The primacy of human autonomy: understanding agent rights through the human rights framework

Bart Kamphorst[1]

Abstract. This paper is concerned with the 'rights' of autonomous agent systems in relation to human users or operators and specifically addresses the question of when and to what extent an agent system may take over control from someone. I start by examining an important ethical code of conduct for system designers and engineers and argue that one would do well to understand it within the human rights framework. I then show that framing the discussion on what agent systems may and may not do in terms of human rights has consequences for intelligent agent systems in that they should be respectful of people's dignity and autonomy. In the remainder of the paper I work out the implications of this for the conditions under which agent systems may take over control. I offer an analysis of control, of delegated control, and of autonomy-respectful delegated control, concluding that for an agent system to justifiably take over control from a user, it should at a minimum offer the user a reliable way to take back control in a timely manner. However, when the user's autonomy is at stake, the system should also know about and act in accordance with the user's goals and core values.

1 Introduction

When and to what extent should an autonomous agent system be able to take over control from a human user? This question is becoming more and more relevant because people are increasingly employing intimate, agent-based support systems that have a profound influence on their personal lives (e.g., in regulating chronic illness [22] or in overcoming obesity [4]). But before one can start formulating an answer to such a question, one needs to have a clear conception of control and of the ethical considerations that come into play when a person hands over control to an agent system. For instance, how does handing over control affect the user in terms of one's well-being or autonomy? Throughout this paper I will propose four guidelines aimed to pave the way towards an answer that takes these issues into account. I will start by examining an important ethical code of conduct for system designers and engineers and argue that in order to properly understand it, one would do well to place it within the human rights framework as codified in the Universal Declaration of Human Rights (UDHR).[2] I will then show that by doing so, it follows that the notion of human autonomy is a value worth protecting when designing intimate systems. Finally, I will set forth some tentative thoughts about the relation between autonomy and control, in order to sketch the outline of an answer to the question, namely that under normal circumstances an autonomous agent system may take over control if and only if the control is willingly delegated to the system and the user retains the possibility to take back control when he or she sees fit.

2 Personal Dignity

In 1992 the council of the Association for Computing Machinery (ACM) adopted a code of ethics that prescribes the ethical professional conduct that is expected of every member of the ACM, an organization with over 100,000 members from over 100 countries, representing the worlds largest educational and scientific computing society. Consider article 3.5 of this code, entitled "As a ACM member I will 3.5: Articulate and support policies that protect the dignity of users and others affected by a computing system" [5]:

> Designing or implementing systems that deliberately or inadvertently demean individuals or groups is ethically unacceptable. Computer professionals who are in decision making positions should verify that systems are designed and implemented to protect personal privacy and enhance *personal dignity*. (emphasis added)

While the gist of this article is clear, one cannot truly understand what it entails unless one has a working idea of what personal dignity is. That is, what is it that needs protecting, or even enhancing? The problem is that because human dignity is an extremely broad concept that people interpret differently in different contexts, there is a danger that in practice the meaning of it is void.[3] To remedy this, given the societal importance of this code of ethics, I suggest understanding personal dignity here in the context of the human rights framework.[4] Within this framework, codified in the UDHR, personal dignity can be understood as a fundamental moral property of people that they are normative agents worthy of respect, a notion that lies at the very core of the framework. Article 1 for instance states that "All human beings are born free and equal in dignity and rights", and dignity also plays a role in positive rights such as social security (art 22) and the right to employment (art 23). In other words, it is through this fundamental moral property of dignity that people have certain rights in the first place, namely those rights that protect personal dignity. In the discussion about what autonomous agent systems may and may not do, then, protection of the user's personal dignity seems like a sensible place to start.

[1] Utrecht University, The Netherlands, email: bart.kamphorst@phil.uu.nl

[2] Because I understand the notion of agent rights in the context of the human rights framework, and I am not inclined to grant autonomous agent systems a human-like status within this framework, I will refrain from speaking of rights and duties of autonomous agents from here on out.

[3] For a critical discussion of the use of the term in medical ethics, see [17]. Interestingly enough in light of this paper, according to Macklin dignity means 'nothing more' than having to respect people's autonomy!

[4] There are also other hints such as "As a ACM member I will 1.1: Contribute to society and human well-being." that indicate that the placement of the code in this context is valid.

Guideline One: Under no circumstances may an autonomous agent ever be harmful to anyone's personal dignity.

While this may seem like a trivial guideline, and an obvious one at that, it may serve as a stepping stone for other guidelines with the human rights conception of personal dignity as their foundation. In the following section I will argue that the next step up is that autonomous agent systems should respect human autonomy.

3 Personal Autonomy

The UDHR does not attribute autonomy to every human being like it does with dignity. Instead, it assumes that people have the capacity for self-rule and strives to lay the groundworks for an environment in which people can develop their autonomy, i.e. become autonomous beings. Personal autonomy, understood here as having the freedom, the capacity and the authority to choose one's own course of action to direct one's life in accordance with one's goals and values, takes a prominent place in (philosophical) discussions about agency. Constituted by a measure of independence and a (minimal) requirement of rationality (see [2] for a thorough discussion), being personally autonomous has normative implications in that one is held responsible for one's choices and actions, but also that others have obligations to respect one's right to decide on and follow a certain course of action. This in turn has a significant impact on the political and legal domain, in that it determines how "the state is permitted to restrict or influence individuals' choices of how to lead their lives" [2]. Personal autonomy is closely related to personal dignity, in that dignity is a necessary condition for leading an autonomous life. Conversely, however, this need not be the case. People can be more or less autonomous than each other or even than themselves viewed over a period of time, without any threat to their dignity. What this observation shows is that autonomy, contrary to dignity, may be viewed as a scale [2, sec. 3.1]. But note that the fact that autonomy can be viewed as a scale does not mean it is negotiable. As Anderson rightly notes, the right to autonomy should be "understood not in terms of ideals of development but rather as a fundamental boundary not to be violated" [2, p. 12, sec 3.2]. Similarly, Oshana remarks that "[t]he fact that it might be morally and legally incumbent upon us to caution others against their own behavior, to warn them of the punitive consequences that might follow their behavior, and to actually take steps to curtail their autonomy, does not mean that autonomy is not a valued state. This ideal remains intact, although uninstantiated in certain cases." [21, p. 126]. The value of autonomy is anchored within the UDHR and individuals as well as society should strive to maximize people's autonomy. Moreover, and very relevant for the discussion at hand, perceived autonomy can be measured, and there is empirical research that shows that diminished autonomy negatively affects personal well-being [e.g. 23]. Therefore, I contend that autonomy may be restricted only insofar as the exercise of autonomy frustrates anyone else's autonomy or the state has compelling reasons to do so.

Guideline Two: Autonomous agent systems should be respectful of people's autonomy. They may not diminish a user's autonomy, unless otherwise directed by law.

At this point I would like to touch upon a possible criticism, voiced but dismissed by Verbeek in a discussion about persuasive technology:

> [A]utonomy was thought to be attacked when human actions are explicitly and consciously steered with the help of technology. This reduction of autonomy was even perceived as a threat to human dignity; if human actions are not a result from deliberate decisions but from steering technologies, people were thought to be deprived from what makes them human. [24]

I agree with Verbeek here that this is not the way to think about the relation between technology and autonomy. What I will come to argue later in this paper is that being steered (better: guided) by an intelligent, autonomous agent system, does not impede one's autonomy per se, but that it might become an issue when one cannot change one's course of action when one has good reasons to do so.

Thus far I have argued that in order to say something meaningful about what intelligent agent systems may and may not do, one would do well to frame the discussion in terms of the human rights framework. I have argued that human autonomy is a central notion within this framework, that it is an important value to protect in all circumstances, and therefore should be regarded as an important value when designing autonomous agent systems that (closely) interact with people.

4 Value Sensitive Design

The idea that software should be respectful of people's autonomy is not new. Most prominently, Friedman has argued at length for the inclusion of human values in the design of computer systems and software agents, specifically respecting and enhancing what she calls 'user autonomy' [11, 10, 12]. In their treatment of human agency and responsible computing, Friedman and Kahn begin by asking the question whether autonomous agent systems can be moral agents like human beings, something they then argue cannot be the case because to date computer systems do not have intentionality, and intentionality is taken to be a prerequisite for morality [11]. But because in some cases people's understanding of this boundary between humans and computational systems is distorted, the argument continues, people's sense of moral agency can be diminished, which in turn causes erosion (their terminology) of their dignity. Friedman and Kahn then propose design strategies to preclude this distortion from happening such as nonanthropomorphic interface design (sharpen the distinction between humans and computers) and participatory design (involving users in defining the problems the system should tackle). In later work, Friedman proposes user autonomy as an important value to take into account when practicing *Value Sensitive Design* (VSD) for developing user-centered systems "because it is fundamental to human flourishing and self-development" [10], citing work by Gewirth (1978) and Hill (1991). Here, Friedman distinguishes between System Capability, System Complexity, Misrepresentation of the System and System Fluidity as aspects of systems that can influence user autonomy [10].[5] While Friedman and I share autonomy as an important value, we seem to differ on the circumstances under which autonomy is in danger. For instance, Friedman and Nissenbaum write that "user autonomy can be undermined when there are states the user desires to reach but no path exists through the use of the software agent to reach those states" [12]. As an illustration they consider a mail agent that has the capability to filter emails by subject header, but does not understand a concept such as urgency. This, Friedman and Nissenbaum argue, leads to the undermining of autonomy for the user who wishes to filter emails by urgency. I think this is too strong: the fact that one's expectations about the capabilities of the agent do not match reality does not threat one's autonomy

[5] Elsewhere, Friedman distinguishes a fifth aspect, namely knowledge about the system: how the (non-)transparency of a system's internal workings may influence user autonomy. See [10, 12].

per se. The software agent was employed by the user and could also be disabled by the user. So, at worst, the user's hopes for autonomy enhancement were unrealized, but there is no loss of autonomy to speak of. It is only when control is handed over but cannot be regained, I will argue in Section 6, that in some cases one's autonomy can be in trouble.

What I do share is Friedman's insight that values such as autonomy (or freedom from bias) do not necessarily override others. So where the right to autonomy is a fundamental boundary not to be violated, the level to which one may exercise this right may under some circumstances be restricted, for instance "to protect against a user with malicious intentions or well-intentioned users guided by poor judgment" [10, p. 22], or "in situations where safety is at stake" [12, p. 6]. Given the status of autonomy however in societies that subscribe to the human rights framework, I think that such judgements should be left to the legislator or the judiciary (as is visible from Guideline 3).[6] Nevertheless, I subscribe to the idea of value-sensitive design and, albeit via a different route I too suggest that human autonomy should play a central role in the discussion of what autonomous agent systems may and may not do.

So, having framed the discussion in terms of the human rights framework and having argued that human autonomy should be considered a central concept, it would now be possible to give a tentative answer to the question posed in the introduction by saying that it is acceptable for a system to take over control, *as long as it doesn't impede one's autonomy*. But what this really only does is reframe the problem, because this answer does not provide any insights about the conditions under which control impacts autonomy. To work towards an actual answer, then, it is now time to spell out what control is, before turning to the relation between control and autonomy.

5 (Delegated) Control

Having control over something roughly means being causally responsible for a particular state of that something. Like autonomy, control is another concept constitutive of and interrelated with human agency. In the first place, people have self-control: "control of the self by the self" [19], which implies being causally responsible for one's own decisions and actions in accordance with one's goals. This type of control is known as executive control, and has been considered (but debated) as the basis on which one can ascribe morality and responsibility to people [3]. Secondly, people can control parts of their environment, such as a soccer ball, or a computer.[7] Importantly, however, being in control need not be limited to human beings. Take the classic example of the thermostat. Without going into the discussion about whether the thermostat can be said to have beliefs and goals about temperature, it is uncontroversial to say it controls the temperature of a room. Now what is important to observe, is that control is transitive: if one controls the thermostat, one controls the temperature of the room by means of the thermostat. This observation, I will argue, plays a crucial role in understanding when delegation of control is justified.

Delegating control, I propose, means handing over immediate causal responsibility over some object or process to another entity, with the provision that one can retake control when one sees fit. The entity may either be another human being or an autonomous system of some sort. The object of control can many things, including decision-making processes that normally would be handled by the self, but do note that handing over self-control as such is a contradictio in terminis. If the delegate system is so intimately coupled with the delegator that the delegator considers it part of the self, then there is no delegation to speak of, only self-control. If, on the other hand, control over the self is delegated to an external entity, we cannot speak of self-control, as it is not the case that the self controls the self. Nevertheless, in principle, control over a great many things can be delegated. But what are the conditions that the delegate should conform to?

In answer of this question I would like to start with a useful distinction made by Fischer and Ravizza in the discourse on the minimal requirements for moral responsibility between guidance and regulative control. Whereas others have held that moral responsibility requires full-blown regulative (cf. executive) control, Fischer and Ravizza contend that guidance control, i.e. "the agent's "ownership" of the mechanism that actually issues in the relevant behavior, and the "reasons-responsiveness" of that mechanism" [7] — meaning that the mechanism is (moderately) sensitive to reasons for acting differently — is both necessary and sufficient. While Fischer and Ravizza's distinction is not directly applicable to a structure of delegated control (as I will show), it provides some useful insights. First, the requirement of ownership over the mechanism that issues the behavior also applies to delegated control: one has to have a certain ownership over the delegate. This does not necessarily entail physical or legal ownership, but a kind of ownership that follows from "taking responsibility for them" [7]. So, for example, reading instructions and knowingly enabling an agent system may be sufficient for this. Secondly, there has to be some sort of mechanism that allows for intervention, i.e. for taking back control. But while moderate reasons-responsiveness is a reasonable requirement for guidance control, it seems to be too strong for delegated control because although some intelligent, autonomous agent systems may be reasons-responsive, others systems (e.g. thermostats) are not. For delegated control, then, I suggest a weaker condition, namely a mechanism that is responsive to *control-retraction*. In other words, the delegate should have a mechanism with which the delegator can take back control.[8] For it is the presence of such a mechanism that determines whether the transitive control relation still holds: if one cannot take back control, then it is not delegation but transference, or attribution. This mechanism has to be reliable as well as responsive in a timely manner. Note that the latter condition is especially important because in some cases it will be paramount that the mechanism will hand back control immediately. At the same time, this criterion leaves room for mechanisms that require a slightly higher threshold to be met — within the limits of reasonableness — for control-retraction, such as having to type in a twenty digit passphrase as opposed to hitting a big red button.

> *Guideline Three*: Delegation of control is valid as long as the delegator has ownership over the delegate, and the delegate offers the delegator a mechanism that is reliably responsive to control-retraction in a timely manner.

To illustrate this idea, consider the case of Alice, an ordinary woman who has set herself the goal to loose a few pounds. Alice is in control of what food she consumes, but when dieting, she finds that it takes a lot of self-control to refrain from eating sweets. Now to help

[6] Notice here the important difference between autonomy and dignity: a court could never justify indignity, that would imply a violation of a fundamental absolute right, and there cannot be a justification for such a violation.

[7] Note that while the object of control differs in both cases, what matters is that one is a dominant causal factor, not necessarily the single cause.

[8] Observe that delegated control is perfectly compatible with guidance control: the internal mechanism that controls the delegation structure should be one's own and be moderately reasons-responsive (guidance control).

herself stay on track, she decides to enable an autonomous, agent-based support system — one that is responsive to control-retraction, i.e. one that can be overridden, for instance when Alice has her friend Bob over for dinner — that draws up a grocery shopping list for her every other day, and orders food online. By doing so, the system is effectively preventing Alice from wandering through the supermarket where she would be confronted with temptation. Now surely Alice has not given up self-control over what she eats: whatever groceries are delivered, she can choose to eat them or not. What she has done, though, is delegate her control over the decision making process to the agent system for what foods to buy.

So, I have argued that control is something that can be delegated, and that the delegation is valid as long as the delegator has taken ownership over the delegate and has a way of reestablishing executive control. In the following section I will discuss how delegation of control relates to autonomy.

6 Delegated Control and Autonomy: Initial Thoughts

Being in control is strongly connected to autonomy. Recall from Section 3 that autonomy consists in part of a measure of independence, which one can only establish if one has control over one's self (decisions, actions) and one's immediate environment. Now to see how delegation of control can work, but also how it can be problematic for one's autonomy, consider the following scenario. Bob, Alice's friend, is an autonomous human being, and as such, he can decide on the temperature of his own home. He can choose to build a fire in the fireplace, or to simply delegate control over the temperature to a thermostat. Should he choose the latter, the delegation of control would be unproblematic, because if Bob finds that it is too cold on a winter's day, he can control the temperature by means of the thermostat. But now consider a thermostat with no off-switch that, once activated, determines what the temperature should be all on its own (it is in fact an autonomous agent system). To make matters even worse, the system is unpredictable, because unbeknownst to Bob, it determines the temperature by taking the word of the day from `http://thesaurus.com/wordoftheday`, taking the number of results that Google's search engine generates for that word, performing a modulo operation on that number with 15, and adding a constant of 10.[9] Since Bob has no control over the thermostat, he therefore lacks control over the temperature in the room, which in turn impacts his autonomy.

To see that delegation of control does not always involve autonomy concerns, take a case where someone hands over immediate control over a soccer ball to a robocup robot that reliably passes the ball back when one asks for it (control-retraction). This is valid delegation. But should the robot decide not to pass the ball back (it may even be reasons-responsive itself, passing the ball instead to another robot in a better position to score!), this surely does not hamper one's autonomy.

Looking back at the original question about when a system may take over control, it thus matters whether the object of control has the capacity to affect one's autonomy. This capacity, which I will call *autonomy-sensitivity*, determines whether delegation of control alone is acceptable in dealing with an autonomous agent system, or that what is required is *autonomy-respectful* delegation of control. To speak of delegated control that is autonomy-respectful, I propose that one more condition must be met, namely goal and value conformance.

As previously mentioned, human autonomy is in part constituted by independence: being the authority over making one's own life choices in accordance with one's goals and core values. What this implies, is that if a delegate is taking control over something that is autonomy-sensitive to the delegator, the delegate has to act in such a way that the delegator perceives the decisions and actions of the delegate as an extension of the self in order to prevent interference with respect to the delegator's independence. To accomplish this, the delegate should know about and act in accordance with the delegator's goals and core values.

> *Guideline Four*: Delegated control is autonomy-respectful if and only if there is valid delegation of control over something that is autonomy-sensitive, and the delegate acts in accordance with the delegator's goals.

There are a number of things to note about this final requirement. The first is that it relates to Friedman and Nissenbaum's notion of 'agent fluidity': "software agents need to take [evolution of the user's goals] into account and provide ready mechanisms for users to review and fine-tune their agents as their goals change" [12]. Indeed, people's goals do change, and to prevent a delegator from feeling alienated from the delegate's decisions and actions, for instance because it is striving to obtain an outdated goal, agent fluidity should be an important element in agent systems design. Secondly, attesting to the importance of the requirement, is that it relates to Ryan and Deci's self-determination theory, the idea that developing a sense of autonomy is critical "to the processes of *internalization and integration*, through which a person comes to self-regulate and sustain behaviours conductive to health and well being" [23]. Especially where agent systems are in a position to instruct and guide the delegator's behavior (e.g. Klein, Mogles, and Van Wissen's eMate), guiding them towards goals the users personally endorse is crucial. Finally, what this requirement highlights, is the importance of individualization, personalization, and tailoring [8]: individuals have different needs, preferences, beliefs, goals, and quite likely, different autonomy-sensitive objects of control. To see that this is so, consider Carol and Dave, who both decide to enable an intelligent agent system that will recommend clothes for them to wear on a daily basis. Carol, who has a great sense for fashion, uses the recommendations to pick and choose her wardrobe, and if she doesn't like the recommendation, she will happily wear something else. Dave on the other hand, has a very poor sense of fashion. In fact, it doesn't take long for Dave to rely on the recommendations of the agent system. But here's the catch: despite his poor sense of fashion, Dave does have certain values about dressing properly for the occasion, and for his new job as assistant professor, he wishes to look presentable, so not to undermine his credibility and authority.[10] Should the system recommend clothes that do not fit this profile, Dave's autonomy will be affected. So, this example illustrates how the object of control can be the same, but the autonomy-sensitivity can differ on an individual basis. This, too, must be accounted for by an autonomous agent system that has been delegated control of something that is autonomy-sensitive to the delegator.

7 Implications

In the previous sections I have argued that in order to say something meaningful about when an autonomous agent system may take over

[9] This would have made it a nice 21 degrees Celsius on May 5th 2012 with the verb 'besot' (approx. 431.000 hits).

[10] Example derived from [20, pp. 177–178].

control from a human user, we would do well to place the discussion within the human rights framework. The implication of this is an emphasis on people's personal autonomy. In order to better understand how having control relates to autonomy — something that is especially important for the design of intimate, agent-based support systems — I have offered an analysis of control, of delegated control, and of autonomy-respectful control delegation. This section elaborates on the implications of the conceptual work. First, very broadly, by framing the discussion in terms of human rights, we get a lot of practical rules for free, in that an agent system may exercise its autonomy (and thus the tasks delegated to it) as long as it does not frustrate anyone else's rights.[11]

Secondly, the question of *when* an agent can take over control, depends on the autonomy-sensitivity of the object of control. But at a minimum, when the object of control is not autonomy-sensitive, control may be taken over when it is willingly delegated — that is, the user should know about and agree with the delegation — and *to the extent* that the delegate has a mechanism that is responsive to control-retraction. One might ask at this point 'Is the control-retraction mechanism necessary?', because one could of course enable an agent system and simply let it run. My response to this question is twofold. Firstly, yes, such a mechanism is necessary in order to speak of delegated control, because without it, the transitive chain of control is broken. Secondly, although technically possible, control transference or control attribution is problematic in terms of responsibility, because on the one hand one cannot rely on the transitive control chain in those cases, and on the other hand it is problematic to consider the agent system a true moral agent with moral accountability [11]. So, normatively speaking, releasing the requirement of control-retraction is undesirable.

Finally, considering a case where the object of control is autonomy-sensitive, an individual's right to autonomy dictates that an agent system may only take over control when it is validly delegated, and the delegate has the capacity to act in accordance with the delegator's (changing) goals and core values. As shown, this is crucial for respecting and protecting the delegator's autonomy. The implications of this is that autonomous agent systems should use personalization and tailoring techniques, and should have access to personal information. Of course, this sparks two separate discussions, on the ethics of persuasive systems and on privacy respectively, but those are beyond the scope of this paper.

7.1 Exceptions

The guidelines laid down in this paper are not without exceptions. One type of exception in particular I would like to mention here, and those are the cases in which a person's autonomy is well below the (minimal) level of autonomy that is presupposed throughout this paper. It is highly conceivable that such cases, for instance that involve people who have very little self-regulatory capacities, should be treated differently. Here, I think, human dignity still plays a central role, but other criteria should be considered as well, such as a person's well-being or a person's prospects for autonomy enhancement. For example, if the use of an autonomous agent system without an overrule mechanism for that user would actually enhance that person's quality of life, then it seems to me such a system should be at least be considered to be allowed (given that it respects the person's dignity).[12] Considering what is at stake in such cases, taken together with common practice regarding lack of autonomy (e.g. legal guardianship), I think such decisions should be left to a specialized institution or a court of law.

[11] Note that in a societal context we may add the clause that actions must be lawful, i.e. legal within the boundaries of the law.

[12] Note that to preclude any issues of responsibility, the system should be responsive to control-retraction from a specialist care-taker or other authority figure.

8 Final considerations

Before concluding, I would like to mention two separate issues that should be addressed in the discussion of what autonomous agent systems may and may not do. The first is concerned with the difference between design and actual use, the second with a dilemma about the limits of autonomy.

8.1 Design versus Use

The main aim of this paper was to provide some preliminary ideas for thinking about the relation between humans and autonomous agent systems, in order to further the discussion of the normative judgements one can make about what such agent systems may and may not do, especially in relation to taking over control. Throughout this paper, though, I have also discussed some design principles that either follow from, or are important for the discussion at hand. I am aware that the relation between design and use is "very complex and principally unpredictable" [1], and I agree in principle that we must not overestimate the correlation between designer intention and actual use. Even so, in designing agent systems that are able to take over control, it seems that providing it with a mechanism that is responsive to requests to relinquish control is sensible and reasonable no matter what domain such an agent system will be used in.

I concede that with regard to goal and core value accordance, the difference between designer intention and actual use may prove more problematic. If actual use of a system is in a totally different domain than it was intended for, personalization and tailoring may fail. For these type of questions (e.g., 'What is the domain?', 'What information does the system need from the user?', 'What type of goal should the system strive for?'), proven methods of design should be used such as stakeholder analysis [9, 8] and empirical investigations as part of Friedman et al. tripartite methodology [13], perhaps accompanied by Verbeek's modified Constructive Technology Assessment to "anticipate possible mediating roles of the technology-in-design" [24]. We cannot always reliably predict actual use, but when a value as important as human autonomy is at stake, we should do our best to err on the safe side.

8.2 Dilemma: Ultimate autonomy?

Finally, I would like to note an interesting dilemma that this paper raises. I have argued that what should be protected is one's capacity to choose one's own course of action, or in other words, to live one's life by one's own standards and desires. Of course, the human rights framework and societal institutions put bounds on this capacity in that one can exercise one's right to autonomy only to the point where one would frustrate someone else's rights. Nevertheless, within that space, one is free: free to go hiking, free to whistle a show tune, even free to mutilate oneself. So why would one not be free to put an agent-based decision support system in place that severely and uncompromisingly restricts one's options? Doesn't the very fact that one is an autonomous being imply this freedom? In response to this dilemma I would like to draw an analogy with the autonomous being wanting to be enslaved, a case discussed in the philosophical discourse on autonomy [e.g. 18, 21]. One way of dealing with such cases

is to hold that the consensually enslaved is no longer autonomous because of the enslavement. As Oshana puts it: "Consensual slavery, regardless of the gains that it might provide and aside from any benefit to the enslaved, transforms the human subject into a possession or object of another and accordingly defiles the enslaved individuals autonomy" [21]. Analogously, one might argue that if someone willingly and knowingly enables an agent system that would place severe strain on that person's autonomy, that person's autonomy is lost. Not even necessarily by the doings of the agent system, but by placing oneself in that situation in the first place. But of course, this is an extreme case, and in practice this is unlikely to happen. People are not out to restrict their autonomy, they wish to reach a particular goal (e.g. having an agent system enforce strict dietary rules to become healthy). In the end, intelligent support systems should strive to help people reach those goals, but again, it is better to err on the safe side and make sure that these systems are respectful of people's autonomy.

9 Conclusion

In this paper I have argued that the discussion about what autonomous agent systems may and may not do should be framed within the human rights framework. I have shown that personal dignity and a human being's right to (personal) autonomy are important values in our society worthy of protection (guidelines 1 and 2), but also how there is empirical evidence that a lack of perceived autonomy negatively influences well-being. I have argued that the primacy of human autonomy should therefore be acknowledged in all discussions about what agent systems may and may not do in relation to their human users or operators. In an attempt to meaningfully answer the question when and to what extent an agent system may take over control, I have made the case that control over something that is autonomy-sensitive may be taken if and only if control is willingly delegated, the delegator assumes ownership over the delegate system, there is a mechanism in place with which to take back control reliably and in a timely manner (guideline 3), and the system acts in accordance with the delegator's goals and core values (guideline 4).

ACKNOWLEDGEMENTS

I thank Joel Anderson and Arlette van Wissen for their helpful comments on an earlier draft of this paper. I also appreciate the suggestions made by the anonymous referees of the RDA2 workshop. This research was supported by Philips and Technology Foundation STW, Nationaal Initiatief Hersenen en Cognitie NIHC under the Partnership programme Healthy Lifestyle Solutions.

REFERENCES

[1] A. Albrechtslund. Ethics and technology design. *Ethics and Information Technology*, 9:63–72, 2007.

[2] J. Anderson. Autonomy. In H. LaFollette, J. Deigh, and S. Stroud, editors, *International Encyclopedia of Ethics*. Wiley-Blackwell, Forthcoming. Expected late 2012.

[3] R.F. Baumeister and J.J. Exline. Virtue, personality, and social relations: Self-control as the moral muscle. *Journal of Personality*, 67(6), 1999.

[4] O.A. Blanson Henkemans, P.J.M. Van der Boog, J. Lindenberg, C.A.P.G. Van der Mast, M.A. Neerincx, and B.J.H.M. Zwetsloot-Schonk. An online lifestyle diary with a persuasive computer assistant providing feedback on self-management. *Technology and Health Care*, 17:253–267, 2009.

[5] ACM Council. Acm code of ethics and professional conduct. `http://www.acm.org/about/code-of-ethics`. Referenced on March 30th 2012.

[6] J.M. Fischer and M. Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, 1998.

[7] J.M. Fischer and M. Ravizza. Précis of responsibility and control: A theory of moral responsibility. *Philosophy and Phenomenological Research*, 61(2):441–445, 2000.

[8] B.J. Fogg. *Persuasive Technology: Using computers to change what we think and do*. Morgan Kaufmann Publishers, San Fransisco, 2003.

[9] R.E. Freeman. *Strategic management: A stakeholder approach*. Pitman, Boston, MA, 1984.

[10] B. Friedman. Value-sensitive design. *Interactions*, 3(6):16–23, 1996. ISSN 1072-5520. doi: 10.1145/242485.242493.

[11] B. Friedman and P.H. Jr. Kahn. Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software*, 17(1):7–14, 1992. ISSN 0164-1212. doi: 10.1016/0164-1212(92)90075-U. Computer Ethics.

[12] B. Friedman and H. Nissenbaum. Software agents and user autonomy. In *Proceedings of the first international conference on Autonomous agents (AGENTS '97)*, pages 466–469, New York, NY, USA, 1997. ACM. doi: 10.1145/267658.267772.

[13] B. Friedman, P.H. Jr. Kahn, and A. Borning. Value sensitive design and information systems. In P. Zhang and D. Galletta, editors, *Human-computer interaction in management information systems: Foundations*, pages 348–372. M.E. Sharpe, 2006.

[14] A. Gewirth. *Reason and morality*. University of Chicago Press, Chicago, 1978.

[15] T.E. Jr. Hill. *Autonomy and self-respect*. Cambridge University Press, UK, 1991.

[16] M.C.A. Klein, N. Mogles, and A. Van Wissen. Why won't you do what's good for you? Using intelligent support for behavior change. In *International Workshop on Human Behavior Understanding (HBU11). Lecture Notes in Computer Science*, volume 7065, pages 104–116. Springer Verlag, 2011.

[17] R. Macklin. Dignity is a useless concept. it means no more than respect for persons or their autonomy. *BMJ*, 327:1419–1420, Dec 2003. doi: 10.1136/bmj.327.7429.1419.

[18] J.S. Mill. *On Liberty*. 1859. Reprint: Filiquarian Publishing, LLC, 2006.

[19] M. Muraven and R.F. Baumeister. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, 126:247–259, 2000.

[20] C. Nass and C. Yen. *The Man Who Lied to His Laptop: What Machines Teach Us About Human Relationships*. Current (Pinguin Group), New York, NY, 2010.

[21] M. Oshana. How much should we value autonomy? *Social Philosophy and Policy*, 20(2):99–126, 2003.

[22] D. Preuveneers and Y. Berbers. Mobile phones assisting with health self-care: a diabetes case study. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services (MobileHCI)*, pages 177–186, 2008.

[23] R.M. Ryan and E.L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000.

[24] P-P. Verbeek. Designing morality. In *Ethics, Technology And Engineering: An Introduction*, chapter 7. Wiley-Blackwell, 2011.