The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots

Matthias Scheutz

1 The rise of social robots

The early 21st century is witnessing a rapid advance in *social robots*. From vacuum cleaning robots (like the Roomba), to entertainment robots (like the Pleo), to robot pets (like KittyCat), to robot dolls (like Baby Alive), to therapy robots (like Paro), and many others, social robots are rapidly finding applications in households and elder care settings. In 2006, the number of service robots world-wide alone outnum-bered industrial robots by a factor of four and this gap is expected to widen to a factor of six by 2010, only fueled by ambitious goals like those of South Korea to put one robot into each household by the year 2013 or by the Japanese expectation that the robot industry will be worth ten times the present value in 2025 (Gates, 2007).

From these expectations alone, it should be clear that social robots will soon be-come an integral part of human societies, very much like computers and the Internet in the last decade. In fact, using computer technology as an analogy, it seems likely that social robotics will follow a similar trajectory: once social robots have been fully embraced by societies, life without them will become inconceivable.

As a consequence of this societal penetration, social robots will also enter our personal lives, and that fact alone requires us to reflect on what exactly happens in our interactions with these machines. For social robots are specifically designed for *personal interactions that will involve human emotions and feelings*: "A sociable robot is able to communicate and interact with us, understand and even relate to us, in a personal way. It is a robot that is socially intelligent in a human-like way." (Breazeal, 2002) And while social robots can have benefits for humans (e.g., health benefits as demonstrated with Paro (Shibata..., 2005)), it is also possible that they could inflict harm, emotional harm, that is. And exactly herein lies the hitherto underestimated danger: the potential for *humans' emotional dependence on social robots*.

As we will see shortly, such emotional dependence on social robots is different from other human dependencies on technology (e.g., different both in kind and quality from depending on one's cell phone, wrist watch, or PDA). To be able to understand the difference and the potential ramifications of building complex social robots that are freely deployed in human societies, we have to understand how social robots are different from other related technologies and how they, as a result, can affect humans at a very basic level.

Social robots are different

Start by comparing social robots to related technologies, namely computers and industrial robots (see Table 1). These two kinds of machines are particularly relevant,

because social robots *contain computers* (for their behavior control) and share with industrial robots the property of *being robots* (in the sense of being machines with motion and/or manipulation capabilities). And both, computers and industrial robots, have been around for decades, while social robots are a recent invention.

Very much like industrial robots, social robots have the capability to initiate motion (of actuators or themselves) and thus exhibit behavior (compared to stationary objects like computers). Different from industrial robots, which are typically confined to factories, social robots are directly targeted at consumers for service purposes (like the Roomba vacuum cleaner) or for entertainment (like the AIBO robo-dog).

Very much like computers, social robots have managed to enter people's homes and thus their private lives, and are becoming increasingly part of people's daily routines (Forlizzi & DiSalvo, 2006). Different from computers, robots can interact with their owners at various levels of sophistication and they can even initiate and terminate those interactions on their own.

And unlike industrial robots and computers, social robots are often mobile and their mobility is driven by different forms of pre-programmed or learned behaviors. Even if behaviors are pre-determined and allow for very limited variability (e.g., as in various robotic toys or the Roomba), current social robots nevertheless change their position in the world. And despite the fact that these behavioral repertoires are very simple, social robots nevertheless can make (limited) decisions about what action to take or what behaviors to exhibit. They base these decisions on their perceptions of the environment and their internal states, rather than following pre-determined action sequences based on pre-programmed commands as is usually the case with robots in industrial automation (Parasuraman, Sheridan, & Wickens, 2000).

The simple rule-governed mobility of social robots, especially when robots are able to adapt and change their behaviors (e.g., by learning from experience), has far-reaching consequences. For it, as will become clear, enables robots to affect humans in very much the same way that humans are affected by animals (e.g., their pets) or even other people. In particular, it allows for and ultimately prompts humans to ascribe intentions to social robots in order to be able to make sense of their behaviors ("It did not clean in the corner because it thought it could not get there..."). The claim is that the autonomy of social robots is among the critical properties that cause people to view robots differently from other artifacts (like computers or cars).

Autonomy + mobility = perceived agency?

There are several intuitions behind applying the notion of autonomy, which has its roots in the concept of *human agency*, to artifacts like robots. These intuitions are derived from ideas about what it means for a human person to be autonomous:

> To be autonomous is to be a law to oneself; autonomous agents are self-governing agents. Most of us want to be autonomous because we want to be accountable for what we do, and because it seems that if we are not the ones calling the shots, then we cannot be accountable. (Buss, 2002)

Clearly, current robots (and those in the near future) will neither be *self-governing agents that want to be autonomous*, nor will they be in a position where they could *be accountable* or *held accountable* for their actions. This is because they will not have the necessary reflective self-awareness that is prerequisite for accountable, self-governing behavior. Yet, there is a sense in which some robots are, at least to some extent, "self-governing" and can thus be said, again in a weak sense, to be

autonomous – a robot, for example, that is capable of picking up an object at point A and dropping it off at point B without human supervision or intervention is, at least to some extent, "self-governing".

A much stronger and richer sense of autonomy, one that comes closest to the notion of human autonomy, is centered around an "agent's active use of its capabilities to pursue its goals, without intervention by any other agent in the decision-making processes used to determine how those goals should be pursued" (Barber & Martin, 1999). This notion stresses the idea of decision-making by an artificial system or agent to pursue its goals and, thus, requires the agent to at least have mechanisms for decision making and goal representations, and ideally also additional representations of other intentional states (such as desires, motives, etc.) as well as non-intentional states (such as task representations, models of other agents, etc.).

Yet, there is also an independent sense in which the autonomy of an artificial system is a matter of degrees: "For example, consider an unmanned rover. The command, 'find evidence of stratification in a rock' requires a higher level autonomy than, 'go straight 10 meters'." (Dorais, Bonasso, Kortenkamp, Pell, & Schreckenghost, 1998) The degrees or levels of autonomy can depend on several factors, e.g., how complex the commands are that it can execute, how many of its sub-systems can be controlled without human intervention, under what circumstances the system will override manual control, and the overall duration of autonomous operation (Dorais et al., 1998, see also Huang, 2004).

There is yet another dimension of robot autonomy, orthogonal to the above conceptual distinctions that focus on functional, behavioral, and architectural aspects, but of clear relevance to human-robot interactions. It is concerned with a human's perception of the (level of) autonomy of an artificial system and the impact the perceived autonomy has on the human's behavior.

The relationship among these different characterizations of robot autonomy has been summarized as a robot's "ability of sensing, perceiving, analyzing, communicating, planning, decision-making, and acting, to achieve its goals as assigned by its human operator(s) through designed human-robot interaction. Autonomy is characterized as involving levels demarcated by factors including mission complexity, environmental difficulty, and level of HRI to accomplish the missions." (Huang, 2004)

There is converging evidence that the degree of autonomy that a robot exhibits is an important factor in determining the extent to which it will be viewed as human-like, where the investigated robots are typically able to move freely, respond to commands, recognize objects, understand human speech, and make decisions (Kiesler & Hinds, 2004, Scheutz, Schermerhorn, Kramer, & Anderson, 2007a). Perceived autonomy is so critical because it implies capabilities for self-governed movement, understanding, and decision-making (Kiesler & Hinds, 2004), capabilities that together comprise important components of how we define the qualities of "humanness" or "human-like" (Friedman, Jr., & Hagman, 2003).

The distinguishing features of mobility and autonomy, therefore, set social autonomous robots apart from other types of robots, computers, and artifacts, and are ultimately a critical factor for shaping the human perceptions of autonomous robots as "social agents".

2 Evidence from HRI Studies

Over the last few years, we have conducted several human-robot interaction experiments to investigate the degree to which humans perceive robots as autonomous agents and to isolate the effects that perceived autonomy can have both on human attitudes towards robots and human behavior. To be able to gain a better understanding of people's true beliefs about robots, we developed a rigorous evaluation framework that encompasses both subjective and objective methods and measures (Rose, Scheutz, & Schermerhorn, 2010). Here we briefly summarize the results from three studies.

Study 1: Dynamic Autonomy

We investigated the extent to which robot autonomy based on independent decision making and behavior by the robot can affect the objective task performance of a mixed human-robot team while being subjectively acceptable to the human team leader (Schermerhorn & Scheutz, 2009, Scheutz & Crowell, 2007). In this task, a human subject worked together with a robot to accomplish a team goal within a given time limit. While both human and robot had tasks to perform, neither robot nor human could accomplish the team goal alone. In one of the task conditions (the "autonomy condition"), the robot was allowed to act autonomously when time was running out in an effort to complete the team goal. As part of this effort, it was able to refuse human commands that would have interfered with its plans. In the other condition (the "no autonomy condition"), the robot would never show any initiative on its own and only carry out human commands. Humans subject were tested in both conditions (without knowing anything about the conditions) and then asked to rate various properties of the robot. Overall, subjects rated the "autonomous robot" as more helpful and capable, and believed that it made its own decisions and acted like a team member. There was also evidence that they found the autonomous robot to be more cooperative, easier to interact with and less annoying than the non-autonomous robot. Surprisingly, there was no difference in the subjects' assessment of the degree to which the robot disobeyed commands (even though it clearly disobeyed commands in almost all subject runs in the autonomy condition while it never disobeyed any command in the no-autonomy condition). We concluded that subjects preferred the autonomous robot as a team partner.

Study 2: Affect Facilitation

We also investigated the utility of affect recognition and expression by the robot in a similar team task (Scheutz, Schermerhorn, Kramer, & Anderson, 2007b, Scheutz, Schermerhorn, Kramer, & Middendorff, 2006). Here, instead of making autonomous decisions, the robot always carried out human orders. However, in one condition (the "affect condition") it was allowed to express urgency in its voice or respond to sensed human stress with stress of its own (again expressed in its voice), compared to the "no-affect condition", where the robot's voice was never modulated. Each subject was exposed to only one condition and comparison were made between subject groups. The results showed that allowing the robot to express affect and respond to human affect with affect expressions of its own – in circumstances where humans would likely do the same and where affective modulations of the voice thus make intuitive

sense to humans – can significantly improve team performance, based on objective performance measures. Moreover, subjects in the "affect condition" changed their views regarding robot autonomy and robot emotions from their pre-experimental position based on their experience with the robot in the experiment. While they were neutral before the experiment as to whether robots should be allowed to act autonomously and whether robots should have emotions of their own, they were slightly in favor of both capabilities after the experiments. This is different from subjects in the no-affect group who did not change their positions as a result of the experiment. We concluded that appropriate affect expression by robot in a joint human-robot task can lead to better acceptability of robot autonomy and other human-like features like emotions in robots.

Study 3: Social Inhibition and Facilitation

While the previous two studies attempted to determine human perceptions and agreement with robot autonomy indirectly through human participation in a human-robot team task (where the types of interactions with the robot were critical for achieving the goal, and thus for the subjects' views of the robot's capabilities), the third study attempted to determine the human-likeness of the robot directly. Specifically, the study investigated people's perceptions of social presence in robots during a sequence of different interactions, where the robot functioned as a survey taker as well as an observer of human task performance (Crowell, Scheutz, Schermerhorn, & Villano, 2009, Schermerhorn, Scheutz, & Crowell, 2008). The experimental design used well-known results in psychology about social inhibition and facilitation that occurs in humans when they are observed performing tasks by other humans (Zajonc, 1965). Our experimental results showed that robots can have effects on humans and human performance that is otherwise only observed with humans. Interestingly, there was a gender difference in subjects' perception of the robot, with only males showing "social inhibition effects" caused by the presence of the robot while they were performing a math task. Post-experimental surveys confirmed that males viewed the robot as more human-like than females.

Together, the above laboratory studies provide experimental evidence about human perceptions of autonomous robots. In particular, they show that humans seem to prefer autonomous robots over non-autonomous robots when they have to work with them, that humans prefer human-like features (e.g., affect) in robots and that those features are correlated with beliefs about autonomy, and that a robot's presence can affect humans in a way that is usually only caused by the presence of another human. The question then arises whether the findings also apply to "robots in the wild", outside of the well-controlled laboratory environment. As the next section will demonstrate, there is already ample evidence for people's susceptibility to the lure of social robots outside the lab, especially when they have repeated longer-term interactions with robots.

3 The Personification of Robots

An increasing body of evidence demonstrates how humans anthropomorphize robots, project their own mentality onto them, and form what seem like deep emotional, yet unidirectional relationships with them. Documented examples, which we will summarize below, range from interviews with soldiers that worked with robots on

defusing improvised explosive devices (IEDs), to ethnographic studies with robot-pet owners (of the AIBO robot dog) and owners of the robotic Roomba vacuum cleaner.

From Garreau's: “Bots on the Ground”

The first story is about a robot developed by roboticist Mark Tilden for the purpose of defusing land mines. The robot achieves the task by way of stepping on them which causes the mine to detonate and destroy the robot’s leg. Hence, the robot was designed with several legs to be able to detonate several mines before becoming useless.

> At the Yuma Test Grounds in Arizona, the autonomous robot, 5 feet long and modeled on a stick-insect, strutted out for a live-fire test and worked beautifully, he says. Every time it found a mine, blew it up and lost a limb, it picked itself up and readjusted to move forward on its remaining legs, continuing to clear a path through the minefield. Finally it was down to one leg. Still, it pulled itself forward. Tilden was ecstatic. The machine was working splendidly. The human in command of the exercise, however – an Army colonel – blew a fuse.
> The colonel ordered the test stopped. Why? asked Tilden. What’s wrong? The colonel just could not stand the pathos of watching the burned, scarred and crippled machine drag itself forward on its last leg. This test, he charged, was inhumane.(Garreau, 2007)

Whether or not “inhumane” was an appropriate attribution, the fact remains that the only explanation for not wanting to watch a mindless, lifeless machine, purposefully developed for blowing up mines, destroy itself, is that the human projected some agency onto the robot, ascribing to it some inner life, and possibly even feelings.

Another example, recounted by a Marine sergeant running a robot repair shop in Iraq, is the technician who returned his IED defusing robot which he had named “Scooby-Doo” for repair. While it is well-known that humans have a tendency to name inanimate things they like and/or use frequently (e.g., their car), naming comes at a price: it automatically generates a kind of intimacy with and connectedness to the named object. And in the case of robots it only re-enforcers what the self-propelled behavior of a robot already does: prompting the inscription of intentionality into an artifact and thus impliciting granting it agency!

> “There wasn’t a whole lot left of Scooby,” Bogosh says. The biggest piece was its 3-by-3-by-4-inch head, containing its video camera. On the side had been painted “its battle list, its track record. This had been a really great robot.” The veteran explosives technician looming over Bogosh was visibly upset. He insisted he did not want a new robot. He wanted Scooby-Doo back. “Sometimes they get a little emotional over it,” Bogosh says. “Like having a pet dog. It attacks the IEDs, comes back, and attacks again. It becomes part of the team, gets a name. They get upset when anything happens to one of the team. They identify with the little robot quickly. They count on it a lot in a mission.” (Garreau, 2007)

In fact, soldiers take pictures of their robots, introduce robots to their friends and family abroad, and even promote them, all indications of treating robots as if they were intentional creatures.

> "When we first got there, our robot, his name was Frankenstein" says Sgt. Orlando Nieves, an EOD from Brooklyn. "He'd been in a couple of explosions and he was made of pieces and parts from other robots." Not only did the troops promote him to private first class, they awarded him an EOD badge – a coveted honor. "It was a big deal. He was part of our team, one of us. He did feel like family." (Garreau, 2007)

Robot dogs are pets too

Even if the above examples seem hardly believable, one might be lenient and justify the soldiers' attribution of human qualities to robots by pointing to the extraordinary circumstances that these soldiers encounter in combat and the huge emotional toll it takes on the human psyche. But surprisingly, being in a deserted remote location dealing with life-threatening situations is not necessary to elicit the kinds of reactions to robots we saw with soliders in Iraq. Ordinary citizens living in the US seem to fall prey to suggestive behaviors of social robots. For example, Peter Kahn and colleagues (Peter H. Kahn, Friedman, & Hagman, 2002) examined the postings of users in AIBO news groups, where robo-dog owners share their experiences with AIBO freely, and identified four categories of postings:

> *Essences* refer to the presence or absence of technological, biological, or animistic underpinnings of AIBO (e.g., "He's resting his eyes"). *Agency* refers to the presence or absence of mental states for AIBO, such as intentions, feelings, and psychological characteristics (e.g., "He has woken in the night very sad and distressed"). *Social standing* refers to ways in which AIBO does or does not engage in social interactions, such as communication, emotional connection, and companionship (e.g., "I care about him as a pal, not as a cool piece of technology"). *Moral standing* refers to ways in which AIBO may or may not engender moral regard, be morally responsible, be blameworthy, have rights or deserve respect (e.g., "I actually felt sad and guilty for causing him pain!"). (Peter H. Kahn et al., 2002)

While they found relatively few references to AIBO's moral standing (12%), people made very frequent references to essences (79%), agency (60%), and social standing (59%). It seems clear that AIBO owners have a strong tendency to form (false) beliefs about (possible) mental states of their robots.

Even the Roomba does the trick

Another example group are owners of Roomba vacuum cleaners that have been interviewed in a variety of studies over the last several years, given that the Roomba is one of the most widely sold autonomous robots. While at first glance it would seem that the Roomba has no social dimension (neither in its design nor in its behavior) that could trigger people's social emotions, it turns out that humans, over time, develop a strong sense of gratitude towards the Roomba for cleaning their home. The mere fact that an autonomous machine keeps working for them day in day out seems to evoke a sense of, if not urge for, reciprocation. Roomba owners seem to want to do something nice for their Roombas even though the robot does not even know that it has owners (it treats humans as obstacles in the same way it treats chairs, tables and

other objects that it avoids while driving and cleaning)! The sheer range of human responses is mind blowing (e.g., see Sung, Guo, Grinter, & Christensen, 2007). Some will clean for the Roomba, so that it can get a rest, while others will introduce their Roomba to their parents, or bring it along when they travel because they managed to developed a (unidirectional) relationship: “I can’t imagine not having him any longer. He’s my BABY! ! ... When I write emails about him which I’ve done that as well, I just like him, I call him Roomba baby... He’s a sweetie.” (Sung et al., 2007).

Not even experienced roboticists are always spared

Somewhat surprisingly, it is even possible for an experienced roboticist to be affected by the suggestive force of apparent autonomous behavior. In our own lab, for example, we found our humanoid robot CRAMER disturbing when it was left on (by accident) and started shifting attention from speaker to speaker (as if it understood what was being said). And, according to Garreau, graduate students at MIT working in the lab with the Kismet robot put up a curtain between themselves and the robot at times because the robot’s gaze was breaking their concentration. In fact, even the creator of Kismet, Cynthia Breazeal, seems to have developed a very personal relationship with her own creation:

> Breazeal experienced what might be called a maternal connection to Kismet; she certainly describes a sense of connection with it as more than “mere” machine. When she graduated from MIT and left the AI Laboratory where she had done her doctoral research, the tradition of academic property rights demanded that Kismet be left behind in the laboratory that had paid for its development. What she left behind was the robot ‘head’ and its attendant software. Breazeal described a sharp sense of loss. (Turkle, 2006)

4 The dangers ahead

The above is only a small set of the ever-mounting evidence that humans are becoming increasingly attached to robots. From seemingly innocuous facts such as the naming of their robots, to more worrisome episodes such promoting robots to military ranks, calling robots “pals”, and exhibiting “shameful” reactions (such as the woman who shut her bedroom door because she was getting undressed and felt that her AIBO was watching her), the personification of social robots is widespread and is becoming a testimony for the human willingness to form unidirectional emotional bonds with these machines.

It is important in this context to note how *little* is required on the robotic side to cause people to form relationship with robots. Consider the case of the AIBO. Clearly, it is modelled after a real dog in that its physical shape resembles that of a dog and its behaviors bear some resemblance to dog behaviors (wagging tail, barking, etc.). Hence, one might argue that it is really a robotic substitute for what otherwise would be legitimate companion. But then, consider the PackBot, which is not even a fully autonomous robot; rather it is under tight remote control from its operator. Moreover, it has tracks and does not resemble any particular biological creature. Yet, it does play a critical role in the soldiers’ daily routines and fight for survival. Hence, one might

argue that these special circumstances make humans forget the very machine-like appearance and lack of autonomy of PackBot. And PackBot has another unique feature that might contribute to the soldier's identification with the robot: soldiers are able to see the world from the robot's perspective (through visual real-time streams from the robot's cameras). This could easily blur the distinction between the robot itself and the human operating it, at least for the human operator (there is evidence from cognitive science that humans view sensory or actuator augmentations as part of their bodies when they have gained sufficient experience using them).

For further contrast, consider now the Roomba, which neither has animal-like appearance, nor allows the human to see the world from its perspective. It is a mere disc that drives around in certain patterns avoiding to bump into things. Yet, it manages to instill the idea of agency in people, and can cause them to even experience gratitude for its service, so much so that they will clean in its stead. One would hardly be able to make that point for dish washers!

It is also interesting to note how *little* these robots have to contribute on their end to any relationship, i.e., how inept and incapable they are to partake as a genuine partner: neither the Roomba nor the PackBot, for example, have any notion of "other"; there are no built-in algorithms for detecting and recognizing people. Rather, anything that causes their contact sensors to be triggered is treated in the same way, namely as an "obstacle" that needs to be avoided.

The false pretense: robots are agents

None of the social robots available for purchase today (or in the foreseeable future, for that matter) *care* about humans, simply because they *cannot* care. I.e., these robots do not have the architectural and computational mechanisms that would allow them to care, largely because we do not even know what it takes, computationally, for a system to care about anything (cp. to Haugeland, 2002). Yet, this fact is clearly getting lost in the increasing hype about social robots. It almost seems as if industry is trying hard to make the case for the opposite, thus enforcing the personification of social robots.

Take, for example, one of the new Hasbro robot dolls, called "Baby Alive", which can say simple phrases like "I'm hungry", "Oh oh, I made a stinky", and "Mommy, I love you". The commercial advertising for the robot emphasizes "how real it is" by explicitly using the phrase "a baby so real". Other companies have been advertising their toys as "recreating the emotions" of a cat, a dog, an infant, etc. (see also Scheutz, 2002).

Even companies like I-Robot that are clearly aware of the computational and cognitive limitations of their products, find it useful, for whatever reason, to create a Facebook page for their PackBot product, where PackBot stories and news are recounted in first person narratives as if there were a single entity called "PackBot" that had experienced all these situations and events.

And finally, academics themselves are often less careful than they ought to be when presenting their research. For example, researchers who work on emotions often say loosely that their robots *have emotions*, *implement emotions*, *use emotions*, etc. This kind of suggestive language (e.g., during research presentations or even in published research papers) makes it easy for non-expert readers to conflate the control processes in these artifacts with similarly labeled, yet substantively very different control processes in natural organisms, particularly humans (e.g., see Scheutz, 2002).

The repeated labeling of control states in robotic architectures and of behaviors exhibited by robots with terms familiar from human and animal psychology helps to create, maintain, and sustain the false belief that “somebody is at home” in current robots. And while people, when asked explicitly, might deny that they think of the robot as a person, an animal, or an otherwise alive agent, this response generated at the conscious level might be forgotten at the subconscious level at which robots can affect humans so deeply. Social robots are clearly able to push our “Darwinian buttons”, those mechanisms that evolution produced in our social brains to cope with the dynamics and complexities of social groups; mechanisms, that automatically trigger inferences about other agents’ mental states, beliefs, desires, and intentions.

The potential for abuse

The fact alone that humans are already anthropomorphizing existing social robots in ways that clearly overstate the robots’ capabilities, is a sufficient indication that the personification of social robots is moving forward quickly, and that more sophisticated future robots will likely be even more anthropomorphized. Features of future robots like human-like appearance, natural language interactions, etc. might prompt people to be even more trusting in them or develop attitudes towards robots that could and likely would be exploited. For example, if it turns out that humans are reliably more truthful with robots than they are with other humans, it will only be a matter of time before robots will interrogate humans. And if it turns out that robots are generally more believable than humans, then it will only be a matter of time before robots are used as sales representatives.

Moreover, it will become even easier and more natural for humans to establish unidirectional emotional bonds with more sophisticated robots, often without noticing, akin to becoming addicted, where one’s realization of one’s addiction always comes after the fact. And with more sophisticated robots that are specifically programmed to exhibit behavior that could easily be misinterpreted as showing social emotions such as sympathy and empathy, it will become increasingly difficult for people to even realize that their social emotional bonds are unidirectional, aside from a basic emotional resistance that we are already seeing today (e.g., when people insist that they get back the very same robot that they sent in for repair and not another copy).

What is so dangerous about unidirectional emotional bonds is that they create psychological dependencies that could have serious consequences for human societies, because they can be exploited at a large scale. For example, social robots that appear “lovable” might be able to get people to perform actions that the very same people would not have performed otherwise, simply by threatening to end their relation with the human (e.g., an admittedly futuristic sounding request of a robo-dog to dispose of a real dog: “Please get rid of this animal, he is scaring me, I don’t want him around any longer.”). More importantly, social robots that cause people to establish emotional bonds with them, and trust them deeply as a result, could be misused to manipulate people in ways that were not possible before. For example, a company might exploit the robot’s unique relationship with its owner to make the robot convince the owner to purchase products they wish to promote. Note that different from human relationships where, under normal circumstances, social emotional mechanisms like empathy, guilt, and others would prevent the escalation of such scenarios, there does not have to be anything on the robots’ side to stop them

from abusing their influence over their owners.

5 We need to act, now!

Despite our best intentions to build useful robots for society, making the case for robo-soldiers, robo-pets, robo-nurses, robo-therapists, robo-companions, and so forth, current and even more so future robot technology poses a serious threat to humanity. And while there is clearly a huge potential for robots to do a lot of good for humans (from elder care, to applications in therapy), any potential good cannot be discussed without reflecting any potentially detrimental consequences of allowing machines to enter our personal social emotional lives.

Some have warned us for quite some time about the dangers of producing increasingly human-like robots:

> "It is also practically important to avoid making robots that are reasonable targets for either human sympathy or dislike. If robots are visibly sad, bored or angry, humans, starting with children, will react to them as persons. Then they would very likely come to occupy some status in human society. Human society is complicated enough already." (McCarthy, 1995)

Yet, it is clear that, as a research community, the fields of artificial intelligence, robotics, and the nascent field of human-robot interaction have not reflected enough on the social and ethical implications of their artifacts. Such a reflection, if considered soon enough, might be able to inform future robotics research in useful ways, for example, on how research should proceed with respect to questions such as the slowly crystallizing perspective of future robotic soldiers (Moshkina & Arkin, 2007) or robotic sex partners (Levy, 2007).

Different from the first discussions about robot consciousness and robot rights in the 1960s, where philosophers thought it opportune to begin reflecting on these subjects, since the existence of such robots was still far off (Putnam, 1964), we are now running out of time. We need to start right away to investigate the potential dangers of social robots, find ways to mitigate them, and possibly develop principles that future law-makers can use to impose clear restrictions on the types of social robots that can be deployed.

For example, one could simply prohibit and stop all research and development on social robots. While this option would certainly solve some of the problems, by avoiding them altogether, it seems completely unreasonable to believe that research and development of social robots could be prohibited and stopped, while other research in robotics and artificial intelligence continuous.

Another option might be to require, by law, that all commercially available robots have some form ethical reasoning built in. For example, some have argued that ethical principles will need to be integrated into the decision-making algorithms in the robotic architecture in such a way that the robot will not be able to alter, ignore, or turn off these mechanisms (e.g., Arkin, 2009). While this option might work for limited domains, where the number of possible actions is clearly constrained and the ethical implications of all actions can be determined ahead of time, it is unclear how general ethical principles could be devised that would work for an unknown number of situations, largely because philosophy in all of its history has not been able to agree on the right set of universal ethical principles, aside from being computationally feasible in real-time given the computational constraints of the robotic platform. Even

if there were a way to encode ethics in a set of universal laws, very much like Asimov conceived of the *Three Laws of Robotics* (in his short story "Runaround" from 1942), there are strong logical reasons why such as system cannot work – it would be straightforward to present a robot with logical paradoxes that would render any rational reasoning system ineffective, e.g., by ordering it to "not obey any orders including this one", an order that, by simply stating it, automatically makes the robot disobedient no matter how sophisticated its control system may be.

Another option might be, again required by law, to make it part of a social robot's design, appearance and behavior, that the robot continuously signal, unmistakenly and clearly, to the human that it is a machine, that it does not have emotions, that it cannot reciprocate (very similar to the "Smoking kills" labels on European cigarette packs). Of course, these reminder that robots are machines are no guarantee that people will not fall for them, but it might reduce the likelihood and extent to which people will form emotional bonds with robots. And it will present the challenge of walking a fine line between making interactions with robots easier and more natural, while clearly instilling in humans the belief that robots are man-made machines with no internal life (at least the present ones). It is currently unclear how effective such mechanisms could be although empirically testing their effectiveness would be straightforward (e.g., add a particular mechanisms to a particular generation of Roombas, repeat the previous ethnographic studies and compare the extent to which people engage in the same behaviors as before).

In the end, what we need is a way to ensure that robots will not be able to manipulate us in ways that would not be possible for another (normal) human beings. And a radical step might be necessary to achieve this: to endow future robots with human-like emotions and feelings. Specifically, we need to do for robots what evolution did for us, namely to equip us with an emotional system that strikes a balance between individual well-being and socially acceptable behavior. By having the same "unalterable affective evaluation" as those realized in humans, future social robots will be able to function in human societies in human-like ways (for all the reasons we are now investigating in HRI and AI/robotics), with the side-effect of having "genuine feelings" that make them just as vulnerable and manipulable as humans.

Some have voiced their reservations about endowing robots with emotions arguing that it would take extra effort to implement human-like emotions in robots (e.g., McCarthy, 1995), while others have maintained that certain types of emotions will necessarily be possible (and even instantiated) in complex robotic architectures with particular architectural properties) (Sloman & Croucher, 1981). Without taking a stance on whether emotions have to be explicitly built in or result as emergent phenomena in certain types of architectures, it is important to appreciate that this suggestion does not apply to *any* type of robot, but only to *certain types of social robots*. We certainly do not need a space exploration robot to be emotional, and nobody would step foot on a plane with an automatic flight controller that can get depressed, if not suicidal. However, if we had a choice between a Terminator 3-type scenario, where intelligent robots take control, despite human efforts to prevent it, and a grouchy household robot that is tired of cleaning up the kitchen floor, the choice is obvious.

References

Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. Chapman and Hall.

Barber, K., & Martin, C. (1999, May). Specification, measurement, and adjustment of agent autonomy: Theory and implementation. *Autonomous Agents and Multi-Agent Systems*.

Breazeal, C. L. (2002). *Designing sociable robots*. MIT Press.

Buss, S. (2002). Personal autonomy. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2002 ed.). Stanford University.

Crowell, C., Scheutz, M., Schermerhorn, P., & Villano, M. (2009). Gendered voice and robot entities: Perceptions and reactions of male and female subjects. In *Iros.*

Dorais, G., Bonasso, R. P., Kortenkamp, D., Pell, B., & Schreckenghost, D. (1998, August). Adjustable autonomy for human-centered autonomous systems on mars. In *Mars society conference.*

Forlizzi, J., & DiSalvo, C. (2006). Assistive robots and domestic environments: A study of the roomba vacuum in the home. In *Proceedings of the 1st acm/ieee human-robot interaction conference hri06.*

Friedman, B., Jr., P. H. K., & Hagman, J. (2003). Hardware companions? : what online aibo discussion forums reveal about the human-robotic relationship. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 273–280).

Garreau, J. (2007, May 6). Bots on the ground in the field of battle (or even above it), robots are a soldier's best friend. *Washington Post*.

Gates, B. (2007). A robot in every home. *Scientific American*, *January*.

Haugeland, J. (2002). Computationalism: New directions. In M. Scheutz (Ed.), (pp. 159–174). MIT press.

Huang, HM. (Ed.). (2004). *Autonomy levels for unmanned systems (alfus) framework, volume i: Terminology*. National Institute of Standards and Technology.

Kiesler, S., & Hinds, P. (2004). Introduction to the special issue on human-robot interaction. human-computer interaction. *Human-Computer Interaction*, *19*, 1–8.

Levy, D. (2007). *Love and sex with robots: The evolution of human-robot relationships*. Harper.

McCarthy, J. (1995). Making robots conscious of their mental state. In *Proceedings of machine intelligence workshop* (pp. 89–96). Oxford: . (Accessible via http://www-formal.stanford.edu/jmc/consciousness.html)

Moshkina, L., & Arkin, R. (2007). *Lethality and autonomous systems: Survey design and results* (Tech. Rep.). : Georgia Tech.

Parasuraman, R., Sheridan, T., & Wickens, C. (2000, May). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Sytems and Humans*, *30*(3), 286-297.

Peter H. Kahn, J., Friedman, B., & Hagman, J. (2002). 'i care about him as a pal': Conceptions of robotic pets in online aibo discussion forums. In *Chi*.

Putnam, H. (1964). Robots: Machines or artificially created life? *The Journal of Philosophy*, 668–691.

Rose, R., Scheutz, M., & Schermerhorn, P. (2010). Towards a conceptual and

methodological framework for determining robot believability. *Interaction Studies*.

Schermerhorn, P., & Scheutz, M. (2009). Dynamic robot autonomy: Investigating the effects of robot decision-making in a human-robot team task. In *The* eleventh international conference on multimodal interfaces and workshop on machine learning for multi-modal interaction *icmi-mlmi*.

Schermerhorn, P., Scheutz, M., & Crowell, C. (2008). Robot social presence and gender: Do females view robots differently than males? In *3rd ACM/IEEE international conference on human-robot interactions* (p. forthcoming).

Scheutz, M. (2002). Agents with or without emotions? In R. Weber (Ed.), *Proceedings of the 15th international flairs conference* (pp. 89–94). AAAI Press.

Scheutz, M., & Crowell, C. (2007). The burden of embodied autonomy: Some reflections on the social and ethical implications of autonomous robots. In *Proceedings of icra 2007 workshop on roboethics.*

Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007a). First steps toward natural human-like hri. *Autonomous Robots*, forthcoming.

Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007b, May). First steps toward natural human-like HRI. *Autonomous Robots*, *22*(4), 411–423.

Scheutz, M., Schermerhorn, P., Kramer, J., & Middendorff, C. (2006). The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st acm international conference on human-robot interaction* (pp. 226–233).

Shibata..., T. (2005). Human interactive robot for pscyhological enrichment and therapy. In *Aisb'05 social intelligence and interaction in animals, robots, and agents*.

Sloman, A., & Croucher, M. (1981). Why robots will have emotions. In *Proc 7th int. joint conference on AI* (pp. 197–202). Vancouver: .

Sung, JY., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). 'my roomba is rambo': Intimate home appliances. In *Ubicomp 2007* (p. 145- 162). Springer.

Turkle, S. (2006, Jul). *A Nascent Robotics Culture: New Complicities for Companionship* (Tech. Rep.). : AAAI.

Zajonc, R. B. (1965). Social facilitation. *Science*, *149*, 269–274.