Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being

Jason Borenstein* and Ron Arkin**

* Director of Graduate Research Ethics Programs, School of Public Policy and Office of Graduate Studies, Georgia Institute of Technology

** Regents' Professor, School of Interactive Computing, College of Computing, Georgia Institute of Technology

Introduction

The time is nearing when robots are going to become a pervasive feature of our personal lives. They are already continuously operating in industrial, domestic, and military sectors. But a facet of their operation that has not quite reached its full potential is their involvement in our day-to-day routines as servants, caregivers, companions, and perhaps friends. It is clear that the multiple forms of robots already in existence and in the process of being designed will have a profound impact on human life. In fact, the motivation for their creation is largely shaped by their ability to do so. Encouraging patients to take medications, enabling children to socialize, and protecting the elderly from hazards within a living space is only a small sampling of how they could interact with humans. Their seemingly boundless potential stems in part from the possibility of their omnipresence but also because they can be physically instantiated, i.e., they are embodied in the real world, unlike many other devices. The extent of a robot's influence on our lives hinges in large part on which design pathway the robot's creator decides to pursue.

The principal focus of this article is to generate discussion about the ethical acceptability of allowing designers to construct companion robots that nudge a user in a particular behavioral direction (and if so, under which circumstances). More specifically, we will delineate key issues related to the ethics of designing robots whose deliberate purpose is to nudge human users towards displaying greater concern for their fellow human beings, including by becoming more socially just. Important facets of this discussion include whether a robot's "nudging" behavior should occur with or without the user's awareness and how much control the user should exert over it.

Influences on Human Behavior

The behavior of human beings is shaped by numerous factors, many of which might not be consciously detected. Marketers are keenly aware of this dimension of human psychology as they employ a broad array of tactics, from music and scents to color schemes and emotional cues, to encourage audiences toward a preferred behavior (oftentimes the purchase of a client's product). Brain imaging and other neurotechnologies are increasingly being used by marketers to identify, and perhaps shape, customer preferences (Morin 2011; Ariely and Burns 2010). Filmmakers and novelists also know how to hone in on underlying veins of human psychology and use sophisticated techniques to subtly affect an audience's mood (Thomas and Johnston 1981). Along these lines, a much publicized study by Kramer and colleagues (2014) illustrates that covertly altering the news stories that *Facebook* users see on their homepages can affect whether they experience positive or negative emotions.

Roboticists, including those from the Georgia Institute of Technology (Arkin et al 2003; Moshkina et al 2011), are making use of empirical findings from psychologists and others to

inform their designs, and by doing so, effectively create robots that elicit strong reactions from users (e.g., some view the rolling robot *Roball* as being "cute" (Salter et al. 2008)). The robotics community is actively creating companion robots with the goal of cultivating a lifelong relationship between a human being and a robotic artifact. The intention behind companion robots is for them to have sophisticated interactions with their human counterparts over a prolonged stretch of time, potentially functioning as friends or caregivers. Of course, the success of such efforts hinges in large part on whether bonding effectively occurs during human-robot interaction (HRI). And the appropriateness of such interactions is partly contingent on whether it is ethically appropriate to deliberately design devices that can subtly or directly influence human behavior, a design goal which some, including Sparrow (2002), consider to be unethical.

The Definition of a Nudge

Thaler and Sunstein (2008) call the tactic of subtly modifying behavior a "nudge". It involves an attempt to shape behavior without resorting to legal or regulatory means. Thaler and Sunstein use numerous examples to illustrate their notion of a "nudge" including the state motto "Don't Mess with Texas" (2008, 60). They claim that eliciting feelings which remind us of a shared group identity can cause a noticeable behavioral change; in this case, it purportedly translates to a greater reluctance to pollute Texas streets and highways. Refocusing residents' attention and awareness to the pride that they feel in their state apparently has had a laudable outcome. In fact, Thaler and Sunstein contend that the motto has been one of the state's most effective anti-pollution initiatives.

Thaler and Sunstein also discuss how ATMs now routinely return a customer's bank card before dispensing cash (2008, 88). The tactic is supposed to provide a remedy for human forgetfulness. Since the customer retrieves the bank card almost automatically (without much, if any, conscious thought), the chance of losing the card should lessen. While customers may not be fully aware that their behavior is being affected, doing so seems fairly innocuous and is mainly for their own benefit. Of course, a nudging strategy can be used for many purposes beyond those that seek to uphold the well-being of an intended recipient. For example, companies often strategically construct pricing models in order to "push" a customer towards the purchase of an expensive product. The relative subtlety of nudges conjoined with the power that they can exert over decision making makes them an obvious source of ethical concern.

Tied into their argument about the usefulness of "nudges", Thaler and Sunstein describe two main types of thinking that they refer as the "Automatic System" and the "Reflective System" (2008, 19-22). This distinction is described in more depth by Norman and Shallice (1986) and has had a profound influence on the design of Arkin's robotic architectures (1990). Hybrid deliberative/reactive robotic software architectures (modeled after automatic and willed cognitive processes) now are the de facto standard for robotic system design.

Simply stated, the "Automatic System" encompasses thoughts and reactions that occur instinctively whereas the "Reflective System" includes rational thought and other related mental processes. Given the limitations of the human brain, the Automatic System may take over and guide many actions, in actuality performing most of the (relatively) unconscious operations of everyday life. The cognitive load on the human brain would likely be too demanding if we had to direct our full attention and awareness to the performance of each individual, mundane task such as pouring a glass of water or climbing a staircase (in some circumstances, doing so can even make it more difficult to complete the relevant task).

Drawing from this distinction, Thaler and Sunstein argue that the usefulness of a nudge is connected to how it can guide or bias the "Automatic System". They claim that "If people can rely on their Automatic Systems without getting into terrible trouble, their lives should be easier, better, and longer" (2008, 22). Yet one does not necessarily have to embrace Thaler and Sunstein's theory of mind in order to appreciate the value of "nudges" as a means for improving human well-being.

Paternalism and Nudges

When examining potential justifications for shaping, modifying, or otherwise influencing the behavior of rational agents, the distinction between "weak" paternalism and "strong" paternalism often emerges (Dworkin 2014). "Weak", or what is sometimes referred to as "soft", paternalism involves preventing self-harm in circumstances where it is presumed that if a person had additional knowledge or was mentally competent, a different decision would have been made. In other words, the intervention is motivated by the goal of preserving an individual's well-being while not trumping what that person presumably would have wanted if fuller access to sources of information was available.

"Strong", or what is sometimes referred to as "hard", paternalism involves implementing a decision to protect someone even if it goes against that person's voluntary choice (e.g., legally requiring a motorcyclist to wear a helmet). Advocates of strong paternalism openly acknowledge that the individuals affected by motorcycle helmet laws or other similar policies may have preferred to act differently. Yet defending this latter type of paternalism requires a greater argumentative burden since it imposes a greater intrusion on the autonomy of rational agents.[1]

Thaler and Sunstein advocate "libertarian paternalism", which they see as being a form of weak paternalism, because the intent is to uphold liberty while at the same time allow efforts that seek to mold human behavior toward more productive ends (2008, 4-6). From their perspective, "Libertarian Paternalism is a relatively weak, soft, and non-intrusive type of paternalism because choices are not blocked, fenced off, or significantly burdened" (2008, 5). However, paternalism in any of its forms has its share of harsh critics, especially those who embrace more traditional forms of libertarianism.

Nudges from Robotic Companions

There is not a definitive and obvious distinction between a robot and other electronic devices. However, Clarke (1993) suggests that "programmability", "mechanical capability", and "flexibility" can shed light on what a robot is. The "sense-think-act" paradigm is also useful as means for defining a robot (Siegel 2003). Being a robot entails that it should have at least some ability to perform self-directed behaviors. But perhaps the hallmark characteristics of embodiment and situatedness are those that are most important (Arkin 1998)[2].

[1] For the purposes of this paper, we are using the term "autonomy" in the sense of how it is normally defined within the realm of ethics (i.e., having the meaningful ability to make choices about one's life); within the realm of robotics, "autonomy" typically refers to a robot or other intelligent system making a decision without a "human in the loop".

[2] Embodiment and situatedness can of course overlap but they are two distinct concepts. Embodiment: A robot has a physical presence (a body). This spatial reality has consequences in its dynamic interactions with the world that cannot be simulated faithfully. Situatedness: A robot is an entity situated and

Numerous types of robots are already in use or under development. For example, the U.S. military uses a vast array of robots for a multitude of purposes including surveillance, bomb disposal, and security (Singer 2009). Moreover, the realm of bionics aims to integrate robotic technology into the human body primarily for medical purposes (Salvini et al. 2008). However, the focus here is on separate, physically-embodied robots; ones of the sort that would not be directly connected to the human body (such as a prosthetic device) nor would they purely exist in a virtual realm (such as a chatbot).

In principle, robots could elicit desirable traits and behaviors from the humans with which they interact. It is not hard to imagine a robot being able to do so. For example, owning a *Roomba* purportedly draws out a desire from the user to be cleaner (Sung et al. 2007). Either through overt verbal behavior, spatial proximity (proxemics), body language (kinesics), or touch, robots could display approval or disapproval of its human companion's actions, which could reinforce certain types of beliefs in the human user (Brooks and Arkin 2007). A robot's approach to accomplish its nudging goal could range from the sophisticated and subtle (such as crossing its arms and tilting its head) to the blunt and obvious (such as voicing the phrase "please stop doing that"). Along these lines, there is some evidence to indicate that a human is more likely to comply with a robot if it exhibits emotion along with a request (Moshkina 2012).

Would Robot Nudges Be Different From Other Types of Nudges?

Human behavior can be nudged in countless ways as already existing tactics for doing so clearly illustrate. Yet would "robot nudges" be different ethically or in other important respects from current nudging tactics, and if so, to what degree? Obviously, smart phones and other similar technologies are having a profound impact on how humans conduct their lives and interact with one another; technological "helps" can remind us that a meeting is upcoming, to pick up dry cleaning, or to place a device back on its charger. However, a robot possesses distinct advantages over other technological artifacts in terms of its potential to mold behavior (which may make the technology more fraught with ethical concerns). Being physically-embodied can give a robot a stronger presence in the world from the user's perspective than a virtual avatar or an app alone and thus is more likely, for better or worse, to have a lasting influence on behavior (Li 2013).

Given that they can be physically-embodied together with a human counterpart and have the capacity to move around, this opens numerous possibilities for robots to mold their surrounding environment. Robots can convey an extensive range of non-verbal messages including through gestures and posture (Brooks and Arkin 2007); they could express themselves verbally as well. More subtle cues however may actually have a greater effect on influencing human behavior than being lectured by the robot, a hypothesis that Arkin and colleagues (2014) are exploring in the context of mediating caregiver patient relationships with a robot.

If robots become genuine companions for human beings, they would interact with us in more facets of our lives than other technological artifacts. Whereas mobile phones, tablet computers, or other similar electronic devices certainly have their useful features, there are serious constraints on the level of engagement users can have with them (e.g., they are not capable of physically providing emotional support during a stressful situation, e.g., hugging or providing

---

surrounded by the real world. It does not operate upon abstract representations of reality, but rather reality itself (Arkin 1998).

physical contact such as a pat of support) . A well-designed companion robot could potentially, for example, engage in physical play such as a game of catch with a user, or charades). In other words, a broader spectrum of physical engagement, and a wider range of opportunities for shaping a human user's behavior, is possible.

Some may argue that behaviors described above are essentially what human beings do. In some sense, that is correct. What companion robots may be encoded to do could in principle closely mimic the nudging behavior that humans display with one another. However, robot nudges could be unique in comparison to nudges from a human being in at least two main ways. First, a robot's designer can control more precisely and predictably which behaviors the robot performs than the control one human can exert over another human. Second, the user of the robot, depending on the pathway encoded by the robot's designer, can selectively decide which behaviors the robot is permitted to perform, which is a level of control that humans do not fully have over their biological companions (and arguably this feature could enhance the user's liberty).

Intentionally designing a robot to act in similar ways to a human is complicated by many contextual factors. For example, it may be permissible at times for parents to nudge their child to behave more appropriately by offering a reward. But we would not typically deem it to be appropriate for a stranger to do the same thing. Analogously, whether a robot should be permitted to perform a particular "nudging" act will be contingent in part of its level of familiarity with a user. This implies that the robot would need to be sophisticated enough from a technical perspective to distinguish between different human beings and possess enough situational awareness to discern when performing certain types of behaviors is appropriate, which is already possible to some degree (Arkin et al. 2003).

Furthermore, even though robots may eventually behave similarly to humans at least to some degree, it does not necessarily follow that what we would permit in human-human interaction should be allowed within the context of HRI. One should keep in mind that merely because a particular nudging behavior is considered to be ethical when it takes place between two human beings, it does not necessarily follow that it would be ethical for a robot to perform said behavior. Tapping someone to draw that person's attention is oftentimes socially acceptable but it is largely an open question whether encoding a robot to make direct physical contact with its human counterparts is prudent or ethical. For instance, doing so might make the robot appear intrusive, threatening or the robot might even accidentally harm a person.

Nudging Humans to Become "More Ethical"

From a technical perspective, it is clearly feasible that robots could be encoded to shape, at least to some degree, a human companion's behavior by using verbal and non-verbal cues. But is it ethically appropriate to deliberately design nudging behavior in such a way so that it increases the likelihood that the human user becomes "more ethical" (however that is defined)? Robots could interact with humans in several ways in "public environments" (Salvini, Laschi, and Dario 2010, 452); and in these situations, a robot could function as a greeter to reinforce polite social etiquette as a passenger enters a train station. Yet the focus here is on private interactions and contexts.

At first glance, a plausible case could be made about the justifiability of nudging human users to perform behaviors that are to their own benefit, such as having an employer automatically pay into a retirement fund instead of relying on an employee to make use of an "opt in" system (Thaler and Sunstein 2008, 106-110). On many occasions, including selecting the option of

having an employer contribute to a retirement fund for no additional cost, humans will fail to act even if it is in their own benefit to do so. Thus, positive reinforcement, reminders, etc. may be needed as means for promoting a person's flourishing. Along these lines, a well-established design strategy encoded within technological devices is to provide prompts that aim to protect the user (e.g., beeping sounds that serve as a warning).

Assuming that a *prima facie* case can be made for the acceptability of nudging when the intent is to promote a person's own well-being, the next step is to examine conditions under which it would be permissible. Yet we sidestep that issue and wade into territory that is even ethically murkier: whether nudging a human user to perform behaviors that are primarily for the benefit of another individual or group is ethical. For instance, a robot could tap its owner in order to redirect that person's attention from completing work to a child that has been sitting alone watching television for a long period of time. Assuming that the intended beneficiary would be a second party (the child in this case), the parent might find the robot's act startling or worse. The hope might be to prevent the child from experiencing loneliness but the owner might feel offended if the robot's action is interpreted as implying he or she is a "bad" parent. Many of us will, often grudgingly, admit that some version of paternalism is necessary at times to preserve an individual's own well-being. But creating robots that encourage users to become "more ethical" in their interactions with other people grows outside the bounds of paternalism, an already controversial mode of thought.

"Positive" vs. "Negative" Nudges

Assuming for the time being that nudging humans for their own betterment is acceptable in at least some circumstances, then the next logical step is to examine what form these nudges may take. An important, but perhaps crude, distinction to draw attention to is between "positive" and "negative" nudges and whether one or both types could be considered ethically acceptable. Roughly stated, a "positive" nudge would make use of positive reinforcement tactics, such a reward or words of encouragement, in order to elicit desirable behavior. On the other hand, a design pathway that makes use of the effects of "negative" feedback, such as mannerisms or words indicating disappointment, could also be pursued.

Psychological and sociological studies can and should inform decision making in terms of which robotic design pathways are the most likely to achieve the desired result of improving a user's well-being and/or the well-being of those with whom the user interacts. However, the sheer effectiveness of a design does not necessarily answer questions about the appropriateness of pursuing that pathway; ethical considerations must be taken into account and given sufficient weight, as the user's autonomy might be constrained. Comprehensive ethical analyses must be performed to determine whether robots should be given equal latitude as a human counterpart to provide "positive" or "negative" feedback. For example, even if negative feedback could deter a user from performing self-destructive behavior (e.g., a robot after detecting smoke repeatedly states "please do not smoke cigarettes in here"), the tactic could be viewed as overly intrusive and run a fairly high risk of angering the user. Furthermore, the user may seek to avoid future interactions with the robot, which casts doubt on the effectiveness of that "negative" feedback strategy anyway.

Objections to Nudges

At least two main categories of objections arise relating to the use of nudging tactics. The first type of objection is largely philosophical in nature; simply put, deliberately manipulating rational agents is usually deemed to be problematic. A "nudging" strategy may undervalue time honored

ethical principles such as respect for persons. In other words, even if one has beneficent motives, that might not be a sufficient justification for intruding upon another person's liberty.

The second type of objection relates to how a nudging strategy could be misused in practice. More specifically, even if one is sympathetic to the notion that nudges are at times justifiable, a wide range of abuses could emerge. History has clearly shown that humans can be manipulated to perform numerous, and sometimes harmful, acts; they can be covertly pushed to vote for certain political candidates, spend more money than they should, and express intense anger against a segment of the population. Recognizing how effective nudges are, as used by marketers and others, only intensifies the depth of this type of concern.

While it can be intrusive, "nudging" is arguably justifiable in at least some circumstances given that humans are forgetful, distracted, act irrationally, and sometimes put themselves at great risk. For example, although these policies are not without controversy, many companies push their employees to exercise more frequently by using financial or other types of incentives (Pear 2013). Moreover, there is widespread agreement that nudging is generally permissible when dealing with children and those who have diminished competence or capacity. Yet intruding on a rational agent's choices is a more difficult position to defend.

Other Concerns about Robotic Nudges

Beyond the aforementioned concerns about nudging tactics, one could argue that robotic nudges may constitute a form of "moral paternalism". According to Harris, "moral paternalism refers to protection of individuals from 'corruption,' moral wickedness, or degradation of a person's character" (1977, 85). While the effort to create "better people" might be well-intended, some will find it repugnant to employ tactics that could be perceived as tampering with personal identity. This concern is especially poignant as it pertains to companion robots considering how many potential avenues they could have to alter or influence their human counterpart's behavior.

The ethics of ethical manipulation discussed in the context of companion robots overlaps to some degree with ongoing debates about "moral bioenhancement." Much of the associated discussion in that realm relates to the ethical appropriateness of using biomedical technology to try to improve human nature, foster the emergence of unique traits, or promote egalitarian aims; a related policy issue is whether these efforts should be state sponsored or not (Sparrow 2014; Persson & Savulescu 2013). Depending on the type of biomedical intervention being proposed, a robotic nudge could be less invasive because a robot's influence would be more reversible in the sense that it can be shut off. Furthermore, we would not advocate allowing governmental entities to coerce citizens into using companion robots even if that policy could result in creating "better" people.

Encoding "Ethical Nudges"

From a technical perspective, designers and others would need to evaluate the feasibility of creating robotic platforms that nudge users in the direction of becoming "more ethical". Intertwined with the technical aspects is whether the pursuit of that particular design goal is ethically appropriate.[3] Assuming that a plausible case can be made for engineering "more

[3] If constructing robots that could promote the aims of ethics, or more specifically social justice, is technically possible, a question arises about whether a moral imperative exists to build the technology (an issue that we will not seek to address here).

ethical" humans, a key question arises: which framework or theory should be used as a basis or foundation for defining what ethical means? Even if the scope was only confined to Western views, the collection of possibilities is varied and extensive; it includes rights-based approaches, deontology, consequentialism, virtue ethics, cultural relativism, and many others. The capabilities approach could also provide insight in terms of how robots could improve human well-being (Borenstein and Pearson 2013; Pearson and Borenstein 2013).

These different perspectives on ethics can and do conflict with one another in terms of which acts they would recommend or require in a given situation (e.g., whether it is appropriate to tell a lie in order to improve someone's well-being). Thus it is a non-trivial decision about which ethical framework(s) to encode; they do not merely provide equivalent answers. Yet we will not seek to wade into age-old disputes about which specific ethical framework or theory is the best supported by reason. Instead, without loss of generality, we focus attention on theories of justice in the belief that underlying principles of justice are of real value in the world. If one accepts the premise that the world would be a better place if more humans acted to promote the goals of social justice, then the next logical step is to determine how to achieve this lofty aim.[4] To frame our discussion, we examine John Rawls's views on social justice. Yet to reiterate, the application of justice principles is mainly for illustrative purposes. It can certainly be envisioned that alternative ethical frameworks could provide a guiding basis for a robot's design architecture (Wallach and Allen 2009).

Nudging Toward Social Justice

Many proposed definitions of social justice are available in the literature. While he has his share of critics (e.g., Nozick 1974; Sen 1982), we focus our attention on the Rawlsian notion of justice and the principles which he claims "free and equal persons" would embrace (1971, 209). According to Rawls, there is a set of primary goods which are "…things that every rational man is presumed to want" (1971, 214). To that end, Rawls articulates the two fundamental principles of justice. The first principle is that "each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others" (1971, 213). The second principle relates to the arrangement of "inequalities" in society (1971, 213). As he states, "social and economic inequalities, for example inequalities of wealth and authority, are just only if they result in compensating benefits for everyone, and in particular for the least advantaged members of society" (1971, 210). Within the confines of Rawls's theoretical framework, one can then ask what role companion robots could play in upholding his principles of justice.

To illustrate how robots could promote social justice, we use the example of sharing in children. Fehr et al. (2008) suggest that "inequality aversion" tends to take root in children between the ages of 3-8. The researchers claim that this psychological phenomenon tends to be linked to "parochialism"; in other words, favoritism towards the child's own social group (2008, 1080-1081). If their findings are correct, robots may be able to perform a useful function; they could nurture "inequality aversion" in young children by reinforcing proper social norms and etiquette during playtime. For example, a robotic companion could smile or display other social cues that encourage the sharing of toys between playmates. Along these lines, the robot could mimic

[4] Emphasizing the importance of the measures needed to address social justice, a United Nations committee states that "The well-being of citizens requires broad-based and sustainable economic growth, economic justice, the provision of employment opportunities, and more generally the existence of conditions for the optimal development of people as individuals and social beings" (2010, 7). Scholarly communities are just beginning to examine what role robots may have in relation to social justice concerns.

expressions of disappointment if a child refuses to share. Furthermore, the robot could nudge a child to interact with other children with whom he/she is not as used to engaging in the effort to avoid "parochialism".

It is an open question whether a robot could effectively temper what may be a natural instinct within children and other age groups towards forming cliques. According to Hyson and Taylor, "Children are more likely to develop empathy and prosocial skills if adults make it clear that they expect (but do not force) them to do so" (2011, 78). If the formation of rigid "in group/out group" structures and negative stereotypes in young children can be disrupted and this carries over to their adult lives, presumably that would be a beneficial outcome. Obviously, adults are not always successful at modeling "prosocial" behavior in part because they can have difficulty suppressing their anger or frustration. Thus, a potential advantage of a robot assisting in this effort is that it will not display negative emotions (unless of course the robot was programmed to do so). Along these lines, the lessons learned from educational contexts can play a crucial role in nurturing prosocial tendencies. Hyson and Taylor state that "Educators can promote prosocial development by building secure relationships, creating classroom community, modeling prosocial behavior, establishing prosocial expectations, and supporting families" (2011, 76). Empirical studies could evaluate whether companion robots can make a significant contribution in this regard.

It is also important to examine whether companion robots could draw out prosocial tendencies from their adult counterparts. According to Rawls, "everyone's well-being depends upon a scheme of cooperation without which no one could have a satisfactory life" (1971, 210). To that end, if adults can be encouraged ("nudged") to contribute to the betterment of their community, for example by giving to charity or participating in service activities, this could enable a greater portion of said community to participate more meaningfully as full community members. For instance, a robot could access its owner's schedule and then nudge her to be involved in adult literacy campaigns when "free time" is available or respond to an emailed emergency charitable donation request (sharing wealth) when that request is deemed legitimate. One could reasonably ask which "prosocial" behaviors or activities are appropriate and worthy of pursuit, but that is not a matter we intend to address here. Rather, our aim is to highlight the possibility that robots could serve in the capacity of trying to nurture such characteristics in human beings. If this goal is deemed to be an ethically appropriate one, then particular design strategies to actualize that goal would have to be systematically evaluated.

Exerting Control Over a Robot's Nudges

There are many possible design pathways to consider in terms of how much awareness of and control over a robot's nudges the user would have. For example, the robotic dinosaur *Pleo* cries out as though it is experiencing pain if pushed over or otherwise "mistreated", which has been shown to have a noticeable effect on human observers (Fisher 2013). Presumably, *Pleo* was designed in that manner in order to elicit certain kinds of emotional responses. Yet one can ask whether this feature of the robot's behavior should be under the user's control (e.g., if the user finds Pleo's cries to be distressing, can those sounds can be turned off?). For simplicity purposes, we will discuss three design pathways related to how much control a user could exert over a robot's nudging behavior: "opt in", "opt out", and "no way out". Potential benefits of and drawbacks to each pathway will be discussed below.

The "Opt In" Pathway

Rejection of a technological device can occur if it is perceived by the user that meaningful control over how it functions is unavailable. Correspondingly, the "opt in" pathway refers to the strategy of having users consciously and deliberately select their preferences. "Opting in" is intended to be consistent with the tenets of autonomy and respect for persons. In short, a conscious, deliberate choice would have to be rendered regarding how a robot would influence a user's behavior. If the user deems that it is acceptable for a robot to provide reminders about giving to charity or performing community service for example, then that setting in the robot's design architecture could be selected.

The "Opt Out" Pathway

The "opt out" pathway refers to the strategy of allowing the robot to perform a behavior as its default until such time as the user decides to consciously modify that setting. The rationale behind this approach is that a predetermined selection is what most users are likely to follow, and if it is an "ethical" selection, it could lead to a greater benefit to society as a whole. Humans have a psychological predilection toward accepting whatever is the original option presented to them; at times, this can even occur if that selection is not the most advantageous one for them. As Thaler and Sunstein point out, many employees fail to enroll in their company's retirement plan when a trivial amount of effort is required of them to do so even though the action would provide them with "free money" (2008, 106-110). Similarly, if less conscious effort is demanded from the user to set up reminders to take medications or complete chores for example, the user may be better off.

However, a key objection can be lodged against the "opt out" pathway; a troubling phenomenon that is not necessarily unique to robots could emerge, which Hansson refers to as "subordination to the technology" (2007, 264). As Hansson states, "There is a risk that users will feel that they are controlled by this technology, rather than using it themselves to control their surroundings" (2007, 265). Assuming this problem should be taken seriously, it would only be intensified by the "no way out" pathway described below.

The "No Way Out" Pathway

If we assume that the betterment and protection of society through the promotion of justice trumps the individual user's autonomy and rights, then perhaps there should be no "off" switch for nudging towards social justice (other than refusal to purchase or use the device in the first place). The "no way out" pathway is on display in cell phones where the user does not have the option of turning off GPS tracking with respect to the police's usage of that feature. Similar parameters, and associated rationale, exist for speed limiters on certain automobiles. The principal justification for such strategies is upholding the good of society. Correspondingly, a case could arguably be made for a companion robot having a duty to warn appropriate authorities if the performance of certain types of antisocial behaviors can reliably be anticipated. However, critics would likely offer counterbalancing perspectives such as how this type of design could intrude upon personal liberty and privacy.

The Designer's Ethical Obligations

Where would the basis for a robot's reinforcement of a human user's "good" habits come from? It must be encoded into the machine by a designer. We have seen instances of designers encoding into robots what constitutes appropriate or inappropriate behavior in the context of military robots where the goal is compliance with the laws of war (Arkin 2009; Brutzman et al. 2013). However, the paramount concern here is not the law; rather, it is morality.

An overarching ethical challenge for designers who are developing companion robots can be expressed in the following manner: does the foremost obligation that a robot possesses belong to its owner or to human society overall? The answer to this question can have a profound impact on the robot's design architecture. It overlaps with an ongoing and persistent ethical concern about whether the interests of the individual or the interests of society should be paramount when they are perceived to be in conflict.

A simplistic framework arises out of Asimov's three laws of robotics, which are provocative and useful literary devices but generally impractical (Anderson 2008). A range of attempted codes specific to roboticists have also been articulated by scholars (e.g., Murphy and Woods 2009; Riek and Howard 2014). Unfortunately, no generally accepted guidelines exist for what constitutes ethical behavior by a robot. Obtaining consensus on this matter is going to remain elusive, but actionable guidance in some form is undeniably necessary, especially considering how many different types of robots are in the process of being created and how much influence those robots may have on our lives. For example, as science fiction books and movies proactively illustrate, including *Robot & Frank*, if a robot is sophisticated enough, a user could presumably ask it to perform criminal acts or other antisocial activities that are perceived to be for the user's benefit but come at the expense of the public's welfare.

Designers should adhere to the technical codes promulgated by their professional societies. Furthermore, societies such as the IEEE (2014), ACM (1992), and NPSE (2007) provide well-intended codes of ethics, but these codes can be too general in scope to guide decision making in many situations relating to the design of robotic systems. For example, the codes do not specifically state whether the manipulation of a user's behavior is appropriate (with the possible exception of situations involving the user's safety).[5] A collection of salient ethical issues enmeshed in the design process of robots might be left up to the individual practitioner to reflect on and apply critical thinking to resolve.

Conclusion

The use of nudging tactics raises a series of serious ethical qualms, including whether they are deceptive, manipulative, or overly paternalistic. The ability to generate nudges generated by companion robots or other forms of technology is feasible and well within reach in the near- to mid-term. Our primary purpose here is to highlight ethical complexities in this realm rather than provide definitive answers about whether allowing robots to nudge a human user to become "more ethical" towards other human beings is an appropriate goal. One of the key complexities is whether and when it might be acceptable to intrude upon a robot user's autonomy in the hopes of making other people's lives better. Designers, and indeed society at large, must decide if and when reshaping human behavior through non-human artifacts is ethical, perhaps even translating into legal restrictions on the robotics community at some point in terms of what it should be allowed to create.

[5] The context we are discussing here relates more directly to professional practice (and not research environments). In the latter case, there are rules and regulations governing whether manipulation is appropriate (for example, those pertaining to IRB review and informed consent).

References

Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. AI & Society 22(4), 477-493.

Ariely, D., & Berns, G. S. (2010). Neuromarketing: The hope and hype of neuroimaging in business. Nature Reviews Neuroscience 11, 284-292.

Arkin, R.C. (1998). Behavior-based Robotics. MIT Press.

Arkin, R. (2009). Governing Lethal Behavior in Autonomous Robots. Chapman-Hall.

Arkin, R.C. (1990). Integrating behavioral, perceptual, and world knowledge in reactive navigation. In P. Maes (Ed.), Designing autonomous agents (pp. 105-122). Bradford-MIT Press.

Arkin, R., Fujita, M., Takagi, T., & Hasegawa, R. (2003). An ethological and emotional basis for human-robot interaction. *Robotics and Autonomous Systems*, 42, 191-201.

Arkin, R.C., Scheutz, M., & Tickle-Degnen, L. (2014). Preserving dignity in patient caregiver relationships using moral emotions and robots. 2014 IEEE International Symposium on Ethics in Engineering, Science and Technology, Chicago, IL.

Association for Computing Machinery (ACM). (1992). ACM Code of Ethics and Professional Conduct. http://www.acm.org/about/code-of-ethics. Accessed 24 October 2014.

Borenstein, J. & Pearson, Y. (2013). Companion robots and the emotional development of children. Law, Innovation and Technology, 5(2), 172-189.

Brooks, A., & Arkin, R.C. (2007). Behavioral overlays for non-verbal communication expression on a humanoid robot. Autonomous Robots, 22(1), 55-75.

Brutzman, D., Davis, D., Lucas Jr., G., & McGhee, R. (2013). Run-time ethics checking for autonomous unmanned vehicles: Developing a practical approach. Proceedings of the 18th International Symposium on Unmanned Untethered Submersible Technology (UUST), Portsmouth, New Hampshire.

Clarke, R. (1993). Asimov's laws of robotics: implications for information technology-part I, *Computer*, 26(12), 53-61.

Dworkin, G. (2014). Paternalism. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy http://plato.stanford.edu/archives/sum2014/entries/paternalism/. Accessed 24 October 2014.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. Nature, 454(7208), 1079-1083.

Fisher, R. 2013. Is it ok to torture or murder a robot? BBC, http://www.bbc.com/future/story/20131127-would-you-murder-a-robot. Accessed 24 October 2014.

Hansson, S. O. (2007). The ethics of enabling technology. Cambridge Quarterly of Healthcare Ethics, 16(3), 257-267.

Harris, C. E., Jr. (1977). Paternalism and the enforcement of morality. Southwestern Journal of Philosophy, 8(2), 85-93.

Hyson, M., & Taylor, J. L. (2011). Caring about caring: What adults can do to promote young children's prosocial skills. YC Young Children 66(4), 74-83.

IEEE. (2014). IEEE Code of Ethics. http://www.ieee.org/about/corporate/governance/p7-8.html. Accessed 24 October 2014.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks PNAS, 111(24), 8788-8790.

Li, J. (2013). The nature of the bots: how people respond to robots, virtual agents and humans as multimodal stimuli. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (ICMI '13) (pp. 337-340). New York, NY.

Morin, C. (2011). Neuromarketing: The new science of consumer behavior. Society, 48, 131-135.

Moshkina, L. (2012). Improving request compliance through robot affect. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2031-2037.

Moshkina, L., Park, S., Arkin, R.C., Lee, J.K., & Jung, H. (2011). TAME: Time-varying affective response for humanoid robots. International Journal of Social Robotics, 3(3), 207-221.

Murphy, R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. IEEE Intelligent Systems, 24(4), 14-20.

National Society of Professional Engineers (NSPE). (2007). NSPE Code of Ethics for Engineers. http://www.nspe.org/sites/default/files/resources/pdfs/Ethics/CodeofEthics/Code-2007-July.pdf. Accessed 24 October 2014.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R. J. Davidson, G. E. Schwartz, & D. Shapiro. (Eds.), Consciousness and self-regulation: Advances in research and theory. Plenum Press.

Nozick, R. (1974). Anarchy, state, and utopia. Basic Books.

Pear, R. (2013). Employers get leeway on health incentives. The New York Times. http://www.nytimes.com/2013/05/30/business/new-rules-give-employers-leeway-on-use-of-health-incentives.html. Accessed 24 October 2014.

Pearson, Y., & Borenstein, J. (2013). The intervention of robot caregivers and the cultivation of children's capability to play. Science and Engineering Ethics, 19(1), 123-137.

Persson, I., & Savulescu J. (2013). Getting moral enhancement right: the desirability of moral bioenhancement. Bioethics, 27(3), 124-31.

Rawls, J. (1971). A theory of justice. Cambridge, MA: Harvard University Press.

Riek, L. D., & Howard, D. (2014). A code of ethics for the human-robot interaction profession. Presented at WeRobot 2014 Conference, University of Miami.

Salter, T., Werry, I., & Michaud, F. (2008). Going into the wild in child–robot interaction studies: Issues in social robotic development. Intelligent Service Robotics, 1(2), 93-108.

Salvini, P., Datteri, E., Laschi, C., & Dario, P. (2008). Scientific models and ethical issues in hybrid bionic systems research. AI & Society, 22(3), 431-448.

Salvini, P., Laschi, C., & Dario, P. (2010). Design for acceptability: Improving robots' coexistence in human society. International Journal of Social Robotics, 2(4), 451-460.

Sen, A. (1982). Equality of what. In A. Sen, Choice, welfare and measurement (pp. 353-369). Oxford: Blackwell.

Siegel, M. (2003). The sense-think-act paradigm revisited, *1st International Workshop on Robotic Sensing* (ROSE' 03), 5.

Singer, P. W. (2009). Wired for war. New York: The Penguin Press.

Sparrow, R. (2014). Egalitarianism and moral bioenhancement. The American Journal of Bioethics, 14(4), 20-28.

Sparrow, R. (2002). The march of the robot dogs. Ethics and Information Technology, 4(4), 305-318.

Sung, J-Y, Guo, L., Grinter, R. E., & Christensen, H. I. (2007). My Roomba is rambo: Intimate home appliances. UbiComp 2007: Ubiquitous Computing. Lecture Notes in Computer Science 4717, 145-162.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. New Haven: Yale University Press.

Thomas, F. & Johnston, O. (1981). The illusion of life: Disney animation. Hyperion.

United Nations, The International Forum for Social Development. (2010). Social justice in an open world: The role of the United Nations. http://www.un.org/esa/socdev/documents/ifsd/SocialJustice.pdf. Accessed 24 October 2014.

Wallach, W., & Allen, C. (2009). Moral machines: Teaching robots right from wrong. New York: Oxford University Press, Inc.