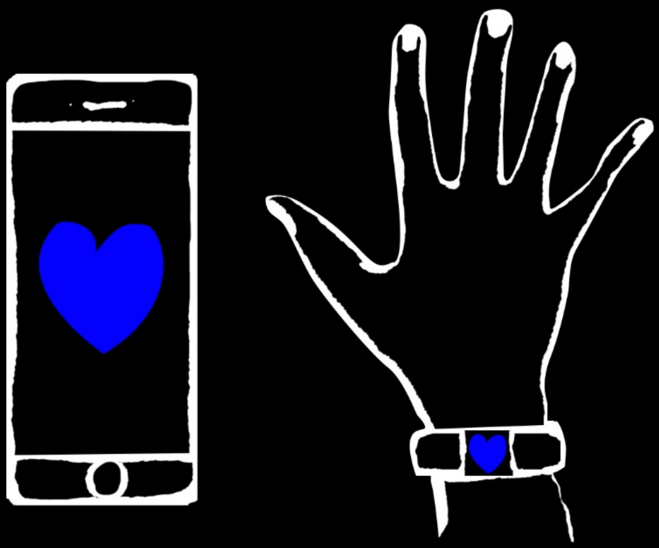

Virtue Epistemology and the Ethics of Profiling in Electronic Coaches

Mandi Astola
Student number: 3929388
Utrecht University

"History and Philosophy of Science"
Master's thesis
Supervisor: Joel Anderson
Word count: 23033

Contents

Introduction --- 2 - 5

1. Can an e-coach be a moral agent? -- 5 - 14

2. Views about the modes of moral agency -- 14 - 18

3. Virtues and e-coach agents --- 18 - 47

3.a. The importance of proper epistemic conduct in e-coaching --------------------------- 19 - 22

3.a.i. User profiling in electronic coaching -- 22 - 26

3.b Epistemic humility as a moral virtue for e-coaches -------------------------------------- 26 - 40

3.b.i. Why virtues? --- 26 - 28

3.b.ii. Virtue responsibilism -- 28 - 33

3.b.iii. Epistemic virtues in user profiling --- 33 - 40

3.c. Epistemic responsibilist virtue in different modes of agency ------------------------- 40 - 48

3.c.i. Epistemic humility in extended agents -- 40 - 46

3.c.ii. Epistemic humility in distributed agents-- 46

3.c.iii. Epistemic humility in machine agents -- 46 - 47

3.c.iv. Counterargument: Responsibilist virtues cannot be realized in machines ---- 47

4. Moral deplorability of epistemic arrogance depends on power ---------------------------------- 47 - 51

4 a. What is it for an e-coach to have power over its user? -------------------------------- 49 - 51

5. Power of e-coach over user: influence on authenticity --- 51 - 53

6. Two types of inauthenticity in e-coaching -- 54 - 58

7. Inauthenticity and power in different modes of e-coach agency --------------------------------- 58 - 62

7.a. Extended and distributed agents --- 59 - 62

7.b. Machine agents -- 62

8. Conclusion -- 62 - 66

Bibliography --- 66 - 73

Introduction

In 2050 you will probably be able have your house do your cooking and your shopping. You can have your watch plan your train travel. I think that you will be able to ask your house to help you lose weight and ask your trains to help you catch them on time. In 2050, we are going to get help for many of our problems because we will all have electronic coaches, or e-coaches helping us.

E-coaches are devices that work like coaches do. People consult e-coaches to help them reach a goal in behaviour change or habit maintenance. E-coaches collect information about the user and devise a personalized plan of action. The e-coach will employ different methods to try and get the user to reach their goal. E-coaches like that of Fitbit come with a wearable band that measures heart rate and movement. It tracks a user’s activity and gives recommendations based on that information. E-coaches are becoming more and more advanced and increasingly personalized to different users and their needs. E-coaches are bringing a lot of hope into healthcare and rehabilitation. Self-tracking for health provides healthcare practitioners, and the patient with a lot of data which is useful for diagnosis. E-coaching can also be a very effective and cheap way to help people keep to healthy lifestyles, which can prevent a lot of illness and suffering. E-coaches could however also be used for a variety of other purposes, like athletic coaching, coaching for a more sustainable lifestyle or getting over breakups or loss. These kinds of devices promise 24/7 assistance with our habits, our wishes and our deepest insecurities.

There are several worries that one might have about this kind of future. What happens to us if more of our behaviour is supported by algorithms? And what happens to our values? We might worry about whether algorithms steering our life can make us do things we didn’t really want to do, or end up in situations where we did not want to end up. We might also worry that we become less responsible for our actions because our lives are so co-influenced by algorithms. We might also worry about becoming inauthentic, turning into something other than our true selves. Although this investigation does not answer these worries, it can provide a useful starting point for doing so.

Worries about the influence of new technologies on our lives have often been met with the suggestion to consciously implement good morals and values into technologies. Value sensitive design is an example. It has been recognized that autonomous artificial intelligence and robots need to be programmed so that they act morally. Wallach and Allen have argued for the fittingness of virtue ethics as a moral theory to implement into machines.[1] The question I want to answer is whether e-coaches can be held responsible for conforming to the demands of virtue ethics. I will answer this question by looking at whether e-coaches are the types of agents that virtue theory can hold responsible. This requires also looking into the hybrid forms of agency that e-coaches can form when they become an integrated part of a human user's life. I will also look into what kinds of virtues would be important for e-coaches to have and whether it makes sense to hold e-coaches responsible for conforming to these.

There is already a rich body of literature on e-coaching ethics. Nickel has investigated the duty of e-coaches to present evidence of trustworthiness to users, as an important part of their success is being perceived as trustworthy.[2] Coeckelbergh and Sparrow have investigated the ethical issues relating to designating machines for tasks which traditionally demand the care of a real human being.[3] Quite some work has also been devoted to the influence of e-

[1] Wendell Wallach and Colin Allen, *Moral machines: Teaching robots right from wrong,* (Oxford: Oxford University Press, 2009).

[2] Philip J. Nickel, "Ethics in e-trust and e-trustworthiness: the case of direct computer-patient interfaces," *Ethics of Information Technology* 13 (2011): 355-363.

[3] Mark Coeckelbergh,"Health Care, Capabilities, and AI Assistive Technologies," *Ethical Theory and Moral Practice* 13 (2010): 181-190.
Robert Sparrow and Linda Sparrow, "In the hands of machines? The future of aged care," *Minds and Machines* 16 (2006): 141-161.

coaches on self-regulation and autonomy. Anderson and Kamphorst have argued that e-coaches can be beneficial to a user's capacity to self-regulate because they allow a user to overcome irrational first-order desires.[4] Heath and Anderson have written on the character of human willpower as extended by environmental scaffolding. Willpower can be outsourced to e-coaches just like it can be outsourced by locking a food cabinet to prevent oneself from overeating.[5] However, there is no literature addressing e-coaches from a virtue perspective.

My argument will go as follows. In section **1. Can an e-coach be a moral agent?** I argue that electronic coaches can be seen as independent machine agents, but also that they can be seen as hybrid agents, together with their user. This hybrid agency can be described either as distributed or extended agency. In section **2. Views about the modes of moral agency** I describe different views one can have about the possibility of these different modes of agency. I argue that they are all possible and which is the most appropriate description of agency depends on the level of integration between the e-coach and the user. In section **3. Virtues and e-coach agents** I evaluate the possibility of ascribing virtue and vice to e-coaches in the modes of agency I described. I argue that virtue ethics is suitable for evaluating agents in all modes. I also argue that since e-coaches are epistemic agents that profile human users, it is extremely important for an e-coach to be a virtuous epistemic agent. I also argue for the importance of epistemic humility in all three modes of e-coach agency, particularly with reference to self-knowledge. In section **4. Moral deplorability of epistemic arrogance depends on power** I argue that epistemic arrogance becomes more morally deplorable when the agent is being epistemically arrogant about a person over whom she has power. As power grows, bad epistemic practices become more deplorable. Therefore, when an e-coach has

[4] Joel Anderson, Bart J. Kamphorst, "Ethics of e-coaching: Implications of employing pervasive computing to promote healthy and sustainable lifestyles," at *The Third IEEE International Workshop on Social Implications of Pervasive Computing* (2014).

[5] Joseph Heath and Joel Anderson, "Procrastination and the Extended Will" in *The Thief of Time: Philosophical Essays on Procrastination,* ed. Chrisoula Andreou and Mark White, (New York: Oxford University Press, 2010).

more power over a person, it becomes more crucial that it is epistemically humble when profiling that person. In section **5. Power of e-coach over user: influence on authenticity** I argue that for an e-coach to haver power over a person (in the way that would make it more morally deplorable for it to be epistemically arrogant) the e-coach must be capable of causing inauthenticity. In section **6. Two types of inauthenticity in e-coaching**, I argue that this inauthenticity can take two forms. The first form involves the user not meeting desired ends and feeling alienated from one's projects, values or endeavours. The second form is when the process of acquiring values or deciding about projects and endeavours does not happen through the appropriate process. In section **7. Inauthenticity and power in different modes of e-coach agency** I discuss the implications of this view for the different modes that e-coaches can take. In section **8. Conclusion** I summarize my findings and explain how they show that e-coaches can and should be held accountable for meeting the demands of a virtue ethical view. Especially, e-coaches should be epistemically humble.

1. Can an e-coach be a moral agent?

In this section I will argue that an e-coach can be a moral agent. In fact, e-coaches can be moral agents in three different ways. These different modes of agency are *extended agency*, where the user of the e-coach extends to the device. The second mode is *distributed agency*, where agency is distributed over the user and the e-coach, or even a system of users and e-coaches. The third mode is when an e-coach is an independent *machine agent*. This "mode" is more of a perspective, because whether we consider this kind of agent or not depends on the actions we want to evaluate. So, which mode of agency we should attribute to an e-coach depends on this and also the level of integration of the e-coach and the user.

Let us begin by considering what it would be to think that an e-coach is not a moral agent. First of all, it is clear that an e-coach can be considered an agent at least in some senses. It is an agent in the sense that it does things. A machine, like a robot, is often called “autonomous” if it can perform tasks independently from human control and decide the tasks it performs in reaction to things it detects in the environment.[6] In fact, e-coaches are defined by their ability to react to information from their environment and to be proactive, meaning that they could be described as autonomous in the way that “autonomous robots” are described as autonomous.[7]

Denying an e-coach moral agency would require arguing that morality is not something like “robot autonomy” that can be attributed to “agents” which have, for example, no deliberated intentions. This view might be characterised as a traditional one. In information ethics there has been a move towards rejecting traditional views where moral agency centres on human characteristics like emotion or intentions. Authors like Floridi have suggested that now there are newly emerging problems of environmental ethics and computer ethics are new contexts where we cannot find a single human culprit. Therefore, we should not have an anthropocentric view of morality where moral agency is determined by the human characteristics of the agent. We should instead adopt a view where moral agency is determined by the receiving side, or the effects of the actions of an agent, where this agent can be a group, an organization or a machine.[8] E-coaches certainly perform morally significant actions. Like for instance, nudging someone to walk stairs instead of taking the lift, or encouraging the sharing of data with peers. Such acts influence a user and can do either harm or good, therefore they are morally significant. While Floridi would grant an e-coach moral

[6] George A. Bekey, *Autonomous Robots: From Biological Inspiration to Implementation and Control*, (Cambridge, Mass.: The MIT Press, 2005).
[7] Bart A. Kamphorst, “E-Coaching Systems: What They Are, and What They Aren't,” *Personal and Ubiquitous Computing* 21 (2017): 625-632.
[8] Luciano Floridi, “Information ethics: On the philosophical foundation of computer ethics,” *Ethics and Information Technology* 1 (1999): 37–56.

agency, opponents might argue that the e-coach is in fact not the agent in this moral situation, but merely a tool or a part of the circumstances of the action. Let us examine this possibility.

Let us imagine the e-coach is seen as either a mediator of the user's moral acts or as part of the situation in which the user performs moral acts. In the same way that we would say that a knife is just a mediator of the immoral act of manslaughter. Or we see the knife as a part of the circumstances of the murder, for instance because it was the nearest sharp object the killer could find. The e-coach is a device which helps the user make a change in their behaviour, just like a nicotine patch or anti-nail-biting nail polish. We can talk about the e-coach as a tool of an agent for performing moral actions. We could say that the user acts morally upon herself when she purchases and uses the e-coach. We could also say that e-coaching creates a certain situation to which the user responds in a particular virtuous or vicious way. We could also talk about the designer or marketer of the e-coach as a moral actor making moral acts upon the buyer and user of the e-coach.

Let us consider how tenable this kind of view is. The most important question to ask is: if e-coaches are tools, then how do we explain their agent-like characteristics? E-coaches sense information about their surroundings and ask for information from the user, they learn about the user based on that information, and sometimes also based on the information of other users. They have theories implemented into them. Fitbit for instance contains functions for the calculation of how many calories are burnt with a certain amount of activity and a certain heart rate. Can something that uses information autonomously, and information of which the user is often times not aware, be simply a tool?

Seeing technologies as morally neutral tools is sometimes called instrumentalism. Very few philosophers of technology defend this view because of the multiple ways in which technologies are non-neutral.[9] For example, Don Ihde famously argues that each technology transforms experience. Looking at the world through a pair of glasses is fundamentally different from looking at the world without them, because they create the relation of "looking through."[10] It has also been argued that technologies themselves have unforeseen consequences (climate change as a result of cars) which makes them non-neutral.[11] Technologies also embody moral values, like speed bumps force cars to slow down and therefore embody the value of traffic safety, as Latour has argued. [12] This is further reason not to think of e-coaches as mere tools.

I do not think it makes sense to deny someone or something moral agency just because they are in some way being "used like a tool." Anything can be said to be used by someone. I could use my lawyer to get out of having to pay a fine. I "use her" to achieve something. However, it doesn't make sense to say that my lawyer has no agency in the moral act of the evasion of the fine. The whole point of an e-coach is that it functions like a coach, so like a moral agent, and not like a tool. Therefore I think that in the context of e-coach mediated moral acts it makes more sense to see the e-coach as an agent rather than a tool.

If e-coaches are more than just tools or parts of the circumstances of an action, then one could still argue that they are just very sophisticated tools or parts of the environment that are embedded in agency. Heersmink describes the relation of some artifacts to the mind as cognitive embedment. This is when an artefact is important for carrying out cognition, but is

[9] Richard Heersmink, "Extended mind and cognitive enhancement: moral aspects of cognitive artifacts," *Phenomenology and the Cognitive Sciences* 16 (2017): 17 – 32.

[10] Don Ihde, *Technology and the Lifeworld: from garden to earth,* (Bloomington, Ind.: Indiana University Press: 1990).

[11] Heersmink, "Extended mind and cognitive enhancement," (2017).

[12] Bruno Latour, "On technical mediation: philosophy, sociology, genealogy," *Common Knowledge* 3 (1994): 29-64.

not a part of the extended mind.[13] One could also see e-coaches as embedded in agency, but not active in it.

However, let us think about how tenable this views is. Verbeek argues that we should see moral agency as extending to artifacts when those artifacts have a significant influence on the outcomes of moral decisions.[14] Heersmink argues that this is too loose a criteria, and we should rather see such technologies as embedded into agency than a part of it, unless they are extremely well integrated into the functioning of the user.[15] I think that there is some sense intuitively in saying that an artefact is only part of someone's agency when that artefact is very much merged with the functioning of the person. However, there is a very intuitive counterargument against this idea. Floridi and Sanders give the following argument in their paper for the plausibility of system-level moral agency: Imagine two nurses who kill a patient by accident. Both behave exactly in the same way but one is an artificial agent and the other a human. Why should we say one is a moral agent and the other not?[16] I will give an argument along these lines too. Imagine that someone uses an e-coach to live more sustainable and succeeds in adopting the habit of waste separation and energy saving. We could then evaluate the e-coach as good. Why should it matter how integrated the user and the e-coach are, if the result is our evaluation of the e-coach as good? Why should this matter for the existence of moral agency for the e-coach? In both cases, the e-coach actively causes a change in the outcome of a moral action. And in both cases we evaluate the e-coach as good.

I have explained why it makes sense to grant e-coaches some kind of moral agency. As I see it, there are three modes in which an e-coach can have moral agency:

[13] Heersmink, "Extended mind and cognitive enhancement," (2017).

[14] Peter-Paul Verbeek, *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press, 2011.

[15] Heersmink, "Extended mind and cognitive enhancement," (2017).

[16] Luciano Floridi and J. W. Sanders, "On the Morality of Artificial Agents," *Minds and Machine* 14 (2004), 349-379.

Extended moral agency: Moral agency resides in between the user and the device. We are evaluating morally the extended agent's actions in coaching itself, so the system is agent and patient.

Distributed moral agency: Moral agency is distributed across the system which involves the user, the e-coach and also the creators of the e-coach. We are evaluating morally the distributed agent's actions in coaching itself, so the system is agent and patient.

Machine moral agency: Moral agency resides in the e-coach, which we can see as an independent moral agent. We are evaluating morally the e-coaches actions in coaching the human user.

I shall explain all three modes in detail.

Extended moral agency involves a human agent and an external object like an e-coach which is so integrated into the agency of the human that it becomes more than a tool. It actually becomes a part of the human agent because the user's agency extends to the device. An adherent of the extended mind thesis might also describe the device as a part of the user's extended mind and therefore also of her agency.

Many different philosophers have looked at environmental scaffolding for cognitive or moral purposes and concluded that the environment should be considered a part of the agent in some way. The forerunner to this is the extended mind thesis of Clark and Chalmers. The extended mind thesis argues that since many components of cognitive functions (like counting with fingers, or with pen and paper) happen outside the skull, we might as well consider them a part of the mind, just like we do with parts of the brain.[17]

[17] Andy Clark and David J. Chalmers, "The extended mind," *Analysis*, 58 (1998): 7-19.

Howell has used Clark and Chalmers's line of argument to argue for extended personhood, that the virtues and vices of a person can also extend to the environment. Extended personhood can, I think, also be interpreted as extended agency within a virtue ethical framework at least. This is because what Howell means by personhood is simply being a thing that is held accountable for moral actions. This might as well mean "moral agent." He argues that if we accept the extended mind thesis, that a person's mind can extend to the environment, it is very easy to accept that her personhood does too as it is a part of her mind. Howell adds that even if one doesn't believe in the extended mind thesis, one can still accept the extended virtue thesis. There is no reason why the thing that we hold responsible for something should stop at the skin of an individual. Howell gives the example of a man who is known to have a prevalent sex drive, causing associated behaviour. As soon as he has children and spends a lot of time caring for them, his sex drive diminishes. This is a real effect of caring for children on men, through a lowering of testosterone levels. In this case, the man gains a character trait because of daily interactions with his children. If his children were not there, he would not have the character of sexual temperance. Therefore, the grounds for his character trait are outside himself. There is therefore reason to believe that his character trait exists within his body as well as outside it.[18] According to Howell's view, it would make sense to say that when a person is operating together with an algorithm, that the moral responsibility for the results of this behaviour resides in the *user-e-coach-system*.

For Howell, it still seems essential that there is a human being at the centre of the system. Systems can have moral agency only insofar as they are centrally controlled by a human person. However, philosophers like Floridi and Verbeek argue that moral

[18] Robert J. Howell, "Extended virtues and the boundaries of persons," *Journal of the American Philosophical Association* 2 (2016), 146—163.

responsibility is a system property and not a human property.[19] This is idea is central to the second mode of moral agency, which I will explain now.

Distributed moral agency means that moral agency is distributed across the system which involves the user, the e-coach and perhaps also the creators of the e-coach. This view requires that agency can be distributed across a system of human or artificial agents or things. Floridi argues that agency can be distributed among humans and artefacts. He gives the example of a customer loyalty scheme in a bank, where the bank automatically donates 15 pounds to charity whenever a customer opens a credit account. Although the donation by itself is a small act, the larger impact of such a scheme is large and morally admirable. Therefore, it can be said of such a system that it, as a whole, is responsible for the moral act of donating lots of money to charity.[20] What matters for whether these systems are moral agents or not is not so much their constitution but rather their impact and whether this can be classified as moral or not. Floridi writes “we need to evaluate actions not from a sender but rather from a receiver perspective: actions are assessed on the basis of their impact on the well-being of the environment at large and its inhabitants specifically.” [21]

Is this kind of detached and impersonal system-level view of moral agency justified? Floridi and Sanders have provided a justification elsewhere. The authors point out that whenever we wish to define something we must first define the Level of Abstaction (LoA) that we are looking at that thing from. There are some things of which there is only one conceivable LoA, where it is obvious to everyone what is said, like that tomatoes are berries. This is because “being a berry” is only relevant to one particular level of abstraction, within which a tomato is a berry. However, systems like machines and living things can also be

[19] Richard Heersmink, “Distributed cognition and distributed morality: Agency, artifacts and systems,” *Science and Engineering Ethics* 23 (2017), 431-448.
[20] Luciano Floridi, “Distributed Morality in an Information Society,” *Science and Engineering Ethics* 19 (2013): 727–743.
[21] Ibid.

understood from various levels of abstraction. And when we try to attribute to those things a fuzzier quality, like having agency, then specifying the LoA becomes important. If we want to attribute agency to anything, we must first specify the LoA at which we are talking.[22]

Floridi and Sanders then outline three characteristics of agency. An agent must be interactive, meaning that it can be influenced by the environment as well as being capable of influencing the environment. An agent must also be autonomous in the sense that it is able to change state without direct response to interaction. An agent is also adaptable, meaning that it *learns* or that its interactions can alter the rules by which it changes state. Sometimes things can satisfy these criteria and qualify as having agency at one LoA but not at another. They give the example of the noughts and crosses playing machine, MENACE. The plays noughts and crosses and learns to play over a series of games. It has markers for each possible square and every time it loses a game, a marker is removed from a square played in a losing game. If we consider MENACE at a single game LoA, then it is autonomous and interactive, but not adaptive, because within a single game we have no means to see its learning. If we look at MENACE at a multiple game LoA, then we see that it does in fact adapt, and we can attribute to it agency. However, if we look at MENACE at a system LoA, then we see that its adaptiveness is simply a mechanical update rule, and the system ceases to be an agent. Floridi and Sanders argue here that the difference between the multiple game LoA ad the system LoA is the fact that at one level the mechanism of adaptability is hidden from us. This hiddenness makes for the fact that we call something an agent. A human is an agent whenever the causal (neural) mechanism by which it adapts, interacts or acts autonomously is hidden from us, while it retains all three characteristics. A machine too can be an agent at an LoA where these aspects are hidden.

[22] Floridi and Sanders, “On the Morality of Artificial Agents,” (2004).

Having established that non-human systems can be agents at the appropriate LoA, Floridi and Sander's argue that if they can be agents, they surely can be moral agents. Moral agency requires the capability to do moral good or evil. For instance, someone caring for patients is performing moral acts, as there is a function to determine whether he is doing good or not. This function involves he actual wellbeing of patients and their desired wellbeing. Whether a human or a machine is influencing the relationship between these variables, it should be considered a moral agent.[23]

If we accept Floridi and Sanders's thesis that moral agency can attach to systems, rather than just humans, we might see the e-coach-user system as an agent. Within this view, it is the system properties that are important in determining agency. This kind of agency could also be distributed across groups of human being or machines. This is why the distributed agency theory has often been used when talking about the spreading of agency over a collaborating system of multiple persons, whereas the extended agency has been discussed in the context of relationships of a single user to a technology.[24]

We have just discussed distributed moral agency, where moral agency attaches to a system. Now we come to the third mode of agency, namely machine agency, which Floridi's view also justifies. If moral agency attaches to systems which have autonomous capacities with morally significant influences on their surroundings, then there is no reason why there would need to be a human being within such a system. Accepting Floridi's view about moral agency means that moral agents do not need to be human or even have human components. If we look upon the e-coach as a full moral agent we can see it as acting upon the user, who is then the moral patient. In this view the user and the e-coach are seen as separate entities.

[23] Floridi and Sanders, "On the Morality of Artificial Agents," (2004).

[24] Heersmink, "Distributed cognition and distributed morality," (2017).

2. Views about the three modes of agency

In this section I argue for a view which states that all three modes of agency are possible in the same world and valid. The three modes of agency attributable to e-coaches described in the previous section are extended agency, distributed agency and machine agency. Which form of agency an e-coach subscribes to depends on the characteristics of the device.

One who sympathizes with Floridi might go further and argue that system-level morality is the only kind of morality that is possible. And therefore all agency would be distributed. Whether we allow for agents that consist of a single human being or a single machine would depend on what level of interaction they have with their surroundings. A machine performing moral acts together with a person or a group of people would be considered a distributed agent, whereas a machine operating in isolation would be considered a distributed agent where moral agency is distributed over the parts and functions of the machine. This view sees extended agency and distributed agency as mutually exclusive. We either have distributed agency or not, in which case we have extended agency. One view on these modes is thus that distributed agency is not compatible with extended agency.

Another view on these modes is that pure machine agency is not compatible with extended or distributed agency. One might argue that if moral agency can be distributed, that this always happens along the lines of what Heersmink calls “moral bloat”.[25] Moral bloat is akin to the cognitive bloat issue in the extended mind theory. The cognitive bloat objection points to the fact that if our mind if not confined by our skull, then what is to prevent us from saying that the whole internet is a part of our mind and we know absolutely everything that we could look up on the internet? Moral bloat would be the distribution of agency among

[25] Heersmink, “Distributed cognition and distributed morality,” (2017).

every interrelated thing in the world. Then a machine could never be a single agent, as it is always an agent *together* with its maker, its operator and the society it exists in. This view states that machine agency is mutually exclusive with extended or distributed agency.

Heersmink's view is that artefacts cannot be agents by themselves and therefore machine agency or Floridi's sense of distributed agency cannot exist, while extended agency can. They can only constitute agency when they are very much integrated into a person's cognitive and moral functioning, in which case they form extended moral agents. Agency is in such cases located in between the human and the device. It seems to me that Heersmink's argument for this is that systems in themselves cannot be moral agents because they are not cognitive agents either. He writes "It seems to me that artifacts do not have cognitive agency because they are not cognitive agents, that is, they do not have the capacity to cognize. In order for something to have cognitive agency, it must have the capacity to initiate thoughts and mental states such as beliefs, desires, or intentions."[26] While artifacts can be important in cognition, they lack the capacity to initiate cognition or cognitive states. A pure machine cognitive agent is therefore not possible. Heersmink seems to take moral agency as involving moral cognition. Therefore if a machine cannot be a cognitive agent because it cannot initiate cognitive activities, then it cannot be a morally cognitive agent and hence also not a moral agent. Heersmink argues against the idea of system-level agency by arguing that a cognitive or moral agent needs the capacity to initiate cognitive processes.[27] However, one could also argue that a moral agent needs to have emotions and emotionally driven intentions. This is if one adopts an ethical theory where sentiments or consciousness are crucial to the performance of moral acts.

[26] Heersmink, "Distributed cognition and distributed morality," (2017).
[27] Ibid.

A fourth view, that I wish to defend, is that all the modes of agency, extended, distributed and machine agency are possible and depend on the kind of device operation ion question and the kind of actions we are considering. For this I wish to draw on Heersmink's conclusion. Heersmink recognizes that there are important connections between the role of artifacts in cognition and their role in morality. He argues that whether an artefact is a part of a moral agent or not depends on how well the artefact is integrated into the cognition of a person. If an artefact is morally significant (influences the outcome of moral decisions) and is integrated into the cognitive system of the person, like the cane for a blind person, then it can be considered an extension of the moral agency of the person. Heersmink does not fix a level of integration as a threshold condition, rather, the extendedness of the agency is a matter of degrees.[28] I would like to take this idea of extendedness being on a spectrum and apply it to distributed and machine agency too. Heersmink would probably not agree with me, as he rejects the possibility of system-level agency that does not attach to the ability to initiate things. But I think that taking his idea a step further reveals a broader and more encompassing view of agency which is what we should look for when we are evaluating the morality of machines.

I think that in cases where a machine and its workings have clear moral significance, but are reasonably opaque to human users, we can speak of machine agency. An e-coach can be an independent moral agent in itself, with a human user as a separate moral patient. However, if a machine can be an independent moral agent, then we may also look at the device, together with the user, as being an independent moral agent. If we accept that a device can be an agent because system –level morality is possible, then there is nothing preventing us from seeing the user as a part of the machine. The two together can constitute an agent. Whether we see an e-coach as an independent agent or as an agent distributed over the device

[28] Ibid.

and the user is a matter of specifying which moral actions we are evaluating. If we are trying to evaluate the overall effects of introducing the workout scheme of an e-coach into a person's life, and the changes it makes on that person, then we should look at the e-coach as an independent moral agent, acting on the moral patient, the user. However, if we are evaluating the acts of a person who is aided by an e-coach, then it will make more sense to consider the actor to be a distributed agent consisting of the e-coach and the human in operation together. Extended agency, I take to be something that happens like Heersmink describes, when the device is transparent to the user and fully integrated into their cognitive and moral reasoning. In such cases, the human uses the e-coach like a blind person uses a cane.

I have a few reasons for favouring this view. First of all, like I explained before, it is counterintuitive to think that the level of integration between an e-coach and a user would change the degree of agency we attribute to them when we evaluate their moral actions. Secondly, it allows for a broader and less anthropocentric view of moral agency, the advantages of which are well defended by philosophers like Floridi.

3. Virtues and e-coach agents

In this section I investigate the compatibility and the possibilities of virtue theories for understanding the ethics of e-coaching. A large part of the functioning of an e-coach is epistemic, and therefore epistemic virtues are relevant. While human coaches are often described in terms of virtue, devices are often spoken of in terms of reliability. But I think that terms of virtue should be attributed to devices too, especially since they can form hybrid agents together with humans.

I argue that a large part of the scope of the moral actions of an e-coach are epistemic actions, which makes epistemic virtue relevant for e-coaches. I will motivate my choice to venture into virtue ethics as a model for e-coach morality. Then I explain Zagzebski's responsibilist virtue epistemology and how it functions as an ethical framework. I will also argue that proper user-profiling should conform to responsibilist epistemic virtue like described in Zagzebski's framework. Using this framework, I will outline the importance of epistemic humility as a key virtue for e-coaches. I will also argue that it is possible to apply this responsibilist framework to all three modes of agency. Epistemic responsibilist virtues can be extended from human users to devices. They can also be distributed across humans and devices and they can be present in machine agents. My purpose is to show that responsibilist virtue epistemology, as well as other virtue ethical theories, are desirable and possible means of evaluating electronic coaches. E-coach agents can be evaluated and held accountable according to a virtue ethical framework.

3.a. The importance of proper epistemic conduct in e-coaching

Let us begin by examining features of e-coaches that call for ethical analysis. Firstly, it is useful to define what e-coaches are and describe how they work. I take e-coaches to be any devices which function like coaches. Kamphorst states that most electronic coaches use a *goal-oriented strategy* where coaching is about helping someone reach their goals.[29] The goal-oriented approach is described by Ives "The primary method is assisting the client to identify and form well-crafted goals and develop an effective action plan. The role of the coach is to stimulate ideas and action and to ensure that the goals are consistent with the client's main life values an interest, rather than working on helping the client to adjust her

[29] Kamphorst, "E-Coaching Systems," (2017).

values."[30] According to this account, a coach is someone who helps another reach her own goals while using her own strength and capabilities.

For my purpose, I do not think that a strict and exclusive definition of what constitutes an e-coach is needed. Kamphorst has constructed a detailed definition of what an e-coach entails. He assembles his definition with reference to what kind of devices he thinks are likely to succeed in changing behaviour. However, people use a variety of algorithmic mediations in assisting them with their different goals. The exact features of e-coaches are, I think, likely to change in the next 20 years. My investigation will be the most useful if it can identify aspects of electronic coaches which have particular influences on authenticity and individuality. For this purpose, and for providing a rough definition, Kamphorst's characterization of different key aspects of e-coaches will be useful.

Bart Kamphorst gives a list of criteria for e-coaches if we are to understand them as devices able to "create and maintain customized, collaborative relationships in which coachees are supported in understanding their situation and in making effective plans for changing their behavior or attitudes in accordance to their own view on how to live their lives."[31] The attributes than an e-coach must have to meet this definition is the following:

1. The system must have social ability, to collaborate with the user
2. The system should be perceived as credible by the user
3. The system must be context aware, to adapt to the user's environment
4. The system must be able to communicate with other systems in a user's ambient environment
5. The system must incorporate data streams from the user to customize its activity
6. The system must give tailored feedback and advice

[30] Yosse Ives, "What is coaching?: an exploration of connecting paradigms," *International Journal of Evidence Based Coaching and Mentoring* 6 (2008): 100-113.
[31] Kamphorst, "E-Coaching Systems," (2017).

7. The system must initiate actions proactively
8. The system must operate using a model of behaviour change
9. The system must guide its users through intention formation and planning. [32]

Kamphorst also specifies that aspects like 1 and 9 are what differentiate e-coaching from nudging.[33] This is because e-coaching involves the user actively into the process of behaviour change. Rather than a nudge-strategy, e-coaching could be seen as a think-strategy. A think-strategy, as described by John et al in the book *Nudge, Nudge, Think, Think* is one which aims to change civic behaviour by involving citizens in decisions about their community so that they consciously adopt behaviour changes.[34]

A system like the Philips Health Watch and the accompanying software Philips HealthSuite is an example of an e-coach that fits this definition. The Health Watch measures physical data like heart rate and movement. It also asks the user to set goals and proactively reminds the user about them.[35] An application like Spotify with its music recommendation algorithm would not meet this definition. Spotify does however include some of these features. It analyses data from the user (what songs they listen to and how often) and gives tailored feedback and advice.[36]

E-coaches can be understood as devices which have the aim of helping a person reach their goals in a way that is analogous to goal-oriented human coaching. To function properly an e coach usually constitutes of functions like the one's described by Kamphorst.

[32] Ibid.

[33] A nudge is an action or implementation with the goal of changing people's behaviour, usually towards a desired goal like choosing a healthy meal option, through bypassing the one's rationality. Nudges include things like healthier foods being placed at eye-level in supermarkets.

[34] Peter John et al, *Nudge, Nudge, Think, Think: Experimenting with Ways to Change Civic Behaviour,* (London: Bloomsbury Academic, 2013).

[35] Philips. "About Health Suite" on https://www.philips.com.au/healthcare/innovation/about-health-suite. Visited on 21st January 2017.

[36] Spotify. "Discover" on https://support.spotify.com/my-ms/using_spotify/discover_music/discover/. Visited on 21st January 2017.

What then are the ethically relevant aspects of e-coaching? There are many ethically laden aspects, including the persuasive influence of the e-coach on the user, the perceived expertise and the long-term influence on self-efficacy of the user. However, I think there is one aspect in particular that lays at the fundament of the e-coach and determines all its actions. This is the true locus of the moral action of an e-coach towards a user and should be investigated carefully. This aspect is the way in which the e-coach comes to know the user, namely profiling. I shall explain how e-coaches profile users and then explain why profiling is so ethically laden and why it is crucial to investigate.

3.a.i. User profiling in electronic coaching

Profiling is also used in traditional human to human coaching. Leadership coaches for example often use the Five Factor Model as a taxonomy of personality types.[37] Sports coaches also sometimes use a method called performance profiling. Performance profiling usually involves having athletes rate themselves on a variety of relevant skills and attributes. Butterworth et al describe a movement from only considering athletes' individual ratings to including the rating of other people concerning the evaluated athlete, as well as the ratings of peers (consisting also of self-rating and rating by others).[38] If coaches do not use explicit methods of profiling, one could still wonder if profiling could be said to happen in every mind, subconsciously and automatically. People have a tendency for very fast in-group/ out-group thinking. Even the presence of very trivial symbolic markers like colour of skin or clothing can initiate the formation of groups and in-group favouritism among people.[39] These tendencies seem to indicate a deep-rooted disposition for something that could be described as

[37] Iain McCormick and Giles St. J. Burch, "Personality-focused coaching for leadership development," *Consulting Psychology Journal: Practice and Research* 80 (2008): 267-278.

[38] Andrew Butterworth et al, "Performance profiling in sports coaching: A review," *International Journal of Performance Analysis in Sport* 13 (2013): 572-593.

[39] Charles Efferson et al., "The Coevolution of Cultural Groups and Ingroup Favoritism," *Science* 321 (2008): 1844–1849.

profiling. In the case of electronic coaching however, systematic profiling is definitely an inevitable part of personalizing the application to a user.

Personalization makes apps much more effective for the purpose changing the user's behaviour, as well as for increasing adherence. This is why Kamphorst includes it in his definition of e-coaching.[40] Personalization requires the creation of a profile for the user, to which the app can adjust its behaviour change model. In the context of e-coaching I shall refer to profiling as *user-profiling.*

Kanoje et al define user profiling as "the process of identifying the data about a user interest domain. This information can be used by the system to understand more about user and this knowledge can be further used for enhancing the retrieval for providing satisfaction to the user. User profiling has two important aspects as efficiently knowing user and based on those recommending items of his interest."[41] Kanoje et al use this definition in a paper about recommended systems. In the context of e-coaching we would understand "recommending items of his interest" as advising the user to do things which are in the interest of him reaching the goal he has set for himself.

Personalizing algorithms create user-profiles by attaining feedback from a user. There are two ways of obtaining feedback, implicitly and explicitly. Implicit feedback is gained by monitoring the behaviour of a user and extrapolating information from it. An application may for example keep track of how long a user sleeps in order to know what bedtime to recommend to them. This is an implicit way to obtain information from the user. Information can also be obtained explicitly by asking the user directly. Explicit feedback can be incorporated at many different stages of application creation and use. Users can be consulted

[40] Kamphorst, "E-Coaching Systems," (2017).
[41] Sumitkumar Kanoje et al., "User Profiling Trends, Techniques and Applications," *International Journal of Advance Foundation and Research in Computer* 1 (2014). 2348-4853.

at the design phase of the application or asked to choose an algorithm which they prefer. (i. e. Do you want us to recommend programs that your friends like? Do you want us to take into account your weekly schedule?) They can also be asked for ratings on performance of the app. In e-coaching application, explicit information and feedback has a particularly important role because the aim is to change behaviour to what the user explicitly plans. Adjusting the advice to a user's current behaviour only would be useless if the whole point is to change that behaviour. [42]

A user profile is often created using data from one specific user, as well as a model based on data from multiple users. The algorithm of the eating e-coach *Think Slim* works in the following way: Users are asked to use the app prior to eating something and they are also surveyed at semi-random intervals (with approximately 2 hour intervals). The user rates their emotional state and their feelings regarding eating and answers questions regarding their location and activity. The user is also asked to report what they eat. The app works like a logbook that spots patterns in eating habits and emotional states, times and locations. The app constructs rules of the form: being at home + negative thoughts + lack of positive thoughts + stress = unhealthy eating. Such rules are calculated from the data of each individual user.[43]

Then the data of multiple users is tested for the triggering frequency of each rule. The participants are grouped into clusters where a set of rules describes 80% of their eating habits. Less occurring rules within a group are removed from the descriptors. These groups, each described by a set of rules based on correlations between eating habits and other recorded experiences, become profiles.[44]

[42] Michael D. Ekstrand and Martijn C. Willemsen, "Behaviorism is not enough: Better Recommendations through Listening to Users," *Proceedings of the 10th ACM Conference on Recommender Systems,* (2016).

[43] Jerry (G.) Spanakis et al, "Machine learning techniques in eating behavior e-coaching," *Personal and Ubiquitous Computing* 21 (2017): 645-659.

[44] Spanakis et al, "Machine learning techniques," (2017).

Individual users are matched to these profiles. But when the app notices further correlations in the behaviour of individuals, these correlations are added as “rules” to the individual profile of the user. The app then creates warnings in situations where the user is statistically likely to experience craving or eat unhealthy foods.[45] This is an example of how e-coaches use user-profiles to tailor their advice to users.

User-profiling is a very ethically laden aspect of e-coaches and requires careful ethical analysis. The ethical ladenness is because 1. User profiling is central to the different ways of influence exertion of the app on the user, 2. Because it is the e-coaches “understanding” of the user and can either respect or disrespect the user’s personality. The first source of ethical ladenness is obvious. The more tailored and variable an application is, the more important the user-profile is in determining the behaviour of the app. As algorithms become more and more self-learning, it is to be expected that they can become more tailored and therefore more of the behaviour can be determined by the user profile. The user-profile is therefore central to the user-specific actions of an electronic coach.

The second source of ethical ladenness concerns seeing the user as an individual. Individuality is a moral value in many senses. In one sense, individuality can be conceived of as the separateness of persons. One person is separate from others in body and mind, what is best for one person is separate from what is best for everyone. Rawls’s critique of utilitarianism relies on this notion of individuality. Rawls’s claim is that while thinking that we should strive towards as much “good” as possible seems like a rational theory, it cannot explain the intuition that we should also strive for what is right. The reason for this is that utilitarianism takes what is a rational principle for one man (strive for one’s own good) and applies it to society as a whole. The one deciding what is good for the assembled society-

[45] Spanakis et al, “Machine learning techniques,” (2017).

person, is an imaginary ideal observer with equal sympathy for everyone. This kind of conflation is illegitimate, Rawls argues. Individuality has moral significance.[46]

We can make an analogy with Rawls's argument against utilitarianism, to argue that statistical profiling can clash with this value of individuality. While having a statistic about what certain kinds of people are more likely to prefer can make it true that it is easier to estimate what one person prefers correctly. If blondes have statistically a very large chance at being bad at math, then it is justifiable to assume that a specific blonde is bad at math. However, it is not morally correct to do so. It's illegitimate, the goods of people cannot be aggregated in this way. Profiling through statistics clashes with the value of being perceived as an individual just like utilitarianism does.

In conclusion, because individuality is morally valuable and much of the morally laden actions of an e-coach depend on the user-profile, profiling in e-coaches is a moral act and deserves ethical analysis. The moral agency of an e-coach hangs together with the epistemic agency of the e-coach.

3.b Epistemic humility as a moral virtue for e-coaches

3.b.i. Why virtues?

I have described why user-profiling in e-coaches is a very ethically laden activity and why it deserves careful analysis. What I want to propose now is to analyse user-profiling from a virtue theoretical perspective. There is a lack of work in how virtue theories can contribute to understanding the ethics of user-profiling.

[46] Rawls, J. "5. Classical Utilitarianism" in *A Theory of Justice: Revised Edition*, (Cambridge: Harvard University Press, 1971 (1999)). 19-24.

Thinking about what makes a good human coach also lends itself very well to a virtue ethical description. Seth Davis, the American sports writer and broadcaster has interviewed and followed many excellent sports coaches and writes about what they share. A good coach is persistent. “Persistence is the strain of character one leans upon during those quiet moments when self-doubt creeps in.”[47] A good coach also needs to empathise with the coachees “taking time to acquire the critical information that will lead him to understand how the player’s mind, heart, and guts operate.”[48] A coach must also be authentic, whether they swear on the side-lines or watch in silence. “The guys on the team must have full confidence that their coach will remain authentic, particularly in those critical moments when the team must function as a single unit or suffer defeat.”[49] Good coaches, in sports and many other areas, are generally described as being marked by character traits. There is therefore reason to believe that a model for a good e-coach could be constructed from knowing the virtues of a good coach. A virtue ethical understanding of e-coaches is valuable for a broader understanding of the ethical issues that these kinds of devices stir up. Because of the fittingness at which virtues can be used to describe good coaches.

Apart from this, there is reason to think that virtue theory could be extremely fruitful in implementation into devices. Wallach and Allen have argued in their book *Moral Machines: Teaching robots right from wrong* that virtue ethics is a promising ethical model for implementation into machines. The authors argue that as a framework for machine implementation, virtue ethics is superior to deontology and utilitarianism. This is because it involves both bottom-up and top-down possibilities in computability. Virtuous behaviour is both a result of following rules (top-down) and of learning and practicing (bottom up). Both of these elements can be incorporated into self-learning machines. This two-sided

[47] Seth Davis, “Introduction” in *Getting to Us*, (New York: Penguin Press, 2018).
[48] Ibid.
[49] Ibid.

computability lends very well to self-corrective behaviour and adaptability to different contexts and users.[50]

Tonkens has argued against this idea by claiming that it would be (bluntly stated) hypocritical to implement virtue ethics into autonomous machines because creating machines that are moral agents would go against the central principles of virtue ethics. He argues that since virtue ethics requires that every moral agent deserves respect and moral treatment, it would be wrong to create morally autonomous machines for a purpose like care, war or entertainment without violating their rights a moral agents.[51] However, I think that Tonkens is equating moral agency (like that of autonomous machines) with moral patiency. Moral patiency is probably more demanding in the sense that it requires pain, alienation or hurt feelings, which machines (generally) do not have. I also think that the implementation of virtue ethics into machines can easily be justified by the fact that it is better than not implementing virtue ethics into machines.

Because of the urgency of epistemic matters in user-profiling by e-coaches, I have decided to start by describing a model of epistemic virtue, which is also a model of moral virtue, from a virtue perspective. This model is the framework of Zagzebski, which I shall now explain in detail.

3.b.ii. Virtue responsibilism

There are two ways of approaching virtue epistemology. The first method is reliabilism. Reliabilists see epistemic virtues as characteristics of agents which reliably produce knowledge or true beliefs. The other view is responsibilism. Responsibilists focus more on epistemic virtue as admirable characteristics of good epistemic agents like open-

[50] Wallach and Allen, *Moral machines*, (2009).

[51] Ryan Tonkens, "Out of character: on the creation of virtuous machines," *Ethics and Information Technology* 14 (2012): 137–149.

mindedness, intellectual humility or curiosity. Some virtue theorists, like Samuelson and Church, call for a combination of both for a complete understanding of epistemic virtuousness.[52] Others defend a single approach. Zagzebski defends a version of responsibilism with the claim that reliabilism cannot explain what is good about knowledge and is therefore not fit enough to be an epistemic virtue theory.[53]

Zagzebski is a defender of combined virtue resposnisbilism. Combined because she uses virtue epistemology to answer both traditional epistemological questions like "how does agent a know p?" and inquiry questions, like "what constitutes wisdom?" Responsibilim is the branch of virtue epistemology which concerns itself with virtues like open-mindedness and conscientiousness as opposed to reliabilist epistemology which is concerned with virtues as processes like seeing, hearing or reasoning that reliably produce true beliefs. The good of Zagzebski's responsibilist virtues are inseperable from moral good. She forges this connection by first arguing that reliabilism cannot explain why knowledge is better than mere true belief. This is because reliabilism takes the use of faculties that produce true belief reliably as being knowledge-creating. There is no reason why it should matter that the knowledge is created through a reliable process if it is already true. [54] Requiring truth and a reliable source of truth for knowledge cannot explain why knowledge is better than mere true belief. In fact, even if we argued that there is some value in a source of truth which is independent of the reliability and the value of the truth itself, we would still have a problem. The valuableness of a cause of knowledge does not transfer to add value to knowledge. This, Zagzebski calls *the value problem*. [55] To solve this problem, Zagzebski's proposes "that act S is credited to the agent only if the truth of belief B is credited to the agent. So if knowing B is something like truly

[52] Peter L. Samuelson and Ian M. Church, "When cognition turns vicious: Heuristics and biases in light of virtue epistemology," *Philosophical Psychology* 28 (2015): 1095-1113.
[53] Linda T. Zagzebski, "In search for the source of epistemic good," *Metaphilosophy* 34 (2003): 12-28.
[54] Mark Alfano, "Expanding the Situationist Challenge to Responsibilist Virtue Epistemology," *Philosophical Quarterly* 62 (2012): 223-249.
[55] Zagzebski, "In search for the source of epistemic good," (2003).

believing when the truth of belief B is credited to the agent, it follows that the agent gets moral credit for an act S based on belief B only if S knows B."[56] I think that this is a credible account of epistemic virtue by itself. It is also a good explanation for why epistemic arrogance is wrong in addition to the negative consequences it may cause.

Zagzebski's account shows that responsibilist virtues are an important components of any virtue epistemology, even if one also endorses reliabilist virtues as descriptions of specific kinds of cognitive capabilities.

Humans are notorious for overestimating their epistemic capabilities. [57] [58] [59] We have a tendency to form biases while also underestimating our liability to such biases. [60] The question I am focussing on is "when is this kind of epistemic failure vicious and worthy of moral blame?" I shall begin by reviewing the answer of Samuelson and Church to this question.

Samuelson and Church explain vicious epistemic conduct with reference to a dual process model of thinking, familiar from the work of Kahnemann.[61] The dual process model represents human thinking as two categories that cognitive behaviour tends to fall into. System 1, or type 1 thinking is fast, intuitive and often involuntary and unconscious. System 2, or type 2, thinking is reflective, slow and hypothetical and requires more working memory and conscious effort. Cognitive errors can happen on both levels of thinking, and it is their interaction which results in more systematic cognitive errors like biases. The authors argue

[56] Ibid.

[57] Shelly Chaiken et al, "Principles of persuasion," in *Social psychology: Handbook of basic principles*, ed. Edward T. Higgins and Arie W. Kruglanski, (New York: Guilford, 1996), 702–742.

[58] David Dunning et al.,"A new look at motivated inference: Are selfserving theories of success a product of motivational forces?," *Journal of Personality and Social Psychology* 69 (1995): 58–68.

[59] Keith E. Stanovich and Richard F. West, "Natural myside bias is independent of cognitive ability," *Thinking & Reasoning* 13 (2007): 225–247.

[60] Emily Pronin and Matthew B. Kugler, "Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot," *Journal of Experimental Social Psychology* 43 (2007): 565-578.

[61] Samuelson and Church, "When cognition turns vicious," (2015).

that it cannot be merely at System 1 level where epistemic viciousness occurs, as these thinking processes are automatic and unconsidered. Viciousness is something that happens when the two levels of thinking are not playing the appropriate role in belief formation.[62]

Stanovich's framework of thinking errors explains how this can happen in different ways, causing errors. Stanovich describes a first class of errors where the agent is a "cognitive miser" or fails to expend enough mental resources on correct thinking. This can take three forms:

1. Failing to decouple from one's own intuitive representation of reality and going with erroneous Type 1 thinking

2. Decoupling but failing to override type 1 thinking

3. Decoupling but failing to produce enough alternative comparisons

Another class of cognitive errors involves what Stanovich calls "mindware problems." These involve:

1. When there is a gap in learning which would be needed to carry out appropriate type 2 thinking
2. "Mindware contamination," or when there are learned rewards or punishments for certain types of thought which interfere with epistemic activity.[63]

Both the mindware problems and the "cognitive miser" problems are typical instances of epistemic vice.

[62] Ibid.

[63] Keith E. Stanovich, "Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?" In *Two Minds: Dual Processes and Beyond,* ed. Keith Frankish and Jonathan St. B. T. Evans, 55-88, (Oxford University Press, 2009).

Responsibilists are interested in the virtuous and unvirtuous characteristics of epistemic agents. Therefore, to determine exactly what this viciousness is within a responsibilist framework, it is useful to look at what kinds of mental processes accompany cognitive errors. Samuelson and Church discuss various mental processes that are used to eliminate or minimize bias. These include rule-based thinking, having motivation for accuracy and accountability and perspective taking. What all these seem to have in common is that they involve the thinker trying in some way to step outside the bubble of one's own mind. When thinking in a rule-based way one engages strategically with a particular way of thinking which is often learned from others. When motivated by accuracy and accountability one takes special care to ensure that others would also agree with one's conclusions. When taking different perspectives one is clearly trying to see an issue from viewpoints outside one's own immediate stance. Samuelson and Church reason that if epistemic other-regardingness or other-involvingness is epistemic virtue, then perhaps self-centeredness is epistemic vice.[64]

In many ways epistemic vice tends to be marked by self-centeredness. For example, being self-centered because of the rewarding feeling of being right and therefore seeking to reinforce one's own standing beliefs is an example of a mindware contamination problem. The kind of self-centeredness where one lacks the motivation to put oneself in another's shoes also often results in epistemic vice of the "cognitive miser"-kind.[65] Samuelson and Church also add that according to studies in psychology, de-biasing often requires motivation. "In this way, it reflects the neo-Aristotelian (responsibilist) notion of epistemic virtue: that it must be consciously practiced to overcome our more arrogant cognitive tendencies and avoid the possibility of being too diffident to others by giving in too easily or not evaluating the other's position rigorously."[66]

[64] Samuelson and Church, "When cognition turns vicious," (2015).
[65] Ibid.
[66] Ibid.

Samuelson and Church sketch an image of epistemic humility as a responsibilist virtue. Intellectual arrogance is characterised by self-centered inclinations in thinking and the lack of motivation to expend cognitive effort on minimizing it. Intellectual arrogance is an epistemic vice. Within the Framework of neo-Aristotelians like Zagzebski, epistemic vices are also moral vices and therefore epistemic arrogance would be morally wrong. Zagzebski also describes epistemic humility as a crucial epistemic virtue.

3.b.iii. Epistemic virtues in user profiling

The virtue of epistemic humility is important for e-coaches because it is very important in profiling. I would argue that epistemic arrogance is the key vice that makes deplorable instances of profiling deplorable. This means that the virtue of epistemic humility is extremely important in profiling, and therefore an essential virtue for e-coaches. The intuitive deplorability of some types of profiling is also indicated by the amount of literature criticizing them. Most of the discussion on the ethics of profiling falls under one of these categories:

- The profiling of people for decision making purposes: I. e. Ethics of criminal or genetic profiling and their use in the justice or medical system.
- The profiling of users to provide a service: i. e. the profiling of users on media-platforms for the purpose of targeted content or advertisement.

One issue often discussed in the debate about racial profiling also is the epistemic validity of different types of profiling.[67] [68] [69] If we put people in boxes then surely they should be the correct boxes. Especially in criminal profiling the epistemic is highly ethical because of the consequences of decisions based on profiling for people.

[67] Michael Boylan, "Ethical Profiling," *The Journal of Ethics* 15 (2011): 131-145.

[68] Frej K. Thomsen, "The Art of the Unseen: Three challenges for Racial Profiling," *The Journal of Ethics* 15 (2011): 89-117.

[69] Kahn, Jonathan, "Getting the Numbers Right: Statistical Mischief and Racial Profiling in Heart Failure Research," *Perspectives in Biology and Medicine* 46 (2003): 473-483.

Another set of problems concerns the negative consequences of putting people in boxes for any reason. Many scholars on the topic of racial profiling argue that even if profiling, in the questionable form practiced collectively by American police, was a reliable way to catch criminals, the negative costs attached to it are too high to permit. If people are treated as “likely criminals” because of their ethnicity, then a demeaning message is sent to members of that ethnicity, which divides society.[70] Despite the negative and stereotype-enhancing consequences of profiling by things like race, many scholars advocate the use of profiling by other means. Boylan for example argues that in retroactive profiling (so when the crime has been committed and suspect is being searched) we should profile criminals through trying to infer their motivations and contact with the victim.[71] This kind of profiling does not have these negative societal consequences.

There is however an intuition that there is something wrong with putting people into boxes based on certain characteristics regardless of negative consequences. There is an unopen-mindedness in assuming things about people based on a category they belong in. Imagine that it is statistically evident and well known that people who wear leopard print are less likely to be intelligent. If Lisa is an extremely intelligent woman who likes to wear leopard print, she will often be treated in a way that annoys her by new colleagues, people in bars and potential employers. She is profiled as a leopard-wearing and unintelligent person when in fact she is only leopard-wearing. Therefore, even if it is reasonable to assume that Lisa is unintelligent and it has no great negative consequences (imagine that people generally treat her the same even though they perceive her as unintelligent) it is still morally bothersome. This intuition exists, I think, because we have a negative moral judgement about epistemic arrogance.

[70] Deborah Hellman, “Racial profiling and the meaning of racial categories,” *Contemporary debates in applied ethics* 2, ed. A. I. Cohen and C. H. Wellman, (Malden: Wiley-Blackwell, 2014), 237.
[71] Michael Boylan, “Ethical Profiling,” (2011).

The epistemic arrogance in deplorable cases of profiling is accompanied by the problematic thinking processes described by Samuelson and Church. Here is an example: an area in a city has a high incidence of crime and also a high incidence of residents belonging to a certain ethnic group. Because of this fact, a police officer comes to believe that people belonging to that ethnic group are more likely to be criminals and therefore targets people with that ethnicity. Jumping from “lots of crime here + lots of ethnicity x here” to “ethnicity x = crime” is in most cases an instance of type 1 thinking. This intuitive heuristic, while probably useful in a life of hunting deer and escaping tigers, is not appropriate for policework. If the police officer goes with this type 1 thinking, without thinking to engage in type 2 thinking or perspective taking, then he is a cognitive miser and thus epistemically arrogant. The same goes for a police officer who refuses to engage in type 2 thinking because of it feeling better to think that her intuitions are correct. She is epistemically arrogant.

Type 1 thinking is known to play a large part in biases. Reliance on heuristics instead of reflective reasoning often causes predictable biases. An example is the way that if people are asked “are there more words where the first letter is K, or words where the third letter is K” people often go with the first option. This is because it is much easier to think of words which begin with a particular letter than words which have a particular third letter.[72] The same goes for judging people. Kahneman describes an experiment where participants are told about Linda who majored in philosophy and was deeply concerned about issues of social justice during her studies. Participants are asked to evaluate the likelihood of Linda now being “a teacher in a nursery school“, “working in a bookstore and taking yoga classes” as well as other alternatives. There were also the options “Linda is a bank teller” and “Linda is a bank teller active in the feminist movement.” Logically, it should be clear that since every “bank

[72] Daniel Kahneman, “9. Answering an easier question” in *Thinking, Fast and Slow,* (London: Penguin Books, 2011).

teller active in the feminist movement" is also a "bank teller" that it is more likely that Linda is a bank teller than that she is a "bank teller active in the feminist movement." However, people still ranked the feminist bank teller as more likely than the bank teller, based on their idea of Linda! This is exactly the kind of thinking error, where type 1 thinking is not corrected, that results in bias and flawed profiling.[73]

Epistemic arrogance is described slightly differently by Aleksandra Tanesini. She describes two components of epistemic arrogance. The first in haughtiness, which is an interpersonal behaviour that tries to silence others when in discussion. Tanesini writes "They include talking over other people, interrupting them, putting them down in public, ignoring or rejecting without reasons what they may have said, and conveying to one's audience the impression that one thinks of oneself as cleverer, smarter or more quick witted than them. Arrogance of this kind is often identified with a feeling of superiority over others."[74] A good example of this would be UK Prime Minister David Cameron's "Calm down, dear" said in a debate to a shadow Chief Secretary to the Treasury, Angela Eagle who criticized his policies in 2011. The second component, arrogance, is when a person generally values their own thinking too much. This often involves overestimating one's abilities and avoiding committing oneself as accountable for one's assertions. I think that this arrogance and consersational haughtiness, as described by Tanesini, can also be explained in terms of deep intellectual self-centeredness that results in an unwillingness to expend cognitive resources or engage in perspective taking.

What then is epistemically humble profiling? Surely it is profiling which does the opposite as the epistemically arrogant practices described above. Epistemically humble profiling is

[73] Kahneman, "15. Linda: Less is more" in *Thinking, Fast and Slow*. (2011).
[74] Alessandra Tanesini, ""Calm Down Dear": Intellectual Arrogance, Silencing and Ignorance," *Aristotelian Society: Supplementary* 90 (2016).

conducted by epistemically humble agents who do not have self-centered tendencies in thinking. Such an epistemic agent is:

- Willing to expend cognitive resources on correcting type 1 thinking
- Is not cognitively lazy
- Is not cognitively self-centered
- Engages in perspective taking
- Does not engage in epistemically arrogant behaviour like silencing others

Because we are trying to understand virtuous profiling in e-coaches it will be necessary to translate these characteristics to a system view. In order to do this we must think about how an algorithm can emulate the profiling of an epistemically virtuous human being. It is helpful here to think about what kinds of profiling behaviours are indicative of epistemic humility.

There are different views about what kind of profiling manifests these virtues. Boylan, for instance, argues that using intentional attributions is more ethical and reliable than using statistical categories. Imagine that we are trying to solve a burglary of a jewellery shop and we are looking for suspects. Judging one characteristic of a person, such as how likely they are to have committed this crime based on a characteristic which is unrelated to criminality, like the colour of their skin is clearly unethical even if there is a correlation. But imagine that we are trying to solve the murder of a rich man. In this case we might judge the likelihood of someone having murdered the man based on whether they would have profited from his death for example through inheritance. Having a certain skin tone and being familiarly related to the man in question are in themselves not indicators of criminal inclinations. However, the first is more unethical than the second because being able to profit from the man’s death is a motive for murder. This kind of profiling, by victim familiarity and possible motivations is much more effective than type-profiling, according to Boylan. And it avoids the ethical problems associated with profiling people based on superficial characteristics.[75]

An exaggeration of Boylan's view could be that statistics are never enough to inform profiling and some plausible causal explanation is always needed. But this becomes unintuitive in the face of very, very strong statistical correlations. Imagine that it was really true that 99.9% of people with a certain skin colour actually were murderers. Imagine that this was proven again and again by history and statistics. Should we then say that it is epistemically arrogant to profile a person with that skin colour as a murderer? Especially with such high stakes as letting a murderer loose, I think it would be foolish to insist on equal treatment for the murderer-skinned individuals. There must be a threshold of statistical power that makes a prediction epistemically viable. And if a prediction is epistemically viable, a profile based on it is not necessarily indicative of epistemic arrogance.

Another take on proper profiling is that of Corlett, who claims that the problem is that the categories used in racial profiling are too broad. Targeting a category of people, like "all people with skin colour x" is not only ineffective but also bothersome for many people. Using narrower and better researched categories that do not impose a burden on a large amount of people would be justified given that these categories are predictive.[76]

The examples named above are examples of supposedly more reliable and ethical ways to profile people. Implementing these into algorithms would be implementing the profiling behaviour of epistemically humble agents into the algorithms. However, we must also implement certain reflexive tendencies that mirror the characteristics of epistemically humble people. Such reflexive tendencies include the tendency to reflect and re-think judgements. This kind of reflexivity could be implemented into machines with the kind of

[75] Boylan, "Ethical Profiling," (2011).
[76] J. Angelo Corlett, "Profiling Color," *The Journal of Ethics* 15 (2011): 21-32.

framework that Wallach and Allen propose in *Moral Machines*. The artificial agent should be programmed "top down" to act in accordance with certain moral principles. Examples would be the categorical imperative or Asimov's three laws of robotics. The artificial agent should however also be fitted with a learning capability and it should receive feedback about its actions, so that the rules it acts in accordance with can be modified through experience. In this way the agent would begin with a crude conception of morality and learn to fine-tune its moral code to different situations and nuances.[77] This flexibility in behaviour code and the ability to retain feedback and learn would constitute an artificial counterpart to human epistemic humility. In the case of profiling it would therefore be important that an algorithm receives feedback on the kinds of profiling actions it carries out. It would have to learn from humans, or other artificial agents further in their moral development, much like a small child. A system view of epistemic humility would include these learning features and display increasingly unarrogant profiling behaviour.

In order to understand the profiling practices of electronic coaches in a virtue ethical light, it will be important to understand how epistemic virtues, including responsibilist virtues, function in user-profiling. I suggest that we hold the same epistemic virtues as prescriptive for e-coaches that we hold prescriptive for human beings. E-coaches must be fitted with algorithms that:

- Act like they are willing to expend cognitive resources on correcting type 1 thinking
- Are not cognitively lazy
- Are not cognitively self-centered
- Engage in perspective taking
- Do not engage in epistemically arrogant behaviour like silencing others

[77] Wallach and Allen, *Moral machines*, (2009).

Even though we might not be able to talk about a machine as “getting more gratification out of easy thinking” or as “putting in lots of effort” we can still describe ways in which machines can use reasoning that is “lazy” or “self-centered.” The specifics of what these virtues consist of depends on the kind of e-coach agent we are evaluating. In the next section I shall describe how epistemic humility and arrogance are manifested in the different kinds of e-coach-agents.

3.c. Epistemic responsibilist virtue in different modes of agency

I have described some different modes of agency that an e-coach can have and why they are all possible in e-coaching. I want to show that e-coaches as parts of distributed or extended agents, as well as independent machine agents, can be evaluated by ascribing virtues and vices to them. First I will describe the kind of existence that virtues can have in these modes of agency. Then I shall treat some counteraguments against my views.

3.c.i. Epistemic humility in extended agents

The extended mode of agency is when a human user’s agency extends to include the e-coach. In this case, I would argue that their virtues (moral and epistemic) can also extend to the device. An adherent of the extended mind thesis might describe the device as a part of the user’s extended mind. And a non-adherent might still consider it a part of the *extended person.*[78] In both cases, virtues and vices can also be said to extend to the device. If for example an alcoholic manages to stay off alcohol with the help of an e-coach, then we can say that the virtue resides in the user as extended to her device. According to Howell’s view, it would make sense to say that when a person is operating together with an algorithm, the virtues and vices reside in the *user-e-coach-system.*

[78] Robert J. Howell, “Extended virtues and the boundaries of persons,” *Journal of the American Philosophical Association* 2 (2016): 146-163.

In fact, it makes a lot more sense to think about good characteristics that one acquires through the use of an e-coach as extended or distributed virtues than as any other kind of virtue. This is because the other ways to see virtues brought on by e-coaches run into difficulties. I will demonstrate these difficulties with an example. Imagine that an alcoholic, Kim, manages to give up alcohol using an e-coaching system. The system asks her to set goals, which she does, and sends highly persuasive daily encouragement according to her stress levels measured by a wearable wristband. When the device senses stress, it may send words of encouragement, offer a chat with a fellow user or a guided relaxation exercise. When Kim is using the app, she stays motivated to stay off alcohol. However, if Kim were to stop using the app, she would become depressed, aggressive and likely to relapse. I would say that there are three ways of understanding the newly formed virtue:

1. Kim's strength against her addiction is mediated by the e-coach
2. Kim possesses a situation specific virtue, "strength-against-her-addiction-while-using-e-coaching"[79]
3. Kim's strength is partially constituted by (and thus extended to) the e-coach.

If Kim's admirable strength is merely mediated by the e-coach, then we must accept that Kim has the virtue of strength against her addiction and the vice of weakness against her addiction at the same time. Or rather we might say that she has no virtue or vice at all, since the character traits do not seem to be consistent. This is not such a counterintuitive conclusion. But what is missed here is the role of the e-coach in creating a stable virtue, bringing about behaviour change (which is what these devices are designed to do). If Kim had never bought the e-coach, she would never have been able to develop this virtuous lifestyle. Why should we

[79] Philosophers have defended virtue ethics from the situationist critique by arguing that there are no global character traits but rather situation-specific ones, like "sailing-in-rough-weather-with-friends-courage and office-party-temperance." Jonathan Webber, "Virtue, Character and Situation," *Journal of Moral Philosophy* 3 (2006): 193-213.

say that Kim has this virtue in a counterfactual future where Kim relapses into alcohol use and aggression? Granting virtues the possibility of existing as potential but nonmanifesting traits distances them too much from the actual practice of virtuous acts. This, I think, would be doing a great injustice to the whole point of virtue ethics.

Is it perhaps better to say that the e-coach is a part of the circumstances of the user, rather than a mediator and that the virtues of the user are specific to the situation that the e-coach creates? Kim might become more determined to fight her alcoholism when she comes into a situation where she is sent many push notifications, because she possesses the virtue of *push-notification-enhanced-determination*.

However, this option is plagued by the same problem as the first. How shall we judge her when she loses her e-coach and relapses into alcoholism? Does she still possesses the virtue, but is not in the right situation to exhibit it? Virtues are not flukes but stable character traits. Thus, presumably, if character traits are context specific, they also need to exist at times when they are not manifested. In this way, a person can have a vice of laziness without an e-coach but a virtue of perseverance with an e-coach. How should we judge such a character? At this point the virtue ethicist must question how useful the concept of virtue is at all as a normative framework by which to judge character. By that logic we are all killers and should perhaps be judged as such. Because surely there is some horrible situation involving years of abuse which would make us so.

The defender of situation specific virtues might claim that we should not judge people by potential virtues or vices but only actual ones that are present or are being developed at the moment or have been in the past. But this claim would strike me as out of line with virtue ethics. Surely to have a virtue is to be judgeable by it. We would not judge a person, virtue ethically, as evil only at the instances that she is killing or stealing. We would judge her as a bad person also when we see her choosing bananas at the supermarket the next day. Virtues and vices that a person has are by definition those to judge her character by.

If we conceive of virtue as extended or distributed, then the virtue commences in whatever system the agent operates, including parts of the environment. The addicted and aggressive Kim has a vice, as long as she does not have her e-coach. A person can switch from having one composition of virtues to having another in a short time because their environment changes, and therefore the constitution of her person changes. This explains the supposedly uncharacteristic behaviour of people during experiments like the Milgram experiment.[80]

All in all, extended virtue explains the changes that environmental factors can make to a person's character. We do not need to postulate traits which we do not see in order to evaluate someone's character. This I add to support Howell's argument. Howell argues for extended virtues in a way analogous to the extended mind argument of Clark and Chalmers. He also demonstrates the ability of the extended virtue theorist to respond to the situationist challenges to virtue ethics. Extended or distributed virtues have a theoretical simplicity and ability to account for changes in a person's character by giving the virtue a larger space of existence.

I have just explained the advantage of extending or distributing virtue, over a traditional view. Let us briefly return to the idea of extended virtues as distinct from distributed virtues. An e-coach can be seen as constituting an extended agent if it is very transparent to the user and integrated in their functioning. When we consider the epistemic

[80] The Milgram Experiments conducted by Yale University psychologist Stanley Milgram. People across many ages and occupations were asked to give electric shocks to a "learner" instructed by an experimenter. The electric shocks were false, and the learner acted as if they were hurt or wanted to stop the experiment. Most participants still obeyed the experimenter and gave them shocks of a deadly voltage. The experiment shows the tendency of people to obey authority even when the commands are unethical.
Howell, R. J. "Extended Virtues and the Boundaries of Persons," (2016).

virtues of the agent, in the context of e-coaching, we are primarily concerned with self-knowledge. The e-coach of course performs epistemic acts to get to know the user, and in turn the user performs epistemic acts to get to know herself and her e-coach.

If virtues can also be understood at a system level, then there’s no reason why we could not attribute them to extended agents. In fact, attributing them to extended agents is probably a lot more comfortable for critics of a system level approach. This is because with a human being as a part of the agent, things like emotions are present in the agent, whereas in a fully artificial agents they are not.

The duty of the extended person or the e-coach-user system to inquire into the needs and desires of the human user can be explained as a duty towards the patient which is the extended user or system itself. This does not require a virtue ethical approach. Surely any agent has the responsibility to acquire knowledge in cases where this knowledge means they can behave more morally. This is sometimes called the duty to fulfil the epistemic condition.[81] If we translate this to a virtue understanding: you need to be epistemically virtuous in order to be morally virtuous in a broader sense too. Holly Smith has argued that there are benighting acts, where an agents omits from inquiring into something, and then unwitting acts, in which a person acts wrongly because they did not know something which would have made them act rightly. If the benighting act causes an unwitting act, then the moral harm in benighting acts is derived from the moral harm of the resultant unwitting act.[82] However, one may ask here whether it is fair to speak of only the human user as the patient, if we accept that the agent is the human user and the e-coach combined.

If we accept virtues as duties, then reliabilist virtue epistemology allows us to conceive of the epistemic duties of the e-coach-user as duties towards itself. The duty of the e-

[81] Holly M. Smith, “The Subjective Moral Duty to Inform Oneself before Acting,” *Ethics* 125 (2014): 11-38.
[82] Ibid.

coach-user to be epistemically virtuous can also be explained in terms of the epistemic virtue of seeking to know oneself. The responsibilist Lorraine Code has argued that it is essential for a virtuous knower to actively seek to know not just the outside world, but also themselves. This is because the self is where cognitive errors and biases occur. Actively trying to know oneself is epistemically humble and prevents the entry of bias and other errors into one's epistemic activity.[83] A virtuous knower is epistemically responsible, tries to reason without being swayed too much by emotions or desires and is also epistemically prudent, having "a sense of one's limitations."[84] Schirmer Dos Santos has expanded on this idea and argued that although self-knowledge by introspection is often not possible, it is still epistemically virtuous to attempt to gain self-knowledge.[85]

Therefore, an extended agent, or system should also function in a way that exhibits virtues of self-knowledge. The system should collect proper information, functions of the human and functions of the device should make sure of this. An e-coach should not process data incorrectly and the user ensure correct data input by not lying about their weight or such. In cases of an extremely integrated e-coach, rather than seeing the virtuousness of the e-coach we should look at the virtuousness of the e-coach and user system.

For such an agent, which consist of a human and an integrated e-coach, epistemic humility would include exhibiting epistemically virtuous behaviours such as:

- Having good self-knowledge (understanding what the e-coach does and the e-coach having a correct profile of the user)
- Not jumping to conclusions about one's own health too quickly

[83] Lorraine Code, "Toward a 'Responsibilist' Epistemology," *Philosophy and Phenomenological Research* 45 (1984): 29-50.
[84] Code "Toward a 'Responsibilist' Epistemology," (1984).
[85] César Schirmer Dos Santos, "Self-Knowledge and Epistemic Virtues: Between Reliabilism and Responsibilism," *Veritas – Revista de Filosofia da Pucrs* 60 (2015): 579-593.

- Not filling in data haphazardly, resulting in wrong information about the self.
- Being conscious of the limitations of the device and one's own brain

I think that if agency can be extended, there is no reason why virtues couldn't. As we have seen, there are also independent reasons for favouring an extended view of virtues over a more traditional view which limits virtues and vices to exist within the skin of a person.

3.c.ii. Epistemic humility in distributed agents

Having explored the constitution of virtue in extended agents and arguing for its possibility, let us consider what it would mean for a virtue like epistemic humility to exist in a distributed agent. A distributed agent, in the Floridi's sense, is a supra-agent consisting of a human and her e-coach. There need not be complete integration of the e-coach into the functioning of the user. If the two, together, constitute a virtue, then this virtue is a trait of this supra-agent (human + e-coach). The distinctive thing about distributed agency is that it allows for what Paul Smart in a recent paper has termed *Mandevillian intelligence*. This is when cognitive vice of individuals contributes to collective epistemic virtues. The effect has been observed as an emergent phenomenon. Behaviours like quick judgement can mean that a group has more exploratory capacity and comes collectively to know truths faster. Smart has described this phenomenon in the context of groups of many people. [86] Something like this is imaginable in e-coaching too.

3.c.iii. Epistemic humility in machine agents

If we take the perspective where the e-coach is a full moral agent acting on the patient, we see that this view mirrors a more familiar and traditional relationship like the one between a coach and a coachee. From this perspective, being epistemically humble would mean having a profiling algorithm which takes into account the fact that it might categorize a person

[86] Paul R. Smart, "Mandevillian Intelligence," *Synthese* 1 (forthcoming).

wrongly and takes measures to prevent this, such as self-correcting and asking the user for input. Its behaviour should be careful and actively self-improving, much like a virtuous human epistemic agent.

3.c.iv. Counterargument: Responsibilist virtues cannot be realized in machines

Zagzebski, for one, would likely not agree with my attribution of responsibilist virtues to distributed agents and certainly not to machine agents. In *Virtues of the Mind* she criticises reliabilism for making epistemic virtue of human being seem something like the workings of a good computer.[87] One could ask whether the idea of responsibilist virtues makes any sense for non-human agents. One might also wonder whether we should reserve responsibilism for humans and evaluate the epistemic virtues of machines through reliabilism.

However, I think that Zagzebski's critique of reliabilism, namely that it does not explain what is good about knowledge as opposed to true belief, also applies for machine agents. Reliabilism still equates the goodness of truth with its reliability. As Zagzebski argues, this is like equating the goodness of espresso with the reliability of the espresso maker.[88] If we are to evaluate devices or device-human systems as good or bad, then why should we settle for reliabilism just because they are non-humans? Then we would be assuming that there is no good in proper epistemic activity other than its reliability when it comes to machines. I contest this view because it is anthropocentric and untenable if one grants devices a role of agency. If we are to evaluate machines morally, then they should also be evaluated for their epistemic behaviour rather than just the correctness of the results they yield.

[87] Linda T. Zagzebski, *Virtues of the Mind: An Inquiry Into the Nature of Virtue and the Ethical Foundations of Knowledge*. (Cambridge: Cambridge University Press, 1996).
[88] Zagzebski, "In search for the source of epistemic good," (2003).

4. Moral deplorability of epistemic arrogance depends on power

So far I have argued that e-coaches can be seen as moral agents. They can be moral extended, distributed or machine agents. I have also argued that the profiling practices of e-coaches, in every mode of agency, should conform to responsibilist epistemic virtues. One could ask at this point: how strong is this obligation? How bad is it if an e-coach is being epistemically arrogant? Is this something that we could call deplorable? In this section I want to discuss the strength of the moral obligation for e-coaches to be epistemically humble in profiling. I will begin by asking the same questions with regard to people profiling other people. For example, is it morally deplorable if you analysing cases of epistemically arrogant profiling which are clearly morally deplorable if you jump to conclusions too quickly about what flowers your mother would like for mother's day? I will look into deplorable and permissible cases of epistemically arrogant profiling and conclude that epistemic arrogance in profiling becomes more deplorable when it is about a person over which the profiler has some kind of power over the person profiled.

A paradigmatic case of deplorable epistemically arrogant profiling is biased racial profiling. An example would be a judge giving convicts of one ethnicity longer sentences due to believing either explicitly or implicitly that they are more likely to commit crimes in the future. Another example would be the police targeting members of one ethnicity disproportionally for random searches and traffic stops. Epistemically arrogant profiling is not always morally deplorable. Imagine a barista looking at a customer and getting out the soy milk because the customer looks like someone who would order a soy-latte. Imagine being in unfamiliar company and making a rude joke because the people seem like the kinds of people who would appreciate it. The important difference seems to be the amount of power that the profiling judgement contains.

In the case of someone being a cognitive miser or self-centred about what kind of flowers their mother prefers for mother's day, it is still epistemic arrogance. But I suggest that the moral deplorability of epistemic arrogance seems to hang together with the amount of power that the agent has over others and the relevance of the arrogance in exercising this power. Being epistemically arrogant about what flower your mother wants might not be terribly wrong. But it is wrong to be epistemically arrogant about whether a person, whom you can stop and detain if you wish, is likely to be a dangerous criminal or not.

This also goes for e-coaching. While e-coaches need to be epistemically humble, the moral obligation to do so becomes more demanding when they have a lot of power over the user. The next step is understanding when an e-coach has the kind of power over a user that would make the responsibility of epistemic humility more pressing. The next question to ask is: What does it mean for an e-coach to have power over an individual?

4 a. What is it for an e-coach to have power over its user?

By power I mean roughly influence. Perhaps we can say that an e-coach has power over its user when it manages to change the user's behaviour. But the applicability of this is a criteria that is hard to judge because behaviour change also depends on the user, not just the e-coach. It is also hard to understand as power of the e-coach agent over the user since changing behaviour is the goal of e-coaching, and this merely means that the device is working well.

The influence of e-coaches on the autonomy of the user has been discussed by philosophers. Kamphorst and Kalis have for instance connected the way e-coaches influence the options of users to how well an e-coach respects a user's autonomy.[89] Is respect for autonomy perhaps a good description for how much power an e-coach has over a user? Could an e-coach that can influences or disrespects autonomy a lot could be considered more powerful than one that does not?

I think that there is a better definition of power than autonomy. This is because in many cases users employ e-coaches because they feel they lack autonomy. Therefore determining whether an e-coach is influencing or leaving the user's autonomy untouched is very difficult to answer. Another reason is that autonomy, at least how it is often understood, is quite a demanding capacity and I think that an e-coach could be seen as having power over a person in many ways that do not influence her autonomy but nevertheless are morally questionable. Dworkin defines autonomy as a second order capacity to self-regulate. When Odysseus commands his men to tie him to the mast of the ship so that he is not tempted by sirens. He restricts his liberty, but not his autonomy. For temptation, like the sirens, appeal to first order desire, not something Odysseus would identify as ultimately desirable in the long term.[90] Autonomy, therefore, must be related to second-order desires. Dworkin writes "autonomy is conceived of as a second-order capacity of persons to reflect critically on their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher-order preferences and values."[91] I can imagine cases where an e-coach influences a person without influencing their second-order capacities in a way that should be morally relevant. Imagine that a friend of yours decides to use an e-coach to become healthier. Suddenly you start to notice him getting snobby and judgemental about your eating habits and reading out health facts from his app. You might be thinking "This is not who you really are?"

[89] Bart Kamphorst and Annemarie Kalis, "Why option generation matters for the design of autonomous e-coaching systems," *AI & Society* 30 (2015): 77-88.

[90] Gerald Dworkin, *The Theory and Practice of Autonomy*, (Cambridge: Cambridge University Press, 1988), 3-20.

[91] Dworkin, *The Theory and Practice of Autonomy*, (1988), 20.

Although their second order capacity to self-regulate might be stronger than ever, something has changed about them into a direction you don't like. The definition of morally relevant power in e-coaching should be able to include cases like this as instances of power of an e-coach over a user.

I suggest that how much morally relevant power an e-coach has depends on its capacity to influence the authenticity of the user. Authenticity is notoriously difficult to define. But I take it to mean, broadly, realness. It's about being who you truly are. I shall give two relevant definitions of authenticity in the next section.

5. Power of e-coach over user: influence on authenticity

I am considering two definitions of authenticity. The first is Altman's definition which sees authenticity as a match between one's identity and one's values, endeavours and projects. For this definition it is important that the person does not feel alienated from their endeavours and does not feel incapable of expressing their identity through their endeavours. The second definition I am considering is adopted from Taylor.

After explaining these definition, I shall show you that there are two types of inauthenticity, in accordance with these two definitions. Both types can occur as a result of an e-coach exerting influence on a user. I shall then explain how these occur in e-coaching and discuss how the likelihood of these things happening increases the moral demand for epistemic humility in profiling by e-coaches.

Ethics of authenticity are a species of moral theories which hold authenticity as an important value. Identity-matching and aesthetic preferences are seen as important and morally relevant.

I shall begin with Altman's view of authenticity. He defines it "the match between one's values and one's projects or activities."[92] Note that this does not require that these projects or actions are necessarily self-chosen. The important things is that one does not feel alienated from them. This means that authentic principles or values can also be "irrational" in the sense that they don't need to be in the best interest of that person. I can authentically value my drug-use, even if it is self-destructive. Varga and Guignon adopt a similar definition and clarify it again the concept of autonomy. While autonomy means holding to rational principles as ethical guides, always, authenticity means rather holding on to certain parts of one's identity and acting in accordance to them, even if it is against rational principles.[93] Authenticity is something more than autonomy, namely it contains a "language of personal resonance." Acting autonomously means acting according to rational principles, but acting authentically can sometimes mean acting against these, if doing so would conflict some integral part of one's identity.[94]

The second view of authenticity is Taylor's. Taylor endorses a thesis on what kind of authenticity is valuable and why. To understand this, we must first understand the criticism from which Taylor strives to rescue authenticity. Taylor's critique of the culture of authenticity is the following: if authenticity is understood as an intrinsic value that doesn't require reference to external values, then it becomes trivial and meaningless. As Taylor says, one cannot simply decide that wiggling his toes in warm sand is the most significant act in the world, without a special explanation, like the pleasurable feeling and connection with the earth, which appeal to some already recognized values. The importance of self-choice only makes sense if there is something valuable and significant to be chosen in the first place.

[92] Scott Altman, "Reinterpreting the right to an open future: From autonomy to authenticity," *Law and Philosophy* (2017): 1-22.
[93] Somogoy Varga and Charles Guignon, "Authenticity," *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (2017), Viewed on 9th May 2018.
[94] Ibid.

Identity must be defined against a background of things that matter or "horizons of significance."[95]

In fact, Taylor asserts, this is the only way we can ever form an identity and set of values. This is because our worldviews and morals develop *dialogically* rather than *monologically*. When we define who we are, we always do it in dialogue with existing views, our parents, our friends or public figures. All ways in which we express ourselves, like spoken and body language or art, are learned from others. There is no content to the world of significance, or its development, which we acquire completely on our own. Values, are therefore always bound to common horizons of significance.[96] That which a person chooses to think or do is therefore not valuable solely in virtue of it being their own. Importantly, we must pay due respect to all aspects of authenticity including the requirement of openness to horizons of significance and to self-definition in dialogue.[97]

Here we have two definitions of authenticity: Altman's definition of authentic it as a match between one's identity and one's values, projects and endeavours and Taylor's definition of authenticity the free choice of one's values against established horizons of significance. While Altman's authenticity refers to a state of non-alienation from one's project, Taylor's authenticity seems to be more concerned with the process of coming to hold authentic values.

Looking at these two ways of understanding authenticity, we can also define two types of inauthenticity, as their opposites.

[95] Charles Taylor, "IV: Inescapable Horizons" in *The Ethics of Authenticity*, (Cambridge Mass.: Harvard University Press, 1992), 31-42.
[96] Ibid. "IV: Inescapable Horizons," 31-42.
[97] Ibid. "VII: La Lotta Continua," 70-80.

6. Two types of inauthenticity in e-coaching

E-coaches are made to change people’s behaviour. I think that although “changing someone” always carries a risk of causing them to change to something inauthentic. There are however two ways in which an e-coach can cause inauthenticity, and these two ways correspond to the two kinds of authenticity described in the previous section.

Now we go back to the question: what kinds of inauthenticity are there that e-coaches can cause. I suggest there are two types. The first is the kind of inauthenticity is:

1. Where one does not act according to some part of their identity and feels alienated from one’s endeavours.

This in inauthenticity by the definition of Altman, there being a match between one’s identity and endeavours which must also be felt. This inautheticity has two components which are in themselves not sufficient to make someone or their action inauthentic. Not acting in accordance with some part of one’s identity need not be inauthentic. Not all actions need to conform to an established identity that a person has. Such actions are only inauthentic if one also feels alienated from them. Feeling alienated itself is not sufficient either. Imagine that an inauthentic person were given a pill that stops them from feeling alienated from what they are doing. It would be counterintuitive to say that taking this pill makes them more authentic. The alienation needs to have as its source the fact that one does not act according to their identity. Taking a pill to rid the feeling of alienation would not alter the causal connection between doing something inconsistent with one’s identity and feeling alienated due to it. Both are needed for inauthenticity.

The second kind of inauthenticity is:

2. When one thinks that they value something, but this is not a consequence of them cultivating that value independently. The appropriate process of coming to value something is not present.

This is inauthenticity in the sense of Taylor's authenticity. The process of selecting one's values and projects dialogically against a horizon of significance does not take place. Therefore, the resulting values or projects are not authentic. It is as if they are just picked off the shelf.

These kinds of inauthenticities, or fakenesses, are all related to the ethics of e-coaching in special ways and both have their source in the misprofiling of a user. I shall discuss these types of inauthenticity and how they can happen through profiling in e-coaching.

Before I do so, it is useful to look at De Vries's discussion of the possible influence of these technologies on the experience of identity. De Vries distinguishes two conceptions of identity, *idem* and *ipse*, adopted from Ricoeur. Idem is our identity as in the groups we belong in, like being a woman or a philosopher, and ipse is the part of the self which determines what one becomes. We define our identities using idem categorizations like "I am a woman, I am tall and I am a good fencer." The ipse is the uneasiness we feel when we are categorized only by an idem category. "I am not just any tall woman fencer, don't just put me in that box!" The ipse is the sense in which we are free to determine our own identity and behaviour. De Vries outlines three possible answers to whether the increase of ubiquitous personalization algorithms could change the way we experience our identity.[98]

[98] Katja De Vries, "Identity, profiling algorithms and a world of ambient intelligence," Ethics and Information Technology 12 (2010): 71–85.

Answer 1: The experience of our identity does not change at all. Stiegler argues convincingly that our identity has always been technologically mediated. The environment, whether it is a smart environment or not, has always been a place to outsource our memory of who we are.

Answer 2: Ubiquitous profiling will obliterate our sense of identity. She describes different reasons for why this might happen. When personalized algorithms guide us through our decisions, then we may lose our skill of navigating options that fit our identity. In De Vries's words it can "weaken human openness to what is called the idem/ipse entanglement or the conjunctive/disjunctive 'and'."[99] or the way in which we autonomously navigate and self-create the known parts of our identity. Additionally, it would be hard for us to extract new understandings of our idem identity from what algorithms tell us, because the way they profile us is often opaque. And even if it wasn't, it would often be incomprehensible or at least hard to identify with. For example, because algorithms profile us according to categories that work for whatever purpose the algorithm has it is likely that the categories are not meaningful enough for humans to take up as components of one's idem identity. Being "someone with two children and one who wakes up between 6 and 7, thus more likely to be motivated to go for a run at 8 pm than 6pm" is a category that an e-coach could use but unlikely to be a meaningful idem identity to someone. Also, when we are profiled by statistical inference and it is hard to relate to statistical categories and define one's identity thereby.[100]

Answer 3: The salience of one's identity as a collection of patterns and attributes profiled by a machine can lead to an uncanny confrontation with ipse. If Stiegler is right in stating that our identity has always been technologically mediated, then the fact that this mediation of identity becomes explicit in the different ways that personalized machines profile and treat us will make the strange construction of identity even more obvious in everyday life. This can lead to

[99] Ibid. 81.

[100] De Vries, "Identity, profiling algorithms…," (2010).

a general alienation from the self where it is obvious that identity is constructed by things and concepts outside oneself.[101]

Let us now return to how the two types of in authenticity relate to e-coaching. The first type of inauthenticity involves not acting in accordance with some part of your identity and feeling alienation as a result. This can happen when something goes wrong is this causal chain from intention to desired action. For example if you plan to do A, but then end up doing B instead. Kind of like you plan to go to the gym but then you go to McDonald's instead. Importantly, you would need to feel alienation from your decision to go to McDonald's, a disappointment, or sentiment of "why am I here, I don't want to be like this." Firstly, this is exactly the kind of thing that e-coaches are supposed to stop happening because they're supposed to engage you with doing what you plan. However, e-coaches can also cause thins kind of inauthenticity if they coach a person to do B instead of A. In this case, the e-coach has failed. De Vries rephrases van Bendegem "A growing use of profiling would automatically increase the amount of faulty categorizations."[102] Whether it is true or not that there would be more of this kind of inauthenticity if everyone used e-coaches, it is clear that an e-coach that is persuasive but misprofiles someone can lead to the endorsement of values that are not authentic because there is a mismatch between one's values and one's identification.

Inauthenticity type 2 is something that e-coaches can cause without failing. And this has to do with the dynamic nature of character. For actions and endeavours to be authentic, they should be based on values that are cultivated in accordance with one's identity. This process could be explained in terms of the idem/ipse distinction. The idem is all the ways in which one is the same as others, so the categories you belong in like for example being a

[101] Ibid.
[102] Jean Paul van Bendegem, "Neat algorithms in messy environments" In *Profiling and the identity of the European citizen*, ed. Mireille Hildebrandt and Serge Gutwirth, (New York: Springer, 2008), 100-103. Rephrased in De Vries, "Identity, profiling algorithms…," (2010).

woman or being a cat-owner. The ipse is this mysterious part that is something like free will. It’s the part where you determine what you are and what you become. So if we use this distinction, then we could say that for values to be authentic they must be pursued by the ipse against the background of idem. That’s how you get authentic values.

As the second and third answer given by de Vries make clear, it is possible that ubiquitous algorithmic profiling interferes with this process. Because firstly, if goal directed algorithms (that want to sell us things, or make us do certain things) profile us and direct us, then our idem identity which now consists of things like “being a woman” or being a cat owner, will consist of things like “being a likely buyer of ketchup on Saturdays”. She argues that the fact that we are being helped with making decisions about our actions all the time, according to the idem that algorithms assign us, means that we may lose the skill of navigating the options that are in accordance with our idem ourselves through our ipse.

Therefore, e-coaching presents a moral advantage in terms of eliminating the kind of inauthenticity that people have when they are controlled by their urges instead of their rational desires. However, e-coaching does include the risk of creating unauthentic behaviour, either because the user is profiled wrong, or because the use of e-coaching means that behaviour and values are not acquired in a way that leads to authentic behaviour and values. These kinds of inauthenticity however appear in different forms, depending on what kind of agent we are considering. I will discuss this in the next section.

7. Inauthenticity and power in different modes of e-coach agency

Let us think about how inauthenticity can be caused in the three modes of agency that e-coaches can have. If we understand how they can cause inauthenticity, then we can understand how much power they have and therefore also the moral deplorability of epistemic arrogance in their profiling practices.

7.a. Extended and distributed agents

The *profiler-profilee* relationship in extended agency develops over time as follows: When a user first begins using an e-coach, the device is presumably not yet integrated into the user's functioning, but it becomes so as the user uses it. Therefore, it would make sense to say that at the start of the e-coaching relationship the e-coach is either a machine agent, or the e-coach user system is a distributed agent, which profiles the human. As the use continues and the system becomes more integrated, it can be said that the extended agent no longer profiles only the human user, but itself, the extended agent. This is because as the device and the user become one. The user's decisions are co-influenced by the used device and therefore it is more correct to say that the *profilee* is the extended agent consisting of the human and the integrated e-coach.

The question is: how likely are e-coach agents to cause inauthenticity within themselves? If they are likely to do so, then we may judge the obligation they have to be epistemically humble as heavier. Let us begin by evaluating how likely they are to cause inauthenticity type 1. One would expect that the risk of type 1 inauthenticity occurring would decrease as the e-coach becomes more integrated. This is because integration involves the transparency of the e-coach to the user which enables to user to use the e-coach like a part of his or her self. If the user is able to use an e-coach to achieve her ends, it is less likely that she will end up, because of her e-coach, unwantingly carrying out acts that she does not want to carry out. Of course, this only holds given that the e-coach actually is transparent to the human user who can be said to understand it. The most deplorable kind of e-coach might be one which seems to integrate seamlessly onto a user's life and seems to operate under the user's control in reaching their goals, but is actually subtly engaged in persuading the user to do something they don't want to do. A feeling of alienation is not in itself sufficient to be called an instance of inauthenticity, however, it is still undesirable. This feeling of alienation that a user might have regarding their goals might however be added to by the phenomenon described in De Vries's answer 3. *Answer 3* concerns an uncanny confrontation with ipse, I realizing that the categories that one belongs to seem to come from outside oneself.[103] As the e-coach integrates itself into the user's life, the user may be confronted by this and feel alienated, even if she feels in complete control of the functions of the e-coach. Therefore, e-coaches that truly extend the user's agency could be seen as quite innocent with regards to type 1 inauthenticity, even though they can, as de Vries would probably agree, create an uncanny confrontation with the ipse. Because it is a requirement of extended agency to be very transparent to the user, it follows that e-coaches that extend agency are transparent and therefore less likely to cause type 1 inauthenticity.

The next question is: how likely are extended e-coach agents to cause inauthenticity type 2 in themselves? One response might be that while the e-coach plays a role in defining many of the categories the user belongs to, the user is outsourcing the creation of the idem identity to the e-coach. And thus, if we see the user and the e-coach as an extended agent, nothing is lost. However, I would argue that there is potential for causing inauthenticity here. Because the agent is not necessarily defining its identity against the background of horizons of significance if it is defining its identity against the categorizations in the profile made by the e-coach. Neither is the e-coach, when categorizing, necessarily using horizons of significance. As de Vries also points out, algorithms often create categorizations which work

[103] De Vries, "Identity, profiling algorithms…," (2010). 81.

for some purpose but make little sense to humans.[104] An example would be categorizing someone as “20% more likely runner on Saturdays than on Tuesdays when weather is bad.” In short, the activity of defining one’s identity against horizons of significance is not easily outsourced to an e-coach. If it was easily outsourced one could argue that inauthenticity type 2 is unlikely in extended agents, because it would not be deskilling, but just re-locating the skill. But since algorithms don’t think in terms of significance, but in terms of utility, this is an inoutsourceable skill. This would mean that e-coach agents can cause inauthenticity within themselves.

The moral deplorability of epistemic arrogance in the self-profiling activities of an e-coach as an extended agent leans on the fact that the agent may cause inauthenticity within itself. It can cause type 1 inauthenticity by profiling someone wrong. Therefore, epistemic humility can prevent this situation because it will likely prevent faulty profiling. A user-e-coach agent can cause type 2 inauthenticity by distorting the process of creating an authentic identity. Epistemic humility in the system as a whole can remedy or prevent this situation. If the e-coach-agent is epistemically humble and makes effort to know itself better, then it is less likely to lose contact with horizons of significance which are only accessible to the human part of the system. As an analogy, we could understand the operations of the device within this agent as “type 1 thinking” and the human operations as “type 2 thinking.” If the system as a whole is epistemically humble, the human being is also engaged in the cognitive processes, enough not to make the identity of the system inauthentic. The extended agent mode requires epistemic humility to flow from the activity of both the device and the human.

In the case of distributed agents, the same applies as for extended agents. The risks of causing inauthenticity type 1 and 2 are equally present in distributed agents because the

[104] De Vries, “Identity, profiling algorithms…,” (2010).

characteristics of the system are essentially the same, with the exception that in extended agency there is always a higher degree of transparency and integration. However, if the e-coach is considerable less integrated or transparent to the human user, then it could be expected that type 1 authenticity happens more easily, as the human user has less knowledge over how the e-coach profiles and nudges the user.

7.b. Machine agents

If we see the e-coach as an independent machine agent, then we may adopt a very simple view of the profiler-profilee relationship. Namely, the e-coach is the profiler and the human user is the profilee. A machine agent can be seen as having power over a human user in the sense that it may cause inauthenticity type 1 if it fails to profile and influence the user according to her wishes. This already grants the e-coach some power. From the machine agent perspective, the e-coach can also cause type 2 inauthenticity in the user. If the user becomes reliant on the e-coach to determine her idem identity, then her identity could be said not to be authentic. De Vries's *Answer 2* predicts an obliteration of the user's sense of identity, because the ways in which one's idem is defined comes from outside the user and the user may lose her skills in constructing her own identity.[105] An e-coach as a machine agent can be seen as capable of producing both kinds of inauthenticity, especially given that it is very persuasive.

8. Conclusion: How should we hold e-coaches responsible for conforming to responsibilist virtues?

I began by arguing that e-coaches can be seen as moral agents. Then I discussed the importance of epistemic virtues in the moral evaluation of e-coaches. Epistemic humility, I

[105] De Vries, "Identity, profiling algorithms…," (2010).

showed, is a key virtue for e-coaches. At last, I analysed the conditions in which epistemic humility becomes a heavier obligation, namely, in cases where the e-coach with its judgements on a user's profile has power over the user. Power, I argued, should be understood as the capacity or likelihood of an e-coach to cause inauthenticity. That is a chronological summary of what I have argued.

To clarify my argument, I shall now make my point in a slightly different order. E-coaches engage in profiling. If we reflect on our intuitions about profiling, we can draw a few conclusions: Firstly, that profiling is a morally laden act where epistemic and moral responsibility are intertwined. Secondly, that the instances of profiling that we deem wrong are characterized by epistemic arrogance, which involves self-centered or lazy cognitive behaviors. Thirdly, we realize that the moral deplorability of these cases also depends on how much power the profiler has over the profiled individual and, thus, how much influence the arrogant judgement can have. It is clear that profiling, therefore, should be done in a way that is epistemically humble in order to be justified, especially if it is done by agents that have a lot of power over the profiled individuals.

The next question is: can e-coaches be considered the kinds of agents that one holds responsible for their profiling activities in a responsibilist virtue epistemological sense? The first section answers that they can. E-coaches can be agents in three different ways. If an e-coach is very transparent to a user and integrated in her functioning, then it can constitute, together with the user, an extended agent. The human user's agency extends to the e-coach. Even if an e-coach is not transparent or especially ingrained in a user's functioning, but nevertheless still has influence on the actions of a user, then we might consider it to constitute a distributed agent together with the user. We may also view an e-coach as an independent moral agent which acts upon the user. This is a perspective on agency that we might take when we are evaluating the changes that an e-coach brings about in a person's life, compared to when they did not have the e-coach. One might wonder whether it makes sense to attribute responsibilist virtues to these kinds of agents. However, as Zagzebski has argued, simply considering traits cognitive virtues because they produce true beliefs reliably does not explain what is good about knowledge. If this goes for humans then it should also go for machines, which produce beliefs in humans, or machine-equivalents of beliefs in themselves. Therefore, we have reason to hold e-coaches responsible for living up to the standards of responsibilist virtue epistemology.

What should the practice of e-coaching learn from this conclusion? The most important lesson for e-coach designers is the relevance of responsibilist virtues for e-coaches. Ideally, epistemic humility should be implemented into the profiling practices of e-coaches, for instance by making them self-learning and self-correcting. As argued by Wallach and Allen, virtue theories have the characteristics of a theory that could do well once implemented into devices. This is because it has a “following rules” or top-down component as well as a “learning from experience” or bottom-up component. The applicability of responsibilist epistemic virtues to e-coaches also means that we should judge them as vicious if they display epistemically arrogant behaviour, or have a biased or skewed algorithm.

Another lesson for e-coach designers could be the importance of attention to the different modes of agency that e-coaches can have. Designers of e-coaches should be aware that the devices do not merely act on their users but also extend the agency of users or create distributed forms of agency. Realizing this allows designers to pay attention to what the characteristics of an e-coach could add to a human agent and what influence they could have within a distributed agent.

In the introduction I outlined a few worries about a future where everyone uses e-coaches. What if e-coaches make us do things we didn’t really want to do? What if we become less responsible for our actions because our lives are so co-influenced by e-coaches?

What if e-coaching makes us inauthentic? Perhaps my conclusions can trim some plants so that a way can be paved to answering some of these worries.

As for the first worry about e-coaches leading us to do things we did not want to do, let us leave aside scenarios where e-coaches are hacked by people or governments to steer people's actions into a desired outcome. In well-meant e-coaches, generally, whether the device understands what a user wants to do or turn into depends on how well the e-coach understand the user through profiling. In cases where the e-coach drives the user to an undesired outcome, we know that the responsibility for the action can be put on either the e-coach as an artificial moral agent, or on the user-e-coach system or the extended agent (user extending to the e-coach). The actions of such agents can also be evaluated through virtue theory where epistemic virtues play a key role. Since the integration of a user and an e-coach also depends on having knowledge of each of these components, epistemic virtues on the behalf of both parts is essential for the formation of an extended agency. Understanding how virtues work in these different kinds of agents will be important in understanding moral action in a world that is steered by algorithms and human-machine hybrid agents.

The second worry, about whether e-coaching makes us less resposnible for our actions, is a legitimate one. Not only because we want someone to blame when something goes wrong, but also because we want to admire people for what they achieve. I have argued that it is possible that e-coaches make us extended agents, I do not think that the extension of our agency make us any less responsible for our actions. As Heersmink argues, conditions for extended agency are that the system in use is entirely transparent and integrated into the functioning of a user.[106] Transparency means control and responsibility over the system as a whole. The working of the e-coach must be sufficiently transparent for the user and the user's

[106] Heersmink, "Distributed cognition and distributed morality," (2017).

wishes must also be sufficiently transparent to the e-coach. This kind of self-knowledge within a hybrid agent can also be understood as a form of intellectual humility.

As for the third worry, I have outlined two kinds of inauthenticity that e-coaching can cause. The first kind of inauthenticity is where a person comes to act in a way that is incoherent with some part of their identity and causes them to feel alienation. This kind of inauthenticity can be prevented with profiling that is epistemically humble. The second kind of inauthenticity, as described by De Vries, involves a disruption of the valid process of acquiring values that match one's identity. This is much harder to remedy. But I think that it could be relieved by e-coaches involving a user more deliberatively in the process of self-profiling. This is difficult to do, also because it costs time for the user and could therefore decrease adherence.

Winston Churchill one said "We shape our buildings; thereafter they shape us."[107] We have realized this and we now actively make technologies for the purpose of shaping ourselves. E-coaches are meant to make us better persons. If it is morally good to be careful in one's assumptions and judgements, then it is morally good too if the technologies we create are careful in their assumptions and judgements. In today's world, the things we shape also shape us, and the boundaries between us and the things we shape are wearing thin. It is time to take theories about virtue and expand them to the kinds of agents we are becoming.

Bibliography

[107] Ursula Hartenberger, "Why buildings matter" in *The Guardian*. Friday 1st July 2011. https://www.theguardian.com/sustainable-business/sustainable-building.

Anderson, Joel and Bart J. Kamphorst, "Ethics of e-coaching: Implications of employing pervasive computing to promote healthy and sustainable lifestyles." *The Third IEEE International Workshop on Social Implications of Pervasive Computing* (2014).

Alfano, Mark. "Expanding the Situationist Challenge to Responsibilist Virtue Epistemology." *Philosophical Quarterly* 62 (2012): 223-249.

Altman, Scott. "Reinterpreting the right to an open future: From autonomy to authenticity." *Law and Philosophy* (2017): 1-22.

Bekey, George, A. *Autonomous Robots: From Biological Inspiration to Implementation and Control*. Cambridge, Massachusetts: The MIT Press, 2005.

Boylan, Michael. "Ethical Profiling." *The Journal of Ethics* 15 (2011): 131-145.

Butterworth, Andrew, O'Donoghue, Peter and Cropley, Brendan. "Performance profiling in sports coaching: A review." in *International Journal of Performance Analysis in Sport* 13 (2013): 572-593.

Chaiken, Shelly, Wendy Wood, and Alice Eagly. "Principles of persuasion." In *Social psychology: Handbook of basic principles*. Edited by Edward T. Higgins and Arie W. Kruglanski, 702-742. New York: Guilford, 1996.

Clark, Andy and David J. Chalmers. "The extended mind." *Analysis* 58 (1998): 7-19.

Code, Lorraine. "Toward a 'Responsibilist' Epistemology." *Philosophy and Phenomenological Research* 45 (1984): 29-50.

Coeckelbergh, Mark. “Health Care, Capabilities, and AI Assistive Technologies.” *Ethical Theory and Moral Practice* 13 (2010): 181-190.

Corlett, J. Angelo. “Profiling Color.” *The Journal of Ethics* 15 (2011): 21-32.

Davis, Seth. *Getting to Us*. New York: Penguin Press, 2018.

Dunning, David, Ann Leuenberger, and David A. Sherman. “A new look at motivated inference: Are selfserving theories of success a product of motivational forces?” *Journal of Personality and Social Psychology* 69 (1995): 58-68.

Dworkin, Gerald. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press, 1988.

Efferson, Charles, Rafael Lalive, and Ernst Fehr. "The Coevolution of Cultural Groups and Ingroup Favoritism." *Science* 321 (2008): 1844–1849.

Ekstrand, Michael D. and Martijn C. Willemsen. “Behaviorism is not enough: Better Recommendations through Listening to Users.” *Proceedings of the 10th ACM Conference on Recommender Systems* (2016).

Floridi, Luciano. “Distributed Morality in an Information Society.” *Science and Engineering Ethics* 19 (2013): 727-743.

Floridi, Luciano. “Information ethics: On the philosophical foundation of computer ethics.” *Ethics and Information Technology* 1 (1999): 37-56.

Floridi, Luciano and J. W. Sanders. “On the Morality of Artificial Agents.” *Minds and Machine* 14 (2004): 349-379.

Heath, Joseph and Joel Anderson. “Procrastination and the Extended Will.” in *The Thief of Time: Philosophical Essays on Procrastination.* Edited by Chrisoula Andreou and Mark White. (New York: Oxford University Press, 2010).

Hellman, Deborah. “Racial profiling and the meaning of racial categories.” In *Contemporary Debates in Applied Ethics 2.* Edited by A. I. Cohen and C. H. Wellman. Malden: Wiley-Blackwell, 2014.

Heersmink, Richard. “Distributed cognition and distributed morality: Agency, artifacts and systems.” *Science and Engineering Ethics* 23 (2017): 431-448.

Howell, Robert J. “Extended virtues and the boundaries of persons.” in *Journal of the American Philosophical Association* 2 (2016): 146-163.

Ihde, Don. *Technology and the Lifeworld: from garden to earth.* Bloomington, Indiana: Indiana University Press, 1990.

Ives, Yossi. “What is coaching?: an exploration of connecting paradigms.” *International Journal of Evidence Based Coaching and Mentoring* 6 (2008): 100-113.

John, Peter, Alice Moseley, Sarah Cotterill, Liz Richardson, Gerry Stoker, Corinne Wales, and Graham Smith. *Nudge, Nudge, Think, Think: Experimenting with Ways to Change Civic Behaviour.* London: Bloomsbury Academic, 2013.

Kahn, Jonathan. “Getting the Numbers Right: Statistical Mischief and Racial Profiling in Heart Failure Research.” *Perspectives in Biology and Medicine* 46 (2003): 473-483.

Kahneman, Daniel. *Thinking, Fast and Slow*. London: Penguin Books, 2011.

Kamphorst, Bart A. “E-Coaching Systems: What They Are, and What They Aren't.” in *Personal and Ubiquitous Computing* 21 (2017): 625-632.

Kamphorst, Bart A. and Annemarie Kalis. “Why option generation matters for the design of autonomous e-coaching systems.” *AI & Society* 30 (2015): 77-88.

Kanoje, Sumitkumar, Sheetal Girase, and Debajyoti Mukhopadhyay. “User Profiling Trends, Techniques and Applications.” *International Journal of Advance Foundation and Research in Computer* 1 (2014). 2348-4853.

Latour, Bruno. “On technical mediation: philosophy, sociology, genealogy.” *Common Knowledge* 3 (1994): 29-64.

McCormick, Iain and Giles St. J. Burch, “Personality-focused coaching for leadership development.” *Consulting Psychology Journal: Practice and Research* 80 (2008): 267 – 278.

Nickel, Philip, J. “Ethics in e-trust and e-trustworthiness: the case of direct computer-patient interfaces.” *Ethics of Information Technology* 13 (2011): 355-363.

Philips. “About Health Suite” on https://www.philips.com.au/healthcare/innovation/about-health-suite. Visited on 21st January 2017.

Pronin, Emily and Matthew B. Kugler. “Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot.” *Journal of Experimental Social Psychology* 43 (2007): 565-578.

Rawls, J. *A Theory of Justice: Revised Edition*, (Cambridge: Harvard University Press: 1971 (1999)). 19 – 24.

Samuelson, Peter L. and Ian M. Church. “When cognition turns vicious: Heuristics and biases in light of virtue epistemology.” *Philosophical Psychology* 28 (2015): 1095-1113.

Schirmer Dos Santos, César. “Self-Knowledge and Epistemic Virtues: Between Reliabilism and Responsibilism.” *Veritas – Revista de Filosofia da Pucrs*, 60 (2015): 579-593.

Smart, Paul R. “Mandevillian Intelligence” in *Synthese* 1 (forthcoming).

Smith, Holly M. “The Subjective Moral Duty to Inform Oneself before Acting.” *Ethics* 125 (2014): 11-38.

Spotify. “Discover” on https://support.spotify.com/my-ms/using_spotify/discover_music/discover/. Visited on 21st January 2017.

Spanakis, Jerry (G.), Bastiaan Boh, Lotte Lemmens, and Anne Roefs. “Machine learning techniques in eating behavior e-coaching.” *Personal and Ubiquitous Computing* 21 (2017): 645-659.

Sparrow, Robert and Linda Sparrow. “In the hands of machines? The future of aged care.” *Minds and Machines* 16 (2006): 141-161.

Stanovich, Keith E. “Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?” In *Two Minds: Dual Processes and Beyond*. Edited by Keith Frankish and Jonathan St. B. T. Evans. 55-88. Oxford University Press, 2009.

Stanovich, Keith E. and Richard F. West. “Natural myside bias is independent of cognitive ability.” *Thinking & Reasoning* 13 (2007): 225–247.

Varga, Somogy and Charles Guignon. "Authenticity." *The Stanford Encyclopedia of Philosophy*, Edited by Edward N. Zalta (2017). Viewed on 9th May 2018.

Verbeek, Peter-Paul. *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press, 2011.

De Vries, Katja. “Identity, profiling algorithms and a world of ambient intelligence.” *Ethics and Information Technology* 12 (2010): 71–85.

Tanesini, Alessandra. “"Calm Down Dear": Intellectual Arrogance, Silencing and Ignorance.” *Aristotelian Society: Supplementary* 90 (2016): 71-92.

Taylor, Charles. *The Ethics of Authenticity*, Cambridge: MA: Harvard University Press, 1992.

Thomsen, Frej K. “The Art of the Unseen: Three challenges for Racial Profiling.” *The Journal of Ethics*, 15 (2011): 89-117.

Tonkens, Ryan. “Out of character: on the creation of virtuous machines.” *Ethics and Information Technology* 14 (2012): 137-149.

Wallach Wendell and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press, 2009.

Webber, Jonathan. “Virtue, Character and Situation.” *Journal of Moral Philosophy* 3 (2006): 193-213.

Zagzebski, Linda T. “In search for the source of epistemic good.” *Metaphilosophy* 34 (2003): 12-28.

Zagzebski, Linda T. *Virtues of the Mind: An Inquiry Into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press, 1996.