Proxy Prudence: Rethinking Models of Responsibility for Semi-autonomous Robots

Jason Millar[1]

jason.millar@queensu.ca

Risks and opportunities often mingle, especially when the risks associated with developing a new class of products are uncertain—when they are difficult to predict, risks can chill opportunity. Managing risks in robotics is complicated by the increasing autonomy of robotic systems. As we program robots to behave in more sophisticated ways, with humans further "removed from the control loop" yet interacting with the robots, it becomes more difficult to determine responsibility for a robot's behaviors. We need to develop theories and frameworks of responsibility for dealing with semi-autonomous robots so we can manage risks and maintain a healthy robotics community, but also for the sake of properly accounting for responsibility when deploying robots in society.

The complicated link between product failures and responsibility has roots in our intuition. When products break we intuitively turn to the products' designers when doling out responsibility since they make the things that break. A bolt whose failure results in harms can be causally linked to the design of the product with relative ease: had the bolt been designed differently it would not have failed. We are used to dealing with responsibility in cases of this kind of failure, where the failure is an undesirable event that can be identified in a relatively clear causal chain, and where we can say that the design feature could, or should, have been otherwise.

Our intuitions are far less helpful when anticipating some of the foreseeable effects of semi-autonomous robots. Consider the following moral dilemma that has been framed in the context of self-driving cars (SDCs).[2] A person is travelling along in her SDC at a speed of 100 km/h (~60 mph). Suddenly a child errantly runs from a busy sidewalk into the street forcing the car to make a quick decision, it can either hit the child and likely kill it, swerve into an oncoming truck likely killing the driver[3] alone,

[1] I wish to thank Ian Kerr and Sergio Sismondo for their considerable help in reviewing drafts of this paper. Elena Ponte also deserves thanks for helping me to sort out the basics of strict liability as it applies to robots.

[2] This is a version of the classic trolley problem in philosophical ethics, which Patrick Lin and others have more recently discussed in the context of driverless cars. For example see Lin, P. (2013). "The Ethics of Saving Lives With Autonomous Cars Are Far Murkier Than You Think." *Wired*. (July 30). Last accessed Nov., 4 2013 at: http://www.wired.com/opinion/2013/07/the-surprising-ethics-of-robot-cars/

[3] I refer here to the person in the car as the driver only to anchor this argument using traditional notions of driving—there is someone in the car who would otherwise be holding the steering wheel. Of course, the person in the car is not driving, and might better be referred to as the passenger or user of the car. For the purpose of this argument I will use the terms interchangeably recognizing that they

or swerve off the road and possibly kill several innocent pedestrians. In this case there is no objectively "right" answer what the car ought to do, only unfortunate outcomes. Regardless of the decision, sorting out responsibility following the outcome of this situation is complicated by the presence of a technology that is relatively autonomous by design. With a traditional car, tracing the causal chain would lead us to think that the driver would be held at least partially, if not wholly responsible for the outcome. But in the case of an SDC the causal chain analysis appears to remove the driver from the equation, shifting our focus to either the car or the designers for having caused the outcome. Assigning responsibility to the car is unsatisfying at best, morally troubling at worst. Yet it also seems problematic to assign too much responsibility to the car's designers who are so far removed from the context of the accident. How do we gain traction on sorting out responsibility in these types of circumstances?

This paper explores the question of responsibility from both philosophical and legal perspectives, by examining the relationship between designers, semi-autonomous robots and users. Borrowing concepts from the philosophy of technology, bioethics and law, I argue that in certain use contexts we can reasonably describe a robot as acting as a *moral proxy* on behalf of some person. If semi-autonomous robots function as moral proxies it is important (and useful) to instantiate the proxy relationship in a morally justifiable way. I examine two questions that are helpful in determining how to appropriately instantiate proxy relationships with semi-autonomous robots, and that we can therefore ask when attempting to sort out responsibility: 1) On whose behalf was the robot acting?; and 2) On whose behalf ought the robot to have been acting?

Focusing on proxy relationships allows us to shift our focus away from a strictly causal model of responsibility and focus also on a *proxy model* informed by an ethical analysis of the nature of the designer-artefact-user relationship. In doing so I argue that we gain some traction on problems of responsibility with semi-autonomous robots.

I examine two cases to demonstrate how a shift towards a proxy model of responsibility, and away from a strictly causal model of responsibility helps to manage risks and provides a more accurate accounting of responsibility in some use contexts. I offer some suggestions how we might decide whom a robot ought legitimately to be acting on behalf of, while offering some thoughts on what legal and ethical implications my argument carries for designers and users.

Semi-Autonomous Technology

The idea of fully autonomous robots, person-like humanoid robots for example, has captured the imagination and interest of philosophers, academics, artists,

---

will likely carry quite different legal and moral implications once SDCs are widespread.

storytellers and the public over the years.[4] More recently, it has become popular to talk of the coming age of "spiritual machines"[5], or those machines capable of human-like intelligence, and the legal and philosophical implications they will carry.[6] In this vein it is not uncommon to come across articles in academic journals and the popular media featuring some version of the term "robot rights" in their titles, or exploring the idea of treating robots as persons in the ethical and legal sense, the implication being just what you would think, when technologies are capable of human-like intelligence we will find ourselves having to consider giving them rights.[7] Along with rights comes responsibility; rights-holding agents capable of making autonomous decisions are held accountable for their actions. Somewhere down the road it could be that your phone, or perhaps your n^{th} generation ultra smart Roomba vacuum will be morally punished and/or praised for its actions. Science fiction? Yes, for the time being. But if the day comes that we cannot tell the difference between robots (machines) and humans in terms of intelligence, and it is reasonable to think it could, we will at least have laid the theoretical groundwork for sorting some of the tough philosophical questions out. In the meantime we are faced with a less fanciful but no less complicated challenge: how to sort out responsibility in situations concerning todays robots, the far less sophisticated ancestors of those potential future rights holders.

---

[4] Aristotle. *Politics*. Book I, Chapter 4; Čapek, K. (1920). *R.U.R. (Rossum's Universal Robots)*. Online at: http://etext.library.adelaide.edu.au/c/capek/karel/rur/complete.html.; Terminator (1984). Orion Pictures.; 2001: A Space Oddyssey (1968). Warner Bros.; Asimov, I. (2004). *I, Robot.* (New York: Bantam Dell).

[5] Kurzweil, R. (2000). *The Age of Spiritual Machines.* (New York: Penguin).

[6] Capek's *Rossum's Universal Robots* featured robots seemingly capable of human-like intelligence. Also see Lin, P., Abney, K., Bekey, G. (2011). "Robot Ethics: Mapping the Issue for a Mechanized World." *Artificial Intelligence 175*(5-6):942-949.; Calo, R. (2010). "Robots and Privacy." In P. Lin, G. Bekey, and K. Abney (Eds.). *Robot Ethics: The Ethical and Social Implications of Robotics.* (Cambridge, Mass: MIT Press).; Petersen, S. "Designing People to Serve." In P. Lin, G. Bekey, and K. Abney (Eds.). *Robot Ethics: The Ethical and Social Implications of Robotics.* (Cambridge, Mass: MIT Press).; Robertson, J. (2010). "Gendering Humanoid Robots: Robo-sexism in Japan." *Body & Society 16*(2):1-36.

[7] Freitas, R. (1985). "The Legal Rights of Robots". *Student* Lawyer 13:54-56; Leiber, J. (1985). *Can Animals and Machines Be Persons? – A Dialogue*. (Hackett Publishing Company).; Sparrow, R. (2011). "Can Machines be People? Reflections on the Turing Triage Test." In P. Lin, G. Bekey, and K. Abney (Eds.). *Robot Ethics: The Ethical and Social Implications of Robotics.* (Cambridge, Mass: MIT Press).; Davoudi, D. (2006). "UK Report Says Robots Will Have Rights". *Ft.com.* (December 19). Online at: http://www.ft.com/cms/s/2/5ae9b434-8f8e-11db-9ba3-0000779e2340.html#axzz2vkLhYzJw.; Levy, D. (2009). "The Ethical Treatment of Artificially Conscious Robots." *International Journal of Social Robotics* 1:209 – 216.; Darling, K. (2012). "Extending Legal Rights to Social Robots". *We Robot 2012 Conference Proceedings.* Online at: http://ssrn.com/abstract=2044797.

We are increasingly interacting with technologies capable of a relatively high degree of autonomy. SDCs, for example, are capable of much more sophisticated decision-making in the use context than are hammers. Perhaps the best-known example of an SDC is the one under development at Google.[8] Google's SDC uses a series of sensors, digital maps, databases, and software to solve the extremely complex problem of driving.[9] To date, Google's SDCs have logged hundreds of thousands of kilometers driving autonomously in regular traffic, with only the occasional need for human intervention.[10] In 2011 the state of Nevada was the first to pass legislation authorizing the licensing of SDCs for the state's roads, a law that went into effect early in 2012.[11] Florida and California followed close on Nevada's heels, a move described as "turning today's science fiction into tomorrow's reality."[12] According to Google, their SDCs are safe: "there hasn't been a single accident under computer control."[13] Many of the major auto manufacturers are developing SDCs, including Volkswagen, Audi, General Motors, and Daimler-Benz, while almost every auto manufacturer is implementing semi-autonomous features such as parallel parking and collision avoidance systems. It is expected that SDCs will be on the market within the next decade, while some predict the SDC will dominate the roads by 2040.[14]

In the case of the errant child pedestrian outlined above, the SDC contributes something to the outcome that the human driver does not, it detects the situation and makes the decision to either continue straight or swerve. Wendell Wallach and Colin Allen describe this kind of autonomy as falling on a continuum with things like

---

[8] Vanderbilt, T. (2012). "Let the Robot Drive." *WIRED*. (Feb):86.

[9] Guizzo, E. (2011). "How Google's Self-Driving Car Works." *IEEE Spectrum.* (Oct. 18). (Last accessed November 29, 2012: http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/how-google-self-driving-car-works).

[10] Ibid.

[11] Slosson, M. (2012). "Google Gets First Self-Driven Car License in Nevada." *Reuters.com*. (May 8). (last accessed November 29, 2012: http://www.reuters.com/article/2012/05/08/uk-usa-nevada-google-idUSLNE84701320120508).

[12] Hayden, E.. (2012). "Speeding into the Future: Self-Driving Cars Are Now Legal in California." *Time: Newsfeed.* (Sept. 26). (Last accessed November 29, 2012: http://newsfeed.time.com/2012/09/26/speeding-into-the-future-self-driving-cars-are-now-legal-in-california/).

[13] Ibid.

[14] Newcomb, D. (2012). "You Won't Need a Driver's License by 2040." *Wired.com: Autopia.* (Sept. 17). (Last accessed February 4, 2014: *http://www.wired.com/autopia/2012/09/ieee-autonomous-2040/*). Also see Laursen, L. (2014). "Self-Driving Car Rules Will Lag Tech, Think Tanks Predict." *IEEE Spectrum.* (Jan. 9). (Last accessed February 4, 2014: *http://spectrum.ieee.org/tech-talk/green-tech/advanced-cars/selfdriving-car-rules-will-lag-tech-predict-think-tanks*).

hammers at the low end, and fully autonomous moral agents at the high end. They describe the middle of the continuum as consisting of machines not with autonomy in the full-blown moral sense of the term, the kind that would warrant attributing rights and responsibilities to the machine, rather in the middle we find at work a colloquial notion of autonomy, or semi-autonomy.[15] Put simply, in the use context the SDC acts without direct human interaction, so it is considered autonomous in a way that hammers are not. For the remainder of the paper I will borrow this colloquial notion of autonomy when using it in relation to technologies—the machines, software, and robots I am interested in discussing are semi-autonomous. They are neither hammers (mere tools) nor human-like in their intelligence. It is this notion of semi-autonomy that we must account for in our theories of responsibility if we are to distinguish between responsibility in cases of full-blown moral autonomy, cases of using mere tools such as hammers, and cases of using semi-autonomous machines like SDCs. We face this challenge because we are presently populating society with semi-autonomous robots.

Moral Proxies in Healthcare

In healthcare, the ideal patient qua decision-maker is the fully rational adult capable of understanding healthcare information as it is presented to her, weighing that information in the context of her personal life, and communicating her preferences and decisions to others. However, patients often find themselves in non-ideal healthcare circumstances due to illness or injury.

It is often necessary to turn to a *moral proxy*[16] for decision-making. A moral proxy is a person responsible for making healthcare decisions on behalf of another. Moral proxies are necessary when a patient is deemed incapable of making, or communicating, his own healthcare decisions. Incapacity of this sort can be age related—a young child cannot make her own decisions regarding care—or can result from a medical condition—a patient might be unconscious, or unresponsive, or generally unable to comprehend the nature of his condition. If a patient is unable either to comprehend the medical information being provided by healthcare professionals, or is unable to communicate his wishes, he cannot give informed consent to medical interventions—a proxy decision-maker is required in such circumstances.

Generally speaking, spouses, family members or next-of-kin are treated as the most appropriate moral proxies given their assumed intimate knowledge of the patient's personal medical preferences.[17] Historically, however, this was not always the case. The rise of modern medicine saw physicians and other health care professionals

---

[15] Wallach, W., Allen, C. (2009). *Moral Machines: Teaching Robots Right From Wrong.* (Oxford: Oxford University Press).

[16] Also referred to as a *proxy decision-maker*, or *substitute decision-maker*.

[17] Kluge, Eike-Henner. (2005). "Consent and the Incompetent Patient." In E-H Kluge (Ed.) *Readings in Biomedical Ethics, A Canadian Focus (3rd ed).* (Toronto: Pearson Prentics Hall).

assuming (indeed granted) decision-making functions in healthcare contexts involving incapacitated patients, but also in ordinary contexts involving fully capable patients.[18] Indeed, it was not uncommon practice for healthcare professionals to act in direct opposition to a patient's stated preferences. Through the latter half of the 20th century this trend was reversed. Critics argued that such a model of decision-making was unjustifiably "paternalistic", and that healthcare decisions were best left to the patient in large part because physicians and other healthcare professionals can not reliably account for the many intimate personal factors that contribute to an autonomous healthcare decision.[19] Because healthcare decisions are among the most intimate and private decisions one can make, few people would easily give up the power to make their own healthcare decisions today. In practice, turning decision-making authority over to a moral proxy tends to be a last resort only after every effort made to engage the patient in his or her own decision-making has failed. Despite physicians' considerable knowledge and expertise regarding medical options, families and next-of-kin are once again viewed as the most ethically appropriate proxy decision-makers.

Not surprisingly, proxy decision-making is not without controversy. In Canada and the U.S. the courts have found that moral proxies must strive to make proxy healthcare decisions that are ideally substitute judgments (exactly what the patient would have wished for) or, if the exact wishes are unknown, decisions that are in the best interests of the patient. In both cases the goal is to make decisions that "[the patient] would choose, if he were in a position to make a sound judgment" autonomously.[20] Thus, the ideal proxy decision-maker makes no decisions at all: the ideal proxy acts merely as a conduit for communicating the precise wishes of the patient given the current circumstances, and helping to ensure that those wishes are carried out. One could say that the ideal moral proxy acts mechanically, robotically even, when faced with decision-making on another's behalf, that if the patient has provided him with adequate information regarding her wishes—a script, perhaps—his role is to act out that script to the letter.

Typically, however, proxy decision-makers find themselves in a position of ignorance with respect to the patient's wishes, which complicates the decision-making process. Still, expectations are such that the proxy, most often a family member or next-of-kin, is required to disentangle his or her own preferences from those of the ailing loved one, a difficult requirement to be sure.[21] It is equally if not more difficult to determine, as an outside observer, whether or not the proxy has

---

[18] Ibid.

[19] The landmark and most widely cited example of this argument is contained in Beauchamp, T., Childress, J. (1979). *Principles of Biomedical Ethics.* (Oxford: Oxford University Press).

[20] *Re S.D.* (1983). 3 W.W.R. 618 (B.C.S.C.).

[21] Kluge, E.H. (2005). "After 'Eve': Whither Proxy Decision Making." In E-H Kluge (Ed.) *Readings in Biomedical Ethics, A Canadian Focus (3rd ed).* (Toronto: Pearson Prentics Hall):186-194.

met this requirement. In *Re S.D.*, a landmark case in Canadian law, parents acting as proxy to their severely disabled 7 year old boy refused consent for a routine surgery to unblock a shunt that drained spinal fluid from the brain. In their estimation their son was enduring a life of suffering that the surgery would only act to prolong. Family and Child Service petitioned the courts on behalf of the boy, resulting in the court case. Ultimately the dispute was over whether or not the parents' decision was in the best interests of the patient. The parents argued that it was, while the Supreme Court of British Columbia found that it was not. So go many of the difficult moral controversies surrounding proxy decision-making.

Pointing out that proxy decision-making is difficult and controversial is not meant to suggest that there is no good way of doing proxy decision-making or of resolving controversies over what is in a patient's best interests. In the context of this argument the controversial nature of moral proxies is meant to underscore the moral complexities associated with delegating proxy decision-making powers to someone other than the patient. Though not without its own controversies[22], having patients make their own healthcare decisions is broadly accepted by healthcare professionals, patients and the courts as morally preferable to having another decide on their behalf.

Technology as Moral Proxy

Consider the following scenario:[23] Jane is at high risk of life-threatening ventricular arrhythmias, a condition that causes her to unexpectedly experience life threatening abnormal cardiac rhythms. She found this out a decade ago after having being admitted to the emergency room with a heart attack. Jane's cardiologist told her that she was lucky to be alive, and that the symptoms could recur unexpectedly at any time. In order to increase her chances of surviving similar cardiac events in the future her cardiologist recommended that Jane be fitted with an Internal Cardiac Defibrillator (ICD). She was told that the ICD is a small implantable device consisting of a power source, electrical leads that are fixed to the heart, and a small processor that would monitor her heartbeat and deliver electrical stimuli (shocks) whenever a dangerously abnormal rhythm was detected, the goal being to return her heart to a normal rhythm. The ICD is a small implantable version of the larger defibrillator that the paramedics had used to save Jane's life. Being otherwise healthy at the time, she agreed to the surgery.

Three uneventful years after her ICD implantation, Jane recalls being in a meeting at work and suddenly feeling lightheaded. She recalls experiencing a painful "jolt" in her chest, and describes it as the rough equivalent of being kicked in the chest by a

---

[22] See, for example, Draper, H., Sorell, T.. (2007). "Patients' Responsibilities in Medical Ethics." In R. Chadwick, H. Kuhse, W. Landman, U. Schüklenk and P Singer (Eds.). *The Bioethics Reader, Editors' Choice*. (Malden, MA:Blackwell):73-90.

[23] This scenario is an adaptation based on first-person accounts of living with ICDs. See Pollock, A. (2008). "The Internal Cardiac Defibrillator." In Sherry Turkle (Ed.) *The Inner History of Devices*. (Cambridge, Mass: MIT Press):98-111.

horse. It was the ICD delivering a shock to her heart. The first shock was followed in quick succession by several others, though she cannot recall how many, though she says each one was as traumatic as the first. After several of these shocks her co-workers called an ambulance. Paramedics arrived to find Jane conscious but in shock. At the hospital she was told that the ICD had delivered a total of seven shocks to her heart. "The ICD performed perfectly," her doctors told her, "it saved your life."

It is worthwhile taking a moment to examine the tremendous work that is accomplished by such a small artefact as the ICD. Without the ICD the efforts of several people, and some luck, are required to prevent disaster. Jane could only hope to be in the company of others during a cardiac episode, for starters someone would need to call the paramedics to come help her. The paramedics, assuming Jane is near enough that they are able to arrive in time, must assess the situation, perhaps without the benefit of knowledge of her preexisting heart condition. (Do those who called in the emergency happen to be privy to Jane's heart condition? Is Jane conscious and alert enough to tell the paramedics of her condition when they arrive?) Assuming the paramedics are able to assess her condition accurately they must then prep both Jane and the defibrillator, and only then can they deliver shocks to Jane's heart. If all goes well, after considerable time, coordination and effort, the human actors surrounding Jane have just successfully performed the medical interventions the ICD is capable of performing on its own almost instantaneously. Thus, Jane's ICD all but eliminates the need for humans in the critical path to the medical intervention. It continuously monitors her heart, detects abnormal cardiac rhythms, and delivers potentially life saving shocks before a single human bystander has the chance to come to Jane's assistance.

Indeed, the ICD is a powerful little artefact! ICDs are capable of accomplishing the work of several humans, a concentration of power that Bruno Latour refers to as *delegation*.[24] Designers delegate the tasks of continuous monitoring, medical assessment and intervention to the ICD. Not only does the ICD perform those functions but also, unlike its human counterparts who are error prone, it performs them with the accuracy and consistency expected of a computer. ICDs also *mediate* our relationship with time and space.[25] With her ICD working diligently in the background, Jane is free to travel farther from medical centres with a certain confidence that in an emergency the time and space between her and medical experts will have less impact on her chances of survival; the ICD promotes Jane's freedom and independence by expanding her geographical safety zones.

---

[24] Latour, B. (1992). "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts." In W.E. Bijker and J. Law (eds.). *Shaping Technology/Building Society.* (Cambridge, Mass: MIT Press).

[25] Verbeek, P-P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things.* (Chicago: University of Chicago Press).; Verbeek, P.P. (2006). "Materializing Morality: Design Ethics and Technological Mediation." *Science, Technology & Human Values* 31(3):361-380.

In addition to the work that is delegated to the ICD, material answers to deeply moral questions can be delegated too.[26] An answer to that life and death question that was asked of Jane prior to her implantation—*Would you like to have potentially life-saving electrical shocks administered in the future event that your heart goes into an abnormal rhythm?*—is implicit in the presence of an activated ICD. *Yes! Shock the heart and sustain Jane's life!* The ICD continuously answers that deeply moral question by the mere fact that it is in Jane, actively monitoring her heart, ready to deliver potentially life-saving electrical shocks at the first sign of an abnormal rhythm. Her ICD works diligently in the background, instantaneously answering an important moral question on her behalf when called upon to act.

It is at this point that we notice something interesting about the relationship that Jane has with her ICD, we see that the ICD can be cast as moral proxy acting on Jane's behalf. Moreover, this proxy relationship approximates that of an ideal moral proxy. In the absence of an ICD, an ideal human moral proxy would be required to provide an answer to that same life and death question in an emergency: *Would Jane agree to have potentially life-saving electrical shocks administered given that her heart has gone into an abnormal rhythm?* An activated ICD provides a material answer to that moral question in addition to supplying the medical intervention. Indeed, we can say generally that in cases where semi-autonomous robots, such as ICDs and SDCs, provide material answers to moral questions in the use context, they function as moral proxies acting on someone's behalf.

To further illustrate how an artefact can function as moral proxy consider Jane's current situation. Just under a year ago she was diagnosed with inoperable cancer. Her initial prognosis suggested she had four to six months to live, and her health has deteriorated to the point that her physicians are beginning to discuss end-of-life palliative measures with her and her family members. As a part of those conversations Jane was asked whether she would want the medical team to attempt to resuscitate her in the event that her heart stopped. Recalling the intense pain she suffered when her ICD fired years ago, and recognizing the gravity of her current medical condition, she decided resuscitation attempts would be futile. Jane asked that no resuscitation efforts be made, only that she be kept comfortable.

Of course, her "do not resuscitate" (DNR) preference, a deeply moral end-of-life decision, would be ineffective if the medical team failed to alert all of the medical staff by way of adding the important DNR note to Jane's medical chart. In that case, healthcare staff that were not a part of the DNR conversation would have no way of knowing what Jane's preferences were, and would likely assume proxy decision-making powers and attempt CPR in an emergency. Her DNR preferences would be equally ineffective if no one alerted the ICD to her preferences by deactivating it. In that case, quite unaware of the information on her chart, in the event that Jane's heart went into an abnormal rhythm the ICD would assume proxy decision-making power and deliver its pre-programmed series of up to nine painful shocks. With the

[26] Ibid.

moral decision to (or not to) attempt cardiopulmonary resuscitation delegated to it, an ICD functions as an efficient moral proxy the moment abnormal heart rhythms are detected, that is, the moment the question whether or not to attempt resuscitation must be answered.

As it stands Jane is waiting for the medical staff to contact the ICD manufacturer to assist with the deactivation (hers is an older model that the hospital is unequipped to deactivate). She is told the deactivation could take several days to accomplish. Delegating to the ICD her new end-of-life preferences involves highly specialized equipment and some official paperwork. Jane hopes that her last moments in life will not involve several painful reminders of the powerful little device within her. As a moral proxy, her ICD is proving to be somewhat uncooperative.

Technological Moral Proxies as Evaluative Tools

We can now begin to understand the motivation behind asking the two questions posed in the introduction of this paper: 1) On whose behalf was the robot acting?; and 2) On whose behalf ought the robot to have been acting? As an evaluative tool the *moral proxy model* focuses our attention on the relationship that is instantiated between robots and users due to the semi-autonomous nature of certain robots. It also introduces a set of normative claims that one can apply to an analysis of semi-autonomous robots in their use context. As I have already pointed out, in the healthcare context there are accepted moral norms surrounding the instantiation of moral proxy relationships. First and foremost, moral proxies, technological or otherwise, should always act on behalf of the patient, in the patient's best interests, owing to the fact that moral proxies are ideally charged with voicing the autonomous healthcare preferences of the patient alone. As such, with respect to ICDs the recipient patient's (or the *user's*) *explicit* healthcare preferences ought to determine the state of whichever device settings provide material answers to moral questions, such as whether the device is active or inactive. The technology ought to be acting as proxy on behalf of the user. To grant that particular decision-making authority to anyone other than the patient (in cases where the patient is competent) would be to subject the ICD recipient to a problematic paternalistic relationship: the proxy would be acting on behalf of someone other than the patient.[27]

Second, even in cases where the decisions of another happen to correspond to those that the patient would have made if given the choice, decisions made by someone other than the patient when the patient is competent are still paternalistic in nature.[28] In philosophical ethics this is referred to as "moral luck", which Thomas Nagel and Bernard Williams both describe as cases in which the consequences of one's actions are inappropriately factored into judgments of one's actions.[29] For

[27] Millar, J. (*forthcoming*). "Technology as Moral Proxy: Autonomy and Paternalism by Design." *IEEE Proceedings: Ethics in Engineering, Science & Technology 2014.*
[28] Ibid.
[29] See Nagel, T. (1979). *Mortal Questions.* (New York: Cambridge University Press)., and Williams, B. (1981). *Moral Luck*. (Cambridge: Cambridge University Press). For

example, it would be inappropriate to praise the doctor who set an ICD to function in a particular mode, without explicit guidance from the recipient patient, even if the choice of settings happens to correspond to that which the patient would have made if asked. That the doctor was lucky in her choice of settings is not praiseworthy, and so does not justify the paternalism. For this reason cases in which designers preprogram "default" settings into devices will tend to subject users to paternalistic relationships, especially when patients are not fully informed of the nature of the default settings so as to make *explicit* autonomous decisions regarding their preferred state, and where those settings instantiate a proxy relationship.

Third, health technology designers should ideally make available options (settings) that allow users to make explicit choices about the proxy relationships the technology instantiates.[30] Here (as in the previous point) we welcome the designer, our third actor in the designer-technology-user relationship, explicitly back into the fold. Recognizing semi-autonomous robots as capable of instantiating proxy relationships highlights the unique role that designers play in characterizing the resulting proxy relationships. Designers are uniquely responsible for enabling delegation in the semi-autonomous robots they design.[31] They are also uniquely responsible for making options available (i.e. settings) that would allow different delegation to be realized in the use context; if options are not designed into a technology then it will tend to function in a single mode of operation which, as we have seen, can lead to paternalism by design. However, and this is the crucial point that we can draw from the healthcare analogy, just as the ideal proxy should not be responsible for deciding which option is in the best interests of the patient when input from the patient is a possibility, designers should not necessarily be made responsible for deciding *which* modes of operation to instantiate in the use context. That decision is most appropriately left to the user.

Fourth, a robust informed consent process should be adopted to help maximize patient autonomy with respect to the choices she explicitly delegates to a proxy. This is how paternalism is avoided in healthcare contexts, and it provides a good model for avoiding paternalism in the context of using healthcare technologies.

To sum up, the moral proxy model is a powerful evaluative tool that can be applied to an analysis of the design and use of semi-autonomous robots. The normative claims that shape appropriate moral proxy relationships in healthcare can be similarly applied to shape moral proxy relationships instantiated in the use context of semi-autonomous robots. Generally speaking, in order to avoid paternalism by design and to maximize user autonomy with respect to proxy relationships

---

another discussion of moral luck in the context of "expert robots" see Millar, J., Kerr, I. (2014). "Delegation, Relinquishment, Responsibility: The Prospect of Expert Robots." *Forthcoming* in *Robot Law* eds. Calo, Froomkin & Kerr. (Northhampton: Edward Elgar).

[30] Millar (n.26).

[31] Latour (n.24); Verbeek (n.25).

instantiated in the use context, design methodologies that aim to identify such proxy relationships should be adopted so that the appropriate design features and user choices can be incorporated up front.[32]

Proxies Beyond Healthcare

So far this argument has focused on examining proxy relationships instantiated via semi-autonomous healthcare technologies, primarily because the moral proxy model finds its normative roots in healthcare contexts. As such, healthcare technologies provide solid anchors for securing the analogy. However, they are be no means the limit to the analogy. By definition proxy relationships instantiated between users and semi-autonomous robots delegate answers to moral questions arising in the use context. This means that each proxy relationship carries with it moral implications for the user. Thus, we can evaluate each instantiated proxy relationship on its own merits, whether or not it arises in a healthcare context, to determine the kinds of decisions it delegates to the robot. Only then can we make decisions regarding the extent to which we must seek informed consent regarding the instantiation of a particular proxy relationship.

Self-Driving Cars

One can apply the moral proxy model to evaluate the hypothetical moral dilemma involving the SDC raised earlier in this paper. Similar to the situation involving Jane's end-of-life decisions and the state of Jane's ICD (should it remain active or be deactivated?), the SDC dilemma has no clear objective answer—what should the car do? Regardless of which path the SDC takes, we can consider the SDC a moral proxy acting on behalf of whoever set it to take that path. If, for example, the engineers at Google programmed the car to keep going straight, in other words to always protect the user's life (and perhaps to minimize other casualties) in situations where the user's vehicle is not at fault for the situation, then we can say that the SDC is acting as moral proxy on behalf of Google. But according to the ethical standards surrounding proxy relationships this situation might turn out to be morally problematic. It might be the case that the user of the car would feel morally obliged to risk her own life in such situations, say by choosing to swerve into the truck, especially where innocent children's lives are concerned. Such a user, if her SDC kept going straight according to Google's decision, would find herself subject to a paternalistic relationship: a moral proxy acting on behalf of Google would thwart her moral preferences. If we consider the owner the morally appropriate decision-maker in this driving context, and it seems we can since her life is directly at risk and she is not at fault (has not broken laws leading to the situation, has not erred, etc.), then hers is the autonomous decision that ought to be represented by the proxy, not Google's.

From a design and use perspective this analysis suggests that SDC owners ought to be offered a choice of settings that would delegate the appropriate decisions to the

[32] Millar (n.26).

SDC how to respond in situations like the one described. As is the case in healthcare contexts, SDC users ought to be informed of the potential consequences of their technology choices so they can appropriately exercise their autonomy. Users should be asked to make autonomous decisions regarding certain cases where it is reasonably foreseeable that the robot will provide material answers to moral questions, in this case risking either the owner's life or the lives of others in particular use contexts. Making that decision on behalf of the user risks subjecting her to a paternalistic relationship in which her autonomy is unjustifiably fettered.

At times ethical arguments prescribing maximizing strategies have a tendency to exasperate those tasked with doing the work of maximizing.[33] Utilitarian arguments for example, which ask people to act in a way that would maximize happiness in the world[34], tend to elicit this response among those who realize suddenly the daunting calculus that would be involved in determining exactly what the relationship is between any act and the net amount of happiness contributed. My argument suggests that designers ought to maximize user autonomy by changing the way they design semi-autonomous robots, and I have piled on top of that a requirement that users exercise their autonomy by making explicit choices regarding certain settings in those technologies. This could seem overly burdensome to designers, who could interpret any number of design features as instantiating a proxy relationship in the use context, many of which would seem relatively unproblematic if left to them. For example, one could argue that a shopping cart's wheels should not be designed to lock up beyond a certain physical boundary by default, owing to the paternalistic relationship that feature subjects users to.[35] Users might also feel unduly burdened by the prospect of having to make complicated (or bothersome) decisions about the modes of operation in which their devices function. They might rightly wonder, *is*

---

[33] I am speaking from experience here, both as a professor who has presented maximizing theories to undergraduate and graduate students, and as someone who has presented similar ideas to diverse audiences at conferences. In both contexts it is common to field questions from individuals convinced that the particular maximizing theory is unduly burdensome in its demands on individuals, and so should be jettisoned. To be sure, if taken to the extremes, any maximizing theory can demand too much. Practically speaking, however, we apply maximizing theories as ideal guides, rounding off and estimating where we can justify doing so, and retaining the decimals where we cannot.

[34] See, for example, Mill, J.S. ([1863]2002). *Utilitarianism*. George Sher (Ed.). (Indianapolis: Hackett), for the classic Utilitarianism argument.

[35] Ian Kerr makes an argument of this sort, though his focus is the effect that digital locks have on an agent's ability to develop a moral character generally. To be sure, any harm resulting from the use of technology warrants consideration by designers. I use this particular example as one the sits somewhere near (perhaps on) the border of those design features that would require users to make explicit choices about their technology use. See Kerr, I. (2010). "Digital Locks and the Automation of Virtue" in Michael Geist (Ed.) From "Radical Extremism" to "Balanced Copyright": Canadian Copyright and the Digital Agenda (Toronto: Irwin Law).

my autonomy really compromised if I don't make an explicit choice about the wheels on every shopping cart I push around a grocery?

My argument is not intended to unduly burden designers or users any more than healthcare professionals and patients are burdened by informed consent requirements and the ethical and legal rules surrounding proxy decision-making in healthcare. Healthcare systems are now designed so that answers to many moral questions are left to the patients, though not all moral questions implicating patients require informed consent. Some questions that arise in the healthcare context are left to the physician. Just as we have worked out frameworks for determining which decisions require informed consent, I suggest we can identify frameworks for determining which proxy relationships instantiated in the use of technology require explicit decision-making on the user's part. Though I will leave much of the work of setting out the specifics of such a framework for further research, I will suggest the following as a starting point: a proportional approach will be helpful in setting thresholds for requiring explicit input from the user. Following this line of thought, each identifiable proxy relationship will need to be evaluated for its mediating effects to determine the particular nature of the answers delegated to the proxy. Though paternalistic, cases involving locking shopping cart wheels might not be deemed sufficiently problematic to warrant changes to the design process that allow for, or require, explicit user input. But it seems cases involving ICD activation settings, SDC settings determining the outcome of "trolley problems", and others will. Indeed, we can anticipate that as more sophisticated decision-making algorithms are incorporated into future technology, the number and complexity of proxy relationships instantiated in the use context will increase, and so too will our requirement to identify, evaluate, and instantiate those relationships appropriately. To sum up the key points of my argument so far, they are that proxy relationships are instantiated in the use of technology, that the normative features of those proxy relationships are shared across healthcare and other technology use contexts, and that we ought to respond appropriately by maximizing user autonomy in a proportional manner both when designing and using semi-autonomous robots.

Proxies and Legal Models of Responsibility

I turn now to examining legal models of responsibility as a first step towards suggesting how the moral proxy model might impact some of the legal reasoning around questions of responsibility in the context of designing and using semi-autonomous robots.

As I sit here wondering how best to begin a section on legal liability, I reach for the latté I purchased at the nearby coffee shop and notice the warnings printed on its plastic travel lid:

> "ATTENTION CHAUD !"
>
> "CAUTION HOT !"

"¡ PRECAUCION CALIENTE !"[36]

In addition to these bold warnings, on the cup there is an inscription that reads, "Careful, the beverage you're about to enjoy is extremely hot."[37]

These warnings (in addition to being informative if somewhat presumptuous) are legal tools intended to satisfy a company's duty to warn the consumer of particular risks associated with using their product. In this case Starbucks® is warning me that the coffee might be extremely hot so that I can take appropriate cautions when drinking it, for example by not immediately popping the lid off and taking large gulps of the contents. If I did engage in such behavior, and if Starbucks® hadn't put those warnings on the lid and cup, and if I had been literally burned by the coffee, Starbucks® could get figuratively burned in court by being found negligent for *failure to warn*. Failure to warn occurs when manufacturers fail to notify consumers of foreseeable risks associated with using a product.[38]

In terms of autonomy a coffee cup is much more like a hammer than a person. Although it is possible to identify ways in which the cup mediates our moral landscape[39], its relative lack of autonomy prevents it from acting as moral proxy in any meaningful way.

Such is not the case with semi-autonomous robots. When we scrutinize the proxy relationships they instantiate we see that questions are raised immediately where a failure to warn is concerned. In theory, each instance of a proxy relationship between the user and the technology carries with it actual and potential harms. First, a user is harmed by virtue of being subjected to a paternalistic proxy relationship. Paternalistic proxy relationships, both inside and outside of healthcare contexts, are generally problematic because of their potential to impose on the user material answers to moral questions other than those autonomously and explicitly expressed by the user. As I have suggested, a proportional approach to evaluating proxy relationships can help determine which among them would be more or less problematic if instantiated paternalistically, but paternalism of this kind requires a justification as a general rule. Second, the outcomes, or effects of, a paternalistic relationship could potentially harm the user. As demonstrated by the ICD and SDC cases, when a technological moral proxy acts paternalistically it does so without express consent of the user. That an SDC user suffers as a result of the proxy is not

---

[36] Starbucks®. (2014). *Grande paper cup with SOLO traveler lid. Canadian Ed..*

[37] Ibid.

[38] Asaro, P. (2012). "A Body to Kick, but Still no Soul to Damn: Legal Perspectives on Robotics." In (P. Lin, K. Abney, and G. Bekey) *Robot Ethics: The Legal and Social Implications of Robotics.* (Cambridge: MIT Press):169-186.

[39] For an interesting discussion on disposable coffee cups and mediation see Verbeek (n.25). Verbeek, building on Latour's notions of *delegation* and *scripting*, describes how a plastic coffee cup contains the script, "throw me away after use, whereas a porcelain cup 'asks' to be cleaned again and again" (p.362). Thus different coffee cups influence human behaviour and even provide scripts for them to follow.

the issue here. Rather, it is the particular way in which the user suffers that is at issue in proxy relationships. I take it to be intuitive that suffering at the hands of others is fundamentally different than suffering of one's own accord.

How should we satisfy a duty to warn given these considerations? Should we devise a litany of warnings to imprint on the pages of SDC user manuals each warning of a potential proxy relationship and its associated potential effects? In principle, each proxy relationship has the potential to harm. But a more reasonable approach than a barrage of warnings is to apply a proportional analysis to each identifiable potential proxy relationship as a first step. Some proxy relationships will not warrant action. An appropriate mechanism for dealing with those proxy relationships that remain serious enough to warrant action is not merely to provide warnings. Warnings leave open the possibility of paternalistic proxy relationships since they fail to address the underlying design issues. As is the case in healthcare contexts, the appropriate response is to adopt a design process that enables a robust informed consent process in the use context, one that provides warnings and information in combination with a requirement of the user to provide explicit consent for certain delegated decisions. It is worth noting that this requirement is likely not adequately being addressed by current design methodologies. Indeed, designers adopting current established design methodologies might intentionally be thwarting meaningful informed consent.[40] A design approach based on the moral proxy model could help close this gap.[41]

Another sort of liability stems from *a failure to take proper care*, which occurs when a manufacturer fails to foresee a risk that they can reasonably be expected to foresee based on a "community standard of reason", or "industry standard of practice" among manufacturers of similar products.[42] Asaro refers to proper care as "perhaps the central issue in practical robot ethics from a design perspective" owing to the complexity associated with anticipating outcomes in complex use contexts.[43]

Again, the moral proxy model has implications for proper care. Though a proxy analysis does not promise to help anticipate all potential harms associated with use (what design methodology could?), it does help to identify and characterize a class of harms that could otherwise go unnoticed by designers: the actual and potential harms of paternalistic proxy relationships. Thus, adopting a proportional approach to proxy analysis in the design phase, and designing for user autonomy through robust informed consent features, helps to legitimately mitigate risk. Just as healthcare professionals have seen their proxy decision-making authority limited by legitimate transfers of responsibility to patients, in large part to mitigate harms associated with paternalism, so too should designers recognize the reasonable

---

[40] Kerr, I., Barrigar, J., Burkell, J., Black, K. (2006). "Soft Surveillance, Hard Consent." *Personally Yours* 6:1-14.
[41] Millar (n.27).
[42] Asaro (n.38):171.
[43] Ibid.

grounds for transferring responsibility for decision-making to users. Borrowing from the healthcare context, proper care in design should include the responsibility to identify problematic proxy relationships that are instantiated in the use context and design choices into semi-autonomous robots that support robust informed consent processes.

This suggestion addresses a point Asaro raises in relation to the "industry standard defense", which allows manufacturers to defend against charges of failure to warn and failure to take proper care so long as they are following "industry standards" as accepted by their peers.[44] The problem with the concept of an industry standard is that "it fails to tell us what sorts of practices *should* be followed in the design of robots".[45] The moral proxy model partially answers that question by providing a normative design framework aimed at managing a specific class of harms. Recognizing the ethical dimensions of moral proxy relationships in the use context effectively raises the bar for industry standards—designers need to respond to the fact of proxy relationships appropriately as part of their standard design process—while providing guidance on how to go about doing the work of design so that the right people end up responsible for making decisions delegated to semi-autonomous robots.

Strict liability in the U.S. covers cases where there is no negligence, yet where there exists a product defect resulting in harm.[46] Of the three categories of defect—manufacturing, design, and warning—design defects are of particular importance in the context of this argument. A product is defective in design when "the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design…and the omission of the alternative design renders the product not reasonably safe."[47] Consider a paternalistic proxy relationship such as that instantiated by an SDC preprogrammed to sacrifice an errant child pedestrian to save the user's life, or that instantiated by Jane's stubborn ICD. The mere fact of the relationship constitutes harm from an ethical perspective, as I have discussed, and in cases where the device settings confound a user's moral preferences the harm is amplified. Of course, harms resulting from the material outcomes of paternalistic proxy relationships—e.g. an ICD that fires without a user's explicit consent but after she has expressed do not resuscitate preferences—are also problematic. In these cases the instantiation of paternalistic proxy relationships in the use context can reasonably be interpreted as a design defect.

A proxy analysis approach to design points to reasonable alternatives for some design choices. Designers can focus on providing users with mechanisms to achieve meaningful informed consent as a reasonable alternative to deciding on modes of operation themselves. One way this could be achieved is by offering users more than

---

[44] Ibid:172.
[45] Ibid.
[46] 59 Cal. 2d 57 (1963).
[47] Restatement (Third) of Torts § 2[c].

one mode of operation to select from (e.g. on/off, MODE_A/MODE_B) with respect to whatever functionality instantiates a proxy relationship. In the case of the ICD this is accomplished by including a switch that allows the active/inactive state to be explicitly selected, while in the case of the SDC a user could be given a choice between tendencies to swerve this way or that in reasonably foreseeable situations. Sometimes, however, it will not be practical to design alternative modes of operation into a device, leaving designers the task of providing a single mode of operation yet requiring informed consent prior to use. This could be the case when technical obstacles stand in the way of providing an alternative solution, or in cases where the cost of providing an alternative is too great. In these cases we can still achieve meaningful explicit informed consent without resorting to default settings (defaults tending to blur the line between paternalism and autonomy). When faced with having to provide only one mode of operation to users, meaningful informed consent can be achieved as it is in healthcare contexts, by providing the user a reasonable amount of information upon which he can base his decision whether or not to use the device given the knowledge of that proxy relationship he will be subjected to, *and* by asking him to knowingly and explicitly activate that mode of operation. This could require designers to put the device in a mode of operation that prevents it from functioning at all until the user makes the appropriate decisions explicitly. Though this suggestion might seem contrary to established design principles that seek to make technologies integrate "out-of-the-box" seamlessly and invisibly in the use context[48], it recognizes moral dimensions of design that become problematic precisely because of that design attitude: default modes of operation can, at times, subject users to problematic paternalistic proxy relationships. Requiring some assembly is one way of helping to ensure users are more informed about how the critical parts of a device actually function in the world.

Treating semi-autonomous robots as proxies suggests that they ought to be considered agents acting on behalf of another person. As such, "*vicarious liability*—when one person or legal entity is liable for the actions of another"[49], for example in cases where a robot acts on behalf of its owner, seems to apply in cases where proxy relationships are established. However, Asaro suggests that vicarious liability would tend to apply by virtue of the property relationship established between a semi-autonomous technology (robot) and its owner, and that the owner would likely be liable for harm caused by the robot.[50] A proxy model complicates this picture. If the owner is the subject of a paternalistic proxy relationship (i.e. one not established via a robust informed consent mechanism) it becomes difficult to assign responsibility based solely on issues of ownership. If an owner/user is subjected to a paternalistic relationship, and if the robot does harm as a result of paternalistic functionality, there is good reason to look elsewhere when assigning responsibility for the proxy's actions. The fact of a paternalistic proxy relationship leading to harm would seem to diminish an owners responsibility by shifting focus to whomever set the robot to act

[48] Norman, D. (2011). Living With Complexity. (Cambridge, Mass.: MIT Press).
[49] Asaro (n.38):178.
[50] Ibid:179.

in that way. If, however, the proxy relationship were adequately established, then standard vicarious liability would seem to apply.

Finally, in cases of differential apportionment of liability, in which some parties are held more or less responsible than others, the moral proxy model has a modifying effect. In all cases of proxy relationships the moral proxy model imposes two important questions:

1) Who was the robot acting on behalf of?; and

2) Who ought the robot to have been acting on behalf of?

Within a causal analysis the moral proxy model has the effect of shifting the focus of the analysis, by suggesting which causal chains ought to have been in effect, rather than focusing simply on which causal chains were actually in effect. Question 2 will modify the analysis of responsibility prior to the final causal chain analysis. If it is the case that the user ought to have been setting the device in such and such a mode of operation, yet the designer did so, then we can say that the designer is responsible for a larger portion of the outcome by virtue of having usurped the user's autonomy. If the user set the device appropriately, then we can say that he did so autonomously and that the proxy was acting on his behalf, resulting in a greater proportion of responsibility landing on his shoulders.

Conclusions: Risks and Opportunities

The moral proxy model, applied using a proportional proxy analysis approach to design, allows both designers and users to better manage some of the risks and leverage the opportunities associated with semi-autonomous robots. From a design perspective, when instantiating proxy relationships it is important to manage informed consent appropriately. Robust informed consent finds its normative roots in healthcare contexts, but extends beyond healthcare in cases where robots stand to answer moral questions as proxies—the user will often be the most appropriate decision-maker in any context where those answers stand to affect him most directly. Where robust informed consent practices are the requirement, as they should be in many cases of moral proxy relationships, it is ethically (and perhaps legally) problematic to claim that consent can be established through the use of coercive methods—consent must be freely obtained. Autonomy by design is one way of managing the risks associated with design. Designers should seek to provide alternative modes of operation where those modes instantiate reasonably foreseeable proxy relationships, and should build robust informed consent mechanisms into their technologies that force the user to make explicit choices regarding those modes of operation. The proxy model allows designers and users to assign responsibility for decisions according to established norms.

As technologies become increasingly autonomous they will also become more capable of instantiating proxy relationships. Thus, the need for a model of responsibility that accounts for the unique relationship instantiated by semi-autonomous robots is pressing. We have jettisoned paternalism in healthcare

because of the risks it poses to patients; we have good reasons to do so in design because of the risks it poses to designers and users alike.