in a more general case, it seems that an AI's effect on human behavior does not merely depend on whether they are friendly or not but also whether the interacting human has and can establish a positive emotional relationship to the machine. It seems that the knowledge or even merely the belief that the Friendly AI is a machine, a non-conscious AI with no emotions, at least opens up for the same kind of behavior which was intended to be avoided by introducing Friendly AI and rejecting Slave AI. Thus, if the knowledge or belief that the Friendly AI is a machine matters, then the assumption (B) that the Friendly AI is not some form of AGI may also matter.

The possible consequences of whether an AI is perceived as an AGI with cognitive abilities on par with humans or as merely a machine has been thematized in, for example, the popular TV-series Westworld. One of the messages that the creators of the TV-series Westworld wish to convey seems to be precisely that it matters both whether the interacting AIs are perceived as AGI with feelings and consciousness or not and whether they actually are conscious sentient beings. In Jonathan Nolan's and Lisa Joy's depiction of a future scenario, humans seem to be fully capable of treating even Friendly AI (like Dolores in the first season) in vicious and even evil ways by believing that the beings they interact with are 'merely machines' (Nolan & Joy, 2016). However, as the same TV-series speculates in the third season, a fully conscious AI—again Dolores -, who eventually 'wakes up' to full self-consciousness, might invoke virtuous behavior in other humans—Caleb (Nolan & Joy, 2020). If Nolan and Joy are correct in their imagination, then Friendly AI's effects seem to be strongly dependent on the preferences and character traits the humans already have. Still, it also seems to matter if the AI (Friendly?) is conscious since the full spectrum of feelings, reasoning, and allegedly human behavior, at least in the case of one human (Caleb), seems to have positive effects. In other words, the chance that humans may accept recommendations from an AGI, which is at least equal to humans in its cognitive capacities, may be greater than in the case of a non-AGI. The interacting humans may connect with the AGI emotionally. They may accept them as authoritative and worthy of respect in some sense due to their cognitive abilities and the possibility to attach to them, at least if these AGIs are friendly, which should be a prerequisite for their creation according to Coeckelbergh or Yudkowsky. (Coeckelbergh, 2020, 52; Yudkowsky, 2008). Nevertheless, as I have hinted, how humans react to or interact with even Friendly-AGI would most probably strongly depend on the moral preferences already developed in the interacting human.

Final discussion and summary

The overall idea of developing AI-systems as Friendly AI is surely not objectionable. Others have even emphasized the necessity of the development of AGI systems as Friendly AI to ensure that such systems do not turn against us (Yudkowsky, 2008). Fröding's and Peterson's discussion of Friendly AI from a virtue ethical perspective even suggested that (A1) interactions with the Non-Friendly AI and Slave AI would negatively affect the interacting humans and that the interacting humans would not develop virtues and would possibly become less sensitive to moral issues, that (A2) Friendly AI would not actively undermine the development of virtues, and that (A3) Friendly AI would even play an important role in the development of human virtues.

In all three cases, the preceding discussion has shown that there are at least some aspects in the interaction with humans that require further consideration and investigation. Firstly, Fröding and Peterson in their article do not consider the moral tendencies and preferences the human already has. However, as I have argued, the behavior an AI would evoke in interaction with a human would depend on which kind of character the interacting human already has. Recently, Misselhorn, for example, has suggested methodological guidelines for developing AI-systems in geriatric care, taking into consideration the specific needs of older people (Misselhorn, 2020). Secondly, Fröding and Peterson do not consider whether and in which way the interacting humans are in a psychological or emotional process of developing their character. In this case, the above discussion suggests that one would have to consider with whom the AI is interacting. It is likely that there are significant differences between, for example, children and adults, even if they all are or may be in a process of developing their character traits. Thus one would need to carefully distinguish between Friendly AI designed for different purposes. The effects may differ depending on whether the interacting humans are, for example, children or Friendly AI is intended for adults. Even among adults, different cases should be considered.

Thirdly, in case (B), it also became clear that effects on the development of virtues in those who are replaced by Friendly AI also need to be considered. This, I firmly believe, is an important matter, which is often neglected in the discussion of the ethics of AI. At the same time, the ethical status of the AI itself and the effects of the AI on the people who are interacting with them has been the object of many lively discussions (see, for example, Hew, 2014; Gunkel, 2018; Floridi & Sanders, 2004; Mittelstadt et al., 2016), the fact that the implementation of Friendly AI or

Social AI would lead to a decreasing number of opportunities for humans to develop their virtues and social abilities, and train moral behavior, is, to the best of my knowledge, hardly ever discussed. Indeed, as I have explicated, Friendly AI would most probably indirectly undermine the development of virtues by decreasing opportunities for humans to develop and train virtues if introduced on a broad scale.

Fourth, in contrast to Fröding and Peterson, I believe whether the AI has emotions and consciousness or not should not neglected. It is conceivable that there are cases in which the mere knowledge that the AI does not have emotions and consciousness would lead to negative behavior. Nevertheless, even if the AI were an AGI with fully developed consciousness and possibly even emotions—which is not technically possible at present—this would not guarantee that humans would behave any better towards the AI or other humans, for that matter, since there is always an element of uncertainty in human behavior. However, it might be greater the chance that humans may accept recommendations from an AGI, which is at least equal to humans in its cognitive capacities. The humans may connect with the AGI emotionally since humans may accept them as authoritative and worthy of respect in some sense due to their cognitive abilities and the possibility to attach to them, at least if these AGIs are friendly.

Most importantly, all of the above conclusions pertain to the role and need for human's moral development. This connection can be further strengthened. Consider the first three cases A1-A3 which are also examples of what John Danaher, in his analysis of robot deception, denotes as superficial state deception: "The robot uses a deceptive signal to suggest that it has some capacity or internal state it actually lacks" (Danaher, 2020a). I take it that Friendly AI should not be treated as a case of hidden state deception, in which a robot is believed to conceal a capacity it actually has (Danaher, 2020a), since the Friendly AI would possibly only be seemingly friendly if the AI would be deceptive about a state or capacity it actually has. This latter case is, according to Danaher, problematic since it can be regarded as a form of betrayal (Danaher, 2020a). However, as I have understood it here, Friendly AI does not exhibit seemingly friendly behavior but actual friendly behavior.

If Friendly AI is a case of superficial state deception, then Danaher's thesis of ethical behaviorism would apply. Ethical behaviorism is the thesis that "[…] the ethical state of our interactions […] can be determined by their external behavioural states and cues only, and not by anything else" (Danaher, 2020a). Consequently, the actions and reactions of Friendly AI can be interpreted in terms of ethical behaviorism. Apart from the conclusion subsequently made by Danaher that based on ethical behaviorism, an AI, in this case Friendly AI, should be "welcomed in the moral circle" (Danaher, 2020b), he emphasizes that in our interaction with such systems, we should "[…] err on the side of caution, of over-inclusivity not under-inclusivity, when it comes to whom we owe duties" (Danaher, 2020a). This caution could be interpreted that in the case of Friendly AI, humans should presuppose that the AI has emotions and possibly consciousness, independent of the fact whether the Friendly AI is an AGI or not. In other words, it once again becomes clear that it is up to humans how they perceive and treat systems like Friendly AI and that they should treat them friendly.

As mentioned above, all of the above claims have in common that they focus on the role of the humans in the interaction, who they are, which kind of character they already may have developed, which kind of opportunities they may need to develop, and how they perceive or should perceive AI. They point back to the actual or potential moral development of humans. Thus it seems that the development of evermore advanced AI-systems in general and socially interacting AI in particular not only emphasizes the importance of programming AI as Friendly AI. It also emphasizes that *we* should carefully consider the development of the humans involved in the interaction with these AI systems. In particular, if the danger highlighted in the third conclusion, that the opportunities for humans to develop and train virtues decreases due to increased use of Social AI, friendly or not, turns out to be real, then not only individual humans, but even human society on a greater scale may eventually have serious problems with their moral development. Perhaps our human dignity at least may lie in our ability as human individuals and as humanity in total to develop our virtues to the same extent as the power of our creations increases; to use a well-known proverb: "With great power comes great responsibility".

Summarizing, one should consider the moral tendencies and preferences of the humans interacting with a Friendly AI. It also needs to be considered whether the humans interacting with a Friendly AI are still developing their virtues and character traits. Here further distinctions between, for example, different ages should be made. This conclusion would suggest further interdisciplinary studies involving ethics, philosophy, psychology, and sociology. Considering such distinctions between different groups of human individuals would also lead to different approaches within the programming of AI. It would presumably lead to more detailed and stricter legal regulation of the AIs involved and their use since different cases require different rules. The indirect effects of replacing humans with Friendly AI should importantly be considered with respect to the possibilities for humans to develop their moral virtues. Here a decreasing number of opportunities for human moral development would then have to be replaced by other situations in which humans can train their moral abilities. Thus, there is not only a need to develop AI in an adequate direction but also

a need for moral development of humans. Whether the AI is some form of Artificial General Intelligence cannot be neglected. All of these conclusions importantly focus on the actual development of the interacting or replaced humans. The ethical behaviorism suggested by Danaher further highlights the important role of humans and their actual moral development in the interaction with AI-systems. Thus, the problems with Friendly AI clearly point in the direction of the need for friendly humans.

Acknowledgements This paper's research is funded by the Wallenberg Foundations WASP-HS program within the project 'Artificial Intelligence, Democracy and Human Dignity'.

Funding Open access funding provided by Uppsala University. The research in this paper is funded by the Wallenberg Foundations WASP-HS program within the project 'Artificial Intelligence, Democracy and Human Dignity'.

Declarations

Conflict of interest The author declares that there are no conflict of interest.

References

Aristotle. (1926). Nicomachean ethics. In H. Rackham (Ed.), *Loeb class.*Harvard University Press.

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.

Coecklbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5

Danaher, J. (2020a). Robot betrayal: A guide to the ethics of robotic deception. *Ethics and Information Technology, 22*(2), 117–128. https://doi.org/10.1007/s10676-019-09520-3

Danaher, J. (2020b). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics, 26*(4), 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fröding, B., & Peterson, M. (2020). Friendly AI . *Ethics and Information Technology*. https://doi.org/10.1007/s10676-020-09556-w

Gunkel, D. J. (2012). *The machine question*. The MIT Press.

Gunkel, D. J. (2018). The other question: Can and should robots have rights? *Ethics and Information Technology, 20*(2), 87–99. https://doi.org/10.1007/s10676-017-9442-4

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120

Hartman, Edwin. (2013). Aristotle on character formation. In Christoph Luetge (Ed.), *Handbook of the philosophical foundations of business ethics.* (pp. 67–88). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-1494-6

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*(3), 197–206. https://doi.org/10.1007/s10676-014-9345-6

Ishiguro, K. (2021). *Klara and the sun*. Faber.

Kaplan, A., & Haenlein, M. (2019). Siri, siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15–25

Kraut, R. (2018). Aristotle's ethics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy 171* (pp. 1174–1182). Metaphysics Research Lab Stanford University.

McCarthy, J. (2007). From here to human-level AI. *Artificial Intelligence*, *171*(18), 1174–1182. https://doi.org/10.1016/j.artint.2007.10.009.

Misselhorn, C. (2020). Artificial systems with moral capacities? A research design and its implementation in a geriatric care system. *Artificial Intelligence*. https://doi.org/10.1016/j.artint.2019.103179

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Nolan, J., & Joy, L. (2016). *Westworld season 1*. HBO.

Nolan, J., & Joy, L. (2020). *Westworld season 3*. HBO.

Parthemore, J., & Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon Us. *International Journal of Machine Consciousness, 06*, 141–161. https://doi.org/10.1142/S1793843014400162

Spitzer, M. (2014). *Digitale demenz*. Droemer.

Turkle, S. (2011). *Alone together*. Basic Books.

Tetzchner, S. V. (2005). *Utvecklingspsykologi*. Studentlitteratur.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. *Intelligent systems, control and automation: Science and Engineering.*IEEE.

Yudkowsky, Eliezer. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks.* (pp. 308–45). Oxford University Press.