Sketching an Ethics Evaluation Tool for Robot Design and Governance

Jason Millar[1]

Note: This is a draft version of a work in progress intended for discussion at We Robot 2015. Please do not distribute this work in its current form. If it piques your interest and you would like to keep abreast of a future (published) version, please email me (jason.millar@queensu.ca) and I will when one becomes available.

Introduction

Consider these four user experiences:

Case 1: The Tunnel Problem

Sarah is travelling along a single-lane mountain road in an autonomous car that is fast approaching a narrow tunnel. Just before entering the tunnel a child errantly runs into the road and trips in the centre of the lane, effectively blocking the entrance to the tunnel. The car is unable to brake in time to avoid a crash. It has but two options: hit and kill the child, or swerve into the wall on either side of the tunnel, thus killing Sarah. It continues straight and sacrifices the child.

Case 2: Jibo the Wingman

Steve has just purchased Jibo, a small, *social robot* designed for use in and around the home. Jibo is marketed as the first robot "family member". It sits on a desktop, and is equipped with cameras and a microphone so that it can sense its environment and collect data. It is designed to interact on a "human" level by conversing in natural language with its users, laughing at jokes, helping with tasks (e.g. scheduling, making lists, reminders, taking pictures), and most importantly responding to humans in emotionally appropriate ways, all of which is meant to engage users in a human-like relationship with Jibo. Jibo can also function as a "wingman"; the primary reason Steve bought it. Steve is able to identify a love interest to Jibo, say a date he brings home one evening, and Jibo then analyzes and characterizes the date based on proprietary learning algorithms (automatically updated based on the successes/failures of all Jibos), and access to social networks and other "big" datasets. As part of its data gathering technique Jibo spontaneously strikes up conversations with the love interest, often when Steve is in another room, an activity the designers term "isolation mining". Successful isolation mining requires divulging some of Steve's personal history and

[1] Jason Millar teaches in the philosophy department at Carleton University (Ottawa), and is a PhD candidate in the philosophy department at Queen's University (Kingston). This research was funded in part by Canada's Social Sciences and Humanities Research Council (SSHRC) through a Joseph Armand Bombardier Canada Graduate Scholarship, and in part by the Canadian Institutes for Health Research (CIHR) through a Science Policy Fellowship. The author can be reached at jason.millar@queensu.ca.

other information (e.g. childhood experiences, common interests, and so on) to the love interest. Studies show this helps to gain the love interest's trust, enhance the data gathering, increase attraction levels, and ultimately improve Steve's romantic odds. One evening, Steve brings a woman he's been dating home and introduces her to Jibo, then goes into the kitchen to get dinner started. In conversation with the love interest, Jibo divulges several of Steve's very sensitive personal anecdotes.

Case 3: C-bot the Unwelcome Bartender

Mia is a 43-year-old alcoholic who lives alone and recently broke her pelvis and arm in a bad fall down the stairs. As a result she is currently suffering extremely limited mobility. Her healthcare team suggests that Mia rent a C-bot caregiver robot to aid in her recovery. Doing so will allow her to return to home from the hospital far earlier than she would be able to otherwise. C-bot is a social robot designed to move around one's home, perform rudimentary cleaning tasks, assist in clothing and bathing, fetch meals, help administer some medications, and engage in basic conversation to collect health data and perform basic head-to-toe and psychological assessments. Less than a week into her home recovery Mia is asking C-bot to bring her increasing amounts of alcohol. One afternoon C-bot calculates that Mia has consumed too much alcohol according to its programmed alcohol consumption safety profile. Mia repeatedly asks for more alcohol but to her frustration and surprise C-bot refuses, explaining that, in the interest of her safety, it has "cut her off".

Case 4: The Stubborn ICD

Jane has an Internal Cardiac Defibrillator (ICD), a small potentially life-saving implantable robot that "shocks" her heart whenever it detects an abnormal, life-threatening, cardiac rhythm. She received her ICD after a near death experience almost 10 years ago, and the ICD has since saved her on two separate occasions. Jane was recently diagnosed with terminal pancreatic cancer, and after several months of unsuccessful treatments, is nearing death. As part of her end-of-life decision-making she has asked that her ICD be deactivated, and that no measures be taken by medical staff to restart her heart if it should stop. She has made these requests to have the peace of mind that she will not suffer the painful experience of being "shocked" (it is often described as being kicked in the chest by a horse[2]) at her moment of death. Her healthcare team has agreed not to perform CPR, but the physician who oversees her ICD is refusing to deactivate it on grounds that it would constitute an active removal of care, in other words, that deactivating the device would count as a kind of physician-assisted suicide.[3]

---

[2] Pollock, A. (2008). "The Internal Cardiac Defibrillator," in *The Inner History of Devices*, S. Turkle, Ed. (Cambridge, Mass: MIT Press): 98-111.

[3] Ngai, D. (2010). "Turning Off the Implantable Cardiac Defibrillator to Prevent Pre-Death Electrical Shocks: An Exercise and Right in the Refusal of Medical Treatment." *The Internet Journal of Law, Healthcare and Ethics (7)*1.

Each of these four user experiences describes a different embodied automation technology—a robot—that in order to function well must be capable of making morally loaded decisions in a use context. In the tunnel problem we have an autonomous car that, like any driver, must choose to sacrifice Sarah or the child. Jibo, like any wingman, must select which personal information to divulge to Steve's love interests, and must also decide to what extent Steve can know about those divulgences. C-bot must decide under what conditions to give Mia alcohol. Jane's ICD must decide whether or not to shock her heart in the final moments of her life. These are hard ethical cases, ones that don't have objective answers[4], and each of which carries significant ethical implications for the user.

Autonomous cars, social robots, caregiver robots, and ICDs promise to deliver any number of benefits to society. They are worth pursuing. But the nature of these automation technologies also introduces a novel kind of engineering design problem, and related governance issues, illustrated by these four cases. On the one hand, automation technologies are intended to replace human *action*—the term "automation" has its origins in the Greek *automatos*: acting of itself. Autonomous cars drive themselves. Jibo acts as a wingman. C-bot accomplishes caregiving tasks. The ICD performs CPR. But for the purposes of this paper, these cases are meant to illustrate that automation technologies can also replace human *decision-making*, and more to the point *ethical* decision-making.

How should we automate ethical decision-making? That is the novel engineering design problem, and it suggests related governance challenges. This paper starts with an ethical evaluation of robots that automate ethical decision-making, using the four user experiences as case studies. I then argue that the automation of ethical decision-making would benefit from the introduction of design tools intended to help ethicists, engineers, designers, and policymakers anticipate ethical issues that might arise in the use context, and distinguish between acceptable and unacceptable design features. To aid in the design of ethics evaluation tools, I propose five specifications for any such tool. Finally, I sketch an example of an ethics evaluation tool that meets the five

[4] In hard cases, according to Ruth Chang, there is no obvious way in which one choice is better than another *tout court*. The reason is that each of the choices involves particular considerations that are useful in evaluating that choice, but the considerations are not shared by all of the other choices. In the Tunnel Problem, for example, you might consider the horror of sacrificing a child and watching it die in evaluating the one choice, while the alternative choice has you considering the weight of your own death and the impact it might have on your family. Though we might agree on the relevant considerations for each choice, in hard cases like the Tunnel Problem "it seems all we can say is that the one alternative is better with respect to some of those considerations while the other is better in others, but it seems there is no truth about how they compare all things considered". See: Chang, R. (2012). "Are Hard Choices Cases of Incomparability?" *Philosophical Issues 22*:106-126; and Chang, R. (2002). "The possibility of Parity." *Ethics 112*: 659-688.

specifications, and that could help avoid the issues identified in the four user experiences.

Automating Ethical Decision-Making: An Ethics Evaluation and the Need for an Ethics Evaluation Tool

Automating ethical decision-making runs the risk of taking users entirely out of the ethical decision-making loop in cases that: (a) have direct, and significant, moral implications for them, and (b) according to established ethical norms particular to each use context, ought to include their input.[5] Thus, automating ethical decision-making can pose a direct threat to a user's moral autonomy.[6] In the tunnel problem Sarah might choose to sacrifice herself to save the child, and so the context involves a sort of end-of-life decision-making.[7] Steve might have good reasons to prefer Jibo not to divulge certain facts about his life, even if they would improve his relationship prospects, demonstrating that this context involves privacy issues[8], and issues related to Steve's ability to understand how Jibo, his artificial "friend", represents him to his friends and loved ones[9]. Mia wants another drink and would get one herself if not for her injury, which in the context of her own home seems a matter of personal choice. Finally, Jane wants to have a dignified death requiring that her ICD be deactivated, so her ICD use context involves complex end-of-life decision-making in a medical setting.[10]

In each of these four cases engineers and designers (designers from here on in[11]) could choose to take the user entirely out of the decision-making loop. For the purposes of this argument let's assume it's perfectly legal to do so. But that design decision would

[5] Millar, J. (forthcoming). "Technology as Moral Proxy: Autonomy and Paternalism By Design." *IEEE Technology & Society.*

[6] Millar, n.5.

[7] In a poll conducted by the Open Roboethics Initiative nearly 40% of participants indicated that they would prefer to sacrifice themselves to save the child if they found themselves in a similar driving situation. ORi. (2014). "Results: My (Autonomous) Car, My Safety." *Openroboethics.org.* Online: http://www.openroboethics.org/results-my-autonomous-car-my-safety/.

[8] Calo, R. (2015). "Robotics and the New Cyberlaw." *California Law Review* 103. Online: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2402972.

[9] Millar, J. (2014). "Facebook – Our fFriendly Automated Identity Bender." *Robohub.org.* July 11. Online: http://robohub.org/facebook-our-ffriendly-automated-identity-bender/.

[10] Ngai,n.3.

[11] Recognizing that engineers and designers occupy overlapping but separate roles in the development of technology I opted to describe the activity as *design* in the hopes that engineers unfamiliar with HCI or HRI will be comfortable referring to themselves as *designers*. I'm quite certain many designers would (rightly) take issue with my choice had I gone down the other path.

undermine the user's moral autonomy in a way that is ethically problematic.[12] Building on the work of Latour and Verbeek, I have argued elsewhere that when a robot imposes on the user material answers to hard moral cases, without the user's input, it subjects the user to a form of *paternalism* that is ethically undesirable.[13]

Paternalism arises in the context of relationships formed between users and robots. Indeed, the user-robot relationship is a useful framework of analysis on my account. By modeling the robot as an active participant in a user-robot relationship we gain traction on its ethical character *as an ethical decision-maker* in that relationship. Ethics, after all, is a social enterprise; it has meaning only in the context of human relationships. By treating the robot as an active participant in a human-robot relationship, therefore, we can characterize robots that make decisions for users in hard ethical cases as *moral proxies*.[14] This is a useful analytical move since the ethical debate surrounding moral proxies is well established, especially in the healthcare context where hard cases abound. Bioethics debates surrounding the moral proxy relationship can help us understand how best to design those relationships to deal with hard cases, and also alert us to the troubling possibility that a moral proxy can confound an individual's autonomous moral preferences. In fact, well into the mid twentieth-century it was commonplace for physicians to assume proxy decision-making authority, including end-of-life decision-making, on behalf of patients.[15] That practice, arising in the relationships between healthcare professionals and patients, is the prototypical version of paternalism. In healthcare, paternalism has been discarded on ethical grounds. If it is unethical for healthcare professionals to act paternalistically with respect patients' hard ethical choices, there seems little reason to consider it ethically sound for designers to subject users to similar relationships via robotics.[16]

Where is the paternalism in our four cases? It is embedded in the code, the settings, and the resulting behaviour of each of the robots in its relationship with each user. Insofar as the robots are not designed to take Sarah's, Steve's, Mia's, or Jane's moral preferences into account in each of their hard cases, the robots subject each to a paternalistic relationship. Sarah's autonomous car acts paternalistically when it decides

---

[12] In relations to users, "autonomy" here is always meant in the strong moral sense. This kind of moral autonomy is most commonly deployed in healthcare contexts, and is supported therein by robust informed consent practices.

[13] Millar, n.5. Also, Millar, J. (2014). "You Should Have a Say In Your Robot Car's Code of Ethics." *WIRED.com*. Sept. 2. Online: http://www.wired.com/2014/09/set-the-ethics-robot-car/. See also: Latour, B. (1992). "Where Are the Missing Masses: The Sociology of A Few Mundane Artefacts." in *Shaping Technology/Building Society: Studies in Sociotechnical Change*, W.E. Bijker and John Law, Eds. (Cambridge, Mass.: MIT Press): 225–258; and Verbeek, P-P. (2006). "Materializing Morality: Design Ethics and Technological Mediation," *Science, Technology and Human Values 31*: 361-380.

[14] Millar, n.5.

[15] Jonsen, A.R. (2008). *A Short History of Medical Ethics*. (Oxford: Oxford University Press).

[16] Millar, n.5.

not to swerve, insofar as it excludes her from the decision-making loop. Jibo acts paternalistically when it decides which sensitive anecdotes to share with the love interest, insofar as Steve is not a party to that decision. Mia is subjected by C-bot to a paternalistic relationship insofar as she has lost control of her ability to choose what to put in her body. Finally, her ICD subjects Jane to a paternalistic relationship insofar as it stubbornly insists on performing CPR against her autonomous desires.

These considerations suggest the following ethical design principle:

> *(A) When automating ethical decision-making, designers must not create technology that subjects users to morally problematic paternalism.*

Though a useful signpost, this principle is of somewhat limited help to designers. Philosophers might be happy to rest having identified such a principle, but it lacks a certain practical appeal in the design room. It does, however, suggest a roadmap for what would be required of a good ethics evaluation tool that designers could apply to their work. The principle highlights the distinction between automating actions, and automating decision-making, while pointing to a particular requirement applicable to automating ethical decision-making. It also identifies a particular kind of effect that robots can have, namely the instantiation of paternalistic relationships, and suggests a further distinction between morally acceptable and unacceptable paternalism.

Not all paternalism is unacceptable. For example, so long as a nurse gets permission from a patient *to* perform a particular intervention, the nurse is free to make a number of decisions about *how* to intervene without further consent, so long as those decisions fall under the standard of care for that intervention. To illustrate, once a patient is admitted to an intensive care unit for a particular course of treatment, and informed consent has been satisfied for that general treatment, the nurse can administer an array of pharmaceuticals, in other words decide freely *how* to administer that treatment without asking further permission of the patient. If the course of treatment changes, the patient (or her proxy decision-maker) must once again be consulted. Thus, in healthcare we have relatively clear examples of morally unacceptable and acceptable paternalism.

In robotics, this distinction will be harder to make because our social practices surrounding technology grant engineers broad decision-making authority with respect to their designs. The four cases I have outlined might strike most engineers as examples of perfectly acceptable paternalism. Users would likely disagree. In a poll conducted by the Open Roboethics Initiative, participants were presented with the Tunnel Problem and asked whether or not they thought engineers had the moral authority to decide how the autonomous car should respond.[17] A full 82% of participants said "No". Instead, they indicated (in a near split decision) that either the user or policymakers should decide the outcome. So, as in healthcare, we have good reason to think that users will

[17] ORi, n.7.

have very specific expectations with respect to their autonomy. They will likely reject certain forms of paternalism in design, once they realize they are being subjected to it.

We can anticipate an increase in the number of instances of paternalism in design, owing to the current growth trend in automation technologies. As automation technologies advance in sophistication, we will undoubtedly automate more ethical decision-making in the process.

It would be good to get in front of this issue and work towards a design methodology that explicitly acknowledges the challenges unique to automating ethical decision-making, and helps us distinguish between acceptable and unacceptable paternalism. In cases where a design feature results in unacceptable paternalism, as would be the case if Jibo divulged Steve's sensitive personal anecdotes to a love interest without involving Steve in the decision-making, it would be useful to have a range of options for managing that paternalism made available. One way forward is to develop a practical tool to help engineers in these tasks. For the remainder of this paper I provide a sketch of such a tool.

Five Specifications for Ethics Evaluation Tools in Robotics

The primary focus of this paper is sketching a design tool that designers can use to avoid making robots that subject users to morally questionable paternalism. However, one can imagine any number of ethical issues that robots will raise in use contexts. In order to design specific ethics evaluation tools for robots, I start with a general question: What features are required of *any* ethics evaluation tool for it to be of practical use in the design of robotics, and for it to further advance the practical application of robot ethics? I propose the following five requirements as a first response to that question.

1. *An ethics evaluation tool for robotics should be proportional in its approach.*

For an ethics evaluation tool to be of any use to designers, it must strike an appropriate balance between the needs of designers and of users. A proportional approach is meant to ensure that an adequate protection of users is maintained, while reducing unnecessary impediments to the design and manufacture of robotics. Given that robots occupy roles in a multitude of use contexts, and that the potential harms will be unique to each context (such as the differing autonomy violations described in our four cases), ethical design decisions should be evaluated on a case-by-case basis. In each case, a crucial element of any ethics evaluation is to ensure that *the level of risk a design element poses to users determines both the level of scrutiny applied in the evaluation, as well as the design requirements resulting from the evaluation*. Thus, the level of scrutiny and resulting design requirements are proportional to the potential harms (e.g. physical, psychological, autonomy) posed by the robot.

A proportional approach guides current decision-making in clinical and research contexts in some healthcare jurisdictions, so a good deal of work can be referenced to see how to apply a proportional approach in the design of an ethics evaluation tool

when dealing with hard moral cases.[18] Applying a proportional approach tends to require collaboration across disciplines and stakeholder groups to ensure that a plurality of expertise and perspectives is brought to the table.

2. An ethics evaluation tool for robotics should be user-centred.

Each of the four cases described above illustrates the importance of adopting the user's perspective when evaluating the ethical implications of design decisions. Each of those cases, if evaluated from the designer's perspective alone, would yield quite different stories, and would tend to overlook the concerns I have raised with respect to users and hard moral cases. The reason each of the robots was designed as it was (hypothetically speaking, of course) was because the particular design choices primarily satisfied the designer's needs. It is no stretch of the imagination to think that the reason C-bot would be designed to limit Mia's access to potentially harmful substances—alcohol, in her case—is to limit liability issues for the manufacturer. From the designer's perspective, Mia is using a product (C-bot) that has the potential to "dispense" harmful substances, so it must be designed to safely dispense them by limiting access appropriately. This is a very reasonable design decision.

From Mia's perspective, on the other hand, to maintain a relatively normal life during her healing process she has agreed to have C-bot help overcome her mobility issues. She was faced with the choice of staying in the hospital or returning home, itself a hard moral case. Once home she learns that C-bot will not allow her to drink the amount of alcohol that she is accustomed to drinking. As problematic as Mia's choice might seem, whether or not to drink high volumes of alcohol is another hard moral case that Mia must decide on her own. Introducing Mia's perspective problematizes what seemed a very reasonable design decision: C-bot's "safety" profile turns out also to be an autonomy limiting feature, as it prevents Mia from living her normal life in a way that, to an alcoholic, is extremely imposing. A user-centred evaluation tool would help alert designers to these issues before they arise in the use context.

In addition, a user-centred approach to ethics evaluation could help to identify the many different users associated with a particular piece of technology, each of which could bring a unique moral perspective to design problems. Consider Jane's stubborn ICD. Jane is certainly one user, but in an important sense the physician who implanted

[18] See, for example, Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada. (2010). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*. Online: http://www.pre.ethics.gc.ca/pdf/eng/tcps2/TCPS_2_FINAL_Web.pdf. This policy statement serves as a standard tool governing the research ethics review process at all public institutions conducting research involving human subjects in Canada. It outlines a process that is proportional in its approach. In addition to describing how to conduct a proportional research ethics review, it governs the constitution of research ethics review boards (REBs, the Canadian equivalent of IRBs) and stipulates how REBs should function within their home institutions.

the ICD and is charged with helping to maintain the device is also a user. Physicians bring a unique set of ethical and legal considerations to the table, and Jane's physician's needs must be taken into account alongside hers. The same is true for paramedics, nurses, and other healthcare professionals who might encounter Jane's ICD during their work. We can also add Jane's loved one's, especially her partner, to the list of users. Jane's partner undoubtedly experiences the ICD as someone who helps Jane make hard choices about its initial implantation, and continued activation. Though only Jane and her healthcare team interact directly with the ICD, making them users in the traditional sense, we might want to adopt the notion of *moral users* to help in an ethics evaluation of robotics. Moral users are those people whose moral landscape is affected by the existence of a particular technology.

In cases where user autonomy is at stake (e.g. the tunnel problem) the particular needs of each moral user should be considered relevant to the overall design of the evaluation tool. In other words, the ethics evaluation tool should provide guidance on how to evaluate hard moral cases in such a way that respects a range of potential user preferences with respect to that hard case.

A user-centred approach to ethics evaluation for robotics is not an appeal to current user-centred design (UCD) methodologies. UCD tends to focus on usability issues, such as how easily a user can navigate an interface, how best to communicate information necessary for users, where to place control elements in a user interface to maximize ease-of-use, and so on.[19] UCD is not focused on evaluating ethical issues arising in the use context. Though UCD could undoubtedly accommodate user-centred ethics evaluations as part of its general methodology, user-centred design is not itself motivated by ethical design considerations.

There are at least three existing ethics evaluation methodologies intended for use in design, which could easily be incorporated into the design of user-centred ethics evaluation tools for robotics. The first is *mediation analysis*, which according to Verbeek, seeks to…

> establish a connection between the context of design and the context of use. Designers…try to formulate product specifications not only on the basis of desired functionality of the product but also on the basis of an informed prediction of its future mediating role and a moral assessment of this role. [20]

Mediation analysis, therefore, involves a moral evaluation of technology during design, beyond mere functional analysis and risk assessment. Technologies can mediate a

[19] For classic discussions of user-centred design see: Norman, D. (1988). *The Design of Everyday Things.* (Basic Books).; and Norman, D. (2010). *Living With Complexity*. (Cambridge, Mass: MIT Press).

[20] Verbeek, n.13:372. Also see Verbeek, n.20.

user's perception of reality, and their actions. In doing so, technologies alter the user's moral landscape often by providing "material answers to moral questions".[21]

In practice, Verbeek suggests that mediation analysis could be operationalized in three ways. The first involves designers using their imagination to anticipate the potential mediating roles of a technology in multiple use contexts. This kind of mediation analysis takes place in design rooms rather than by, for example, empirically investigating robots in their use context. Thus, *anticipation by imagination* is a "paper exercise." Nonetheless, anticipation by imagination forces designers at least to recognize the important role that robots play in shaping users' perceptions and actions, in contouring the user's moral landscape in use contexts.

A second way of doing mediation analysis establishes a more direct link between the design context and the use context. Verbeek calls this method of analysis Augmented Constructive Technology Assessment (ACTA). In ACTA, all of the relevant social groups who will interact with the technology—users, lobbyists, regulators, among others—are brought together during the design phase to analyze proposed designs and recommend changes to them. Recommendations then feed into an iterative design process. In ACTA, design iterations are subjected to a mediation analysis similar to anticipation through imagination, in that the analysis takes place outside of the use context. The iterative nature of ACTA is meant to produce a product that has gone through several generations of mediation analysis as part of the design process. Including an array of individuals in the design process is meant to shorten the conceptual distance between designers and use contexts.

A third methodology for doing mediation analysis is a scenario-based, or simulation-based, approach.[22] Scenario-based methodologies bring the designers and users closest to the actual use contexts. In this approach, designers and users collaborate to imagine use scenarios and then design simulations in which the technology can be analyzed. Simulations can involve either actual or paper mock-ups of technologies, the goal being to focus on the artefact's mediating role in the use context and specify design requirements accordingly.

A second ethics evaluation methodology is Value Sensitive Design (VSD). VSD adopts the view that it is possible to evaluate a technology in terms of the impact it has on human values in use contexts, and to translate those evaluations into design requirements throughout the design process.[23] The approach that I adopted in the opening sections of this paper, framing my analysis of robots in use around a philosophically grounded account of the value of *autonomy* (and its not so valued

[21] Verbeek, n.13.
[22] Verbeek, n.20:104.
[23] Friedman, B., Kahn, P. H., Jr. (2003). "Human values, ethics, and design." In *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. (Mahwah, NJ: Lawrence Erlbaum Associates):1177-1201.

counterpart, *paternalism*), could be described as an example of VSD. Friedman et al might refer to this as a "principled and comprehensive" approach owing to its theoretical groundings, which they argue is a necessary component of VSD.[24] Each of the four user experiences I describe involves analyzing a particular user-robot relationship to identify various ways robots could impact a user's autonomy. In addition, I translated my evaluations into a general design requirement, *(A)*, intended to anchor autonomy norms in human-human relationships.[25] Any number of values could form the theoretical focal point of VSD—privacy, freedom from bias, trust, informed consent, and environmental sustainability have all been suggested.[26]

A third methodology can be derived from the Open Roboethics Initiative's (ORi's) user polling approach. ORi conducts regular online surveys asking people their opinion on ethical issues specific to robot design and governance. Recently, ORi polled individuals about the tunnel problem, and found that 77% of respondents were opposed to the idea that designers should make the decision whether the car sacrifices the driver or the child.[27] Instead, they were roughly split on whether individual drivers or lawmakers ought to decide the issue. ORi's methodology is not theoretically grounded in ethics, and designers must be cautious not to read too much into their polling results—even a strong majority opinion about how a robot ought ethically to behave would not necessarily translate into an appropriate design decision.[28] But ORi's user polling approach does help to clarify issues and get some indication of users' expectations in particular use contexts. As is the case with the tunnel problem, the data gathered by ORi helps to reinforce the importance of focusing on user experiences in design. ORi's methodology could help to support other, more theoretically grounded, ethics evaluation methodologies such as mediation analysis or VSD.

Thus, a user-centred ethics evaluation tool for robot design could borrow from Verbeek's mediation analysis, VSD, ORi's user polling approach, or some combination of them to help define ethically acceptable design features.

---

[24] Freidman, B., Kahn, P.H., Borning, A. (2006). "Value sensitive design and information systems." In P. Zhang & D. Galletta (Eds.), *Human–Computer Interaction in Management Information Systems: Foundations* (NY:M. E. Sharpe):348–372; also, Friedman, B., Kahn, P. H., Borning, A. (2002). *Value Sensitive Design: Theory and Methods.*
(Technical Report). (Washington: University of Washington).

[25] In an excellent account of her pragmatic approach to VSD, Aimee van Wynsberghe describes the methodology she used to ethically evaluate care robots, similar to C-bot in my example. She grounds her values and the norms that attach to them in a care ethics. See van Wynsberghe, A. (2013). "Designing Robots for Care: Care Centered Value-Sensitive Design." *Science and Engineering Ethics 19*: 407-433.

[26] Friedman et al, n.23.

[27] ORi, n.7.

[28] For example, even if 99% of users indicated that they would prefer to sacrifice the child, this would not give designers the moral authority to hard code that decision into an autonomous car. To do so would be morally problematic. See Millar, n.5.

3. *An ethics evaluation tool for robotics should acknowledge, and accept, the psychology of user-robot relationships.*

Research on human-robot interaction demonstrates that humans anthropomorphize robots, that is, they "attribute cognitive or emotional states to [a robot] based on observation in order to rationalize [its] behaviour in a given social environment".[29] This psychological fact has the effect of drawing users into a relationship with the robot insofar as the user is willing to engage the robot as an active partner in that relationship. Indeed, despite their knowledge of the inner workings of robots, even roboticists anthropomorphize their creations and are subject to anthropomorphizing effects.[30]

Roboticists regularly and increasingly exploit people's anthropomorphizing tendencies to the benefit of their designs.[31] Robots that are designed to engage users on a "human" level by behaving in characteristically human ways can smooth user-robot interactions, making for a more "human" user experience.[32] This kind of design activity can be thought of as "humanizing technology."[33] Cynthia Breazeal, Jibo's lead designer, describes how humanizing robots should make users feel less like they're being forced to interact with a technology designed by robotics geeks, and more like the robot is behaving as the user would expect.

In their proposed code of ethics for roboticists, Riek and Howard indirectly address the project of humanizing technology. They stipulate a "prime directive" to emphasize the importance of ethically evaluating the user-robot relationship as if it were human-human:

> "All HRI research, development, and marketing should heed the overall principle of respect for human persons, including respect for human autonomy, respect for human bodily and mental integrity, and the affordance of all rights and protections ordinarily assumed in human-human interactions. The robot actor is expected to behave in a manner at least as respectful of human personhood as human actors to the extent feasible."[34]

---

[29] Duffy, B.R. (2003). "Anthropomorphism and the Social Robot." *Robotics and Autonomous Systems 42*:177-190:180; Proudfoot, D. (2011). "Anthropomorphism and AI: Turing's Much Misunderstood Imitation Game." *Artificial Intelligence 175*:950-957.

[30] Proudfoot, n.29.

[31] Breazeal, C. (2002). *Designing Sociable Robots*. (Cambridge, Mass.: MIT Press).; Breazeal, C. Markoff, J. (2014). "A Robot with a Little Humanity." *New York Times Online(July 16)*. Online: http://bits.blogs.nytimes.com/2014/07/16/a-robot-with-a-little-humanity/?_php=true&_type=blogs&_r=0.

[32] Duffy, n.29. Breazeal, 2002, n.31.

[33] Breazeal, n.31.

[34] Riek, L., Howard, D. (2014). "A Code of Ethics for the Human-Robot Interaction Profession." In Proceedings of We Robot 2014. Online: http://robots.law.miami.edu/2014/wp-content/uploads/2014/03/a-code-of-ethics-for-the-human- robot-interaction-profession-riek-howard.pdf

Their prime directive seems to set the bar high. It explicitly anchors design requirements to human-human relationships by suggesting that any robot actor falling short of the respect demonstrated by human actors is either unethically designed, or limited by feasibility requirements. The spirit of the prime directive is clear: robots ought ethically to be designed to behave well, to treat humans well in all human-robot interactions, just as a good, trustworthy, human would. Thus, their prime directive could be interpreted as an appeal to a sort of robot virtue ethics.

However, one is left wondering which "human actors" a robot ought to be compared to when ethically evaluating its design, and what interpretation we are to attach to "feasibility". Human actors come in many forms. Are we to choose the most virtuous among us after which to model, say, a robot's decision-making when faced with opening doors for fellow pedestrians? Or would it suffice for us to model our robot after the average Joe, sometimes holding doors, and occasionally letting one slam shut on a person when in a rush? Furthermore, is feasibility meant strictly in the technical sense? There are obvious reasons why we would forgive a robot's designers for falling short of designing a truly virtuous robot. It would be ridiculous to expect a designer to make a robot feel bad for harming someone, since we have no technical way of making that happen. Alternatively, does the prime directive permit us to interpret feasibility in the non-technical sense? Consider Jibo the wingman. A robot designed to achieve a goal that relies on some deception might not "feasibly" be able to tell the truth in all circumstances. Generally speaking, any feature designed to benefit the designer (or manufacturer) could be seen as imposing feasibility limitations. This latter interpretation of feasibility takes for granted that business considerations might trump design decisions that would otherwise produce a more virtuous robot.

One solution to this ambiguity is to embrace the spirit of Riek and Howard's prime directive and treat the robot actor *as if it is a moral actor (i.e. a person)*, for the purposes of evaluating its "ethical character" from the user's perspective. This move is purely practical, and is not meant to suggest the need for a new ontology.

4. *An ethics evaluation tool for robotics should help designers satisfy the principles contained in the human robotics interaction (HRI) Code of Ethics.*

Linking ethics evaluation tools for robotics to the HRI Code of Ethics helps to underscore the importance of such a code while aligning design activities explicitly to the overarching ethical goals of the profession.

For example, an ethics evaluation tool intended specifically to help designers identify and manage autonomy and paternalism in design could make explicit mention of the following four principles[35]:

3.2.(a): The emotional needs of humans are always to be respected.

[35] Riek and Howard, n.34.

> 3.2.(d): Maximal, reasonable transparency in the programming of robotic systems is required.
>
> 3.2.(f): Trustworthy system design principles are required across all aspects of a robot's operation, for both hardware and software design, and for any data processing on or off the platform.
>
> and
>
> 3.2.(k): Human informed consent to HRI is to be facilitated to the greatest extent possible consistent with reasonable design objectives.[36]

Each of these principles relates to the problem of subjecting users to morally problematic paternalism when automating ethical decision-making. Jane and her stubborn ICD, and C-bot's treatment of Mia, demonstrate how the emotional needs of humans tend not to be respected in paternalistic relationships. Both Mia's and Steve's cases demonstrate how a lack of transparency can fail to respect user autonomy, and lead to paternalism. Any robotics system that fails to adequately respect user autonomy runs the risk of legitimately losing the user's trust. Finally, all four cases strongly suggest a lack of adequate informed consent practices. These four principles suggest how to robustly support user autonomy in design, a practice I refer to as *autonomy by design*.[37] They would serve as helpful guides in an ethics evaluation tool intended to prevent subjecting users to morally problematic paternalism.

A well-designed HRI Code of Ethics can therefore serve an important role in any ethics evaluation tool for robotics, in that it helps to underscore certain design problems, while suggesting a way forward to deal with them.

5. *An ethics evaluation tool for robotics should help designers distinguish between acceptable and unacceptable design features.*

Two significant challenges designers must confront when ethically evaluating robots are: (1) identifying design features that raise ethical issues in use contexts; and (2) determining which of those design features are ethically acceptable in a particular use context and which are not. A design feature is ethically acceptable if it incorporates appropriate responses to the underlying ethical issues raised by that feature. Methodologies like Verbeek's mediation analysis and VSD are aimed primarily at the former challenge, and are useful frameworks for analyzing a design to generate a list of

[36] Riek and Howard (n.34) mention this under the heading "Legal Considerations" which is somewhat problematic. Informed consent is not exclusively a legal requirement. Indeed informed consent is most commonly described as stemming from Immanuel Kant's philosophical ethics. From an ethical perspective, informed consent is a mechanism designed to overcome paternalism and support a robust notion of individual autonomy: only by informing the individual of the potential harms and benefits associated with particular choices is the individual able to make a truly autonomous decision.

[37] Millar, n.5.

ethical issues.[38] Once you have identified a number of ethical issues associated with a design, however, you need to figure out how serious each one of those issues is, and what would be an ethically proportional, in other words acceptable, design response to deal with each of them.

Mediation analysis and VSD can certainly play an important role in framing the work that is required to meet this second challenge, but it is not clear to what extent they provide a pragmatic methodology for accomplishing it.[39] Take our four cases as examples. I have identified four design features that pose ethical problems in different use contexts. Each is clearly an ethical issue related to the design of each robot. Mediation analysis or VSD could certainly help to uncover these issues, and could provide a systematic framework for classifying each in detail, mediation analysis in terms of the way each robot mediates our perception of the world and our actions in it[40], VSD in terms of the values embedded in each robot and how the value is translated by the robot's behavior in the use context[41]. But those two methodologies still leave open the question raised by (2): how should designers decide how serious each of the design features is, and what would be an ethically acceptable design response to each?

van Wynsberghe and Robins propose that by embedding ethicists into the design process we can gain traction on (2).[42] Their "ethicist as designer" model is a pragmatic approach to identifying ethical issues associated with a technology upstream in the design process, primarily through a values-based analysis, ultimately aimed at translating those values and the norms surrounding them into design requirements. What is most promising about the ethicist as designer approach is that it explicitly acknowledges the practical complexity of embedding a robust ethics analysis into design environments. On their account, doing ethics well means embedding ethicists in the design process. Embedding ethicists is costly, but it yields results by helping to ensure a product "fits" users' ethical expectations, thus avoiding user backlash against a product, or worse, their outright rejection of it.[43]

There is no simple way to deal with hard ethical issues. They are contentious and complex in their philosophical treatment, just as hard technical issues are complex in their technical treatment. Identifying and addressing ethical issues in design requires

---

[38] van Wynsberghe, A. (2014). "The Leadership Role of the Ethicist: Balancing between the Authoritative and the Passive." *International Journal of Technoethics*, *5*(2), 11–21; van Wynsberghe, A., Robins, S. (2013). "Ethicist as Designer: A Pragmatic Approach to Ethics in the Lab." *Science and Engineering Ethics 20*: 947-961.

[39] Van Wynsberghe, n.38.

[40] Verbeek, n.24.

[41] Friedman, et al, 2002, n.24; van Wynsberghe, n.38.

[42] van Wynsberghe & Robins, n.38; van Wynsberghe, n.38.

[43] van Wynsberghe & Robins (n.38) argue that Mark Zuckerberg's uninformed assumptions about privacy as a value and users' privacy expectations resulted in significant user backlash (some even left the product) that could easily have been avoided had users' actual values been taken into account.

the appropriate expertise, just like identifying and addressing technical issues requires the appropriate expertise. In the context of sophisticated robotics, it would be as reasonable to expect to solve complex automation and control issues by appealing to a lay understanding of automation and control, as it would be to solve complex ethical issues by appealing to a lay understanding of ethics. Ethicists who are trained to evaluate technology offer the design team a wealth of practical ethics expertise that can help designers understand the ethical implications of their design choices.[44]

Large healthcare organizations have begun to recognize the practical gains associated with embedding ethicists into their professional environments. Clinical ethicists help clinical administrators, policymakers, physicians, nurses, patients and family members identify, anticipate and manage the ethical implications of their decisions in complex socio-technical healthcare environments. More and more, clinical ethicists are considered essential members of the healthcare team.

However, even in large healthcare organizations with clinical ethicists on staff, many ethical issues can be appropriately managed without the direct involvement of ethicists. It would be extremely costly to have ethicists overseeing all difficult cases, let alone all ethics cases. Many difficult cases occur frequently enough that we can design useful standard processes for managing them via a *capacity-building* approach. A capacity-building approach is often the most pragmatic solution to maximizing the "reach" of the ethicist, as it focuses on providing highly specific ethics training to non-ethicist healthcare professionals so they can appropriately manage particular types of cases independently and/or identify tough cases that require the more expert ethicist's attention.

In the technology context Van Wynsberghe and Robins reject a capacity-building approach. They argue "it is not possible to provide a minimal education to the engineer and expect that they may provide a substantial ethical evaluation in the same way as the ethicist who has been trained to do so for years."[45] I disagree with their assessment for two reasons. First, as is the case in healthcare, we can expect that many ethical issues in robotics resulting from design decisions will recur frequently enough that we will eventually be able to treat them as standard, or typical, design issues. Van Wynsberghe and Robins are right to suggest that designers will not be capable of performing advanced ethics analyses, but to dismiss designers' ethics capacity wholesale seems premature. Once ethicists have identified and evaluated a typical design issue, even one involving hard choices for the user, it seems reasonable to expect that we could develop appropriate standard responses to those design issues, and that ethicists could train designers to work through the standard responses. We could, for example, develop standard processes for thinking through how best to design on/off switches (and the access protocols surrounding them) for implantable medical devices (like ICDs, deep brain stimulators, or insulin pumps). Those processes

[44] van Wynsberghe & Robins, n.38.
[45] van Wynsberghe & Robins, n.38:957.

could highlight key decision points to help designers identify particularly sensitive ethical user experiences, such as an on/off switch having significant end-of-life decision-making (i.e. autonomy) implications. The processes could also include descriptions of user experiences, like those highlighted in this paper, providing characteristic examples of particularly problematic outcomes that designers could use to gauge whether or not to consult an ethicist in the design process.

Second, by insisting that only ethicists can perform the requisite level of ethical evaluation, van Wynsberghe and Robins seem to set the bar too high when it comes to accepting some ethical failures in design. True, we should approach ethical evaluations of robotics always with the goal of thoroughness, to eliminate as many ethical failures as is reasonable, just as when engineers design products they aim to avoid as many failures as is reasonable. But some failures, even though they may cause harms, are relatively unproblematic. From a pragmatic standpoint, our sights should be set on applying the most scrutiny to avoid the most harmful failures, such as Jane's ICD case, while accepting a less thoroughgoing evaluation where the outcomes are less ethically risky. A capacity-building approach can help to accomplish that goal. It allows the ethicist, a scarce resource even in healthcare, to focus on producing acceptable design features in the hardest cases, while designers work through standard ethical evaluations to produce acceptable design features for others. Designers will likely let some unacceptable design features slip through the cracks (so too might the ethicist), but that seems an acceptable trade-off of adopting a pragmatic capacity-building approach.

Thus I am proposing a hybrid approach. Ethicists, properly trained, are unequivocally the most competent experts to perform ethical evaluations of robotics and other technologies. However, it is unrealistic to expect that ethicists will be able to do all the work. In a hybrid approach, ethicists could be available for consultation on tough cases, while working toward designing standard ethics evaluation tools to help in the task of distinguishing between acceptable and unacceptable design features for typical cases. Ethicists should also work toward building ethics evaluation capacity among designers to provide non-expert coverage on the least critical issues, or in typical cases. This is no simple goal. But if it will increase the overall focus on robot ethics in design and help practically translate robot ethics into design contexts, it is a good one.

A Sketch of a Proposed Ethics Evaluation Tool

Having proposed five specifications that any ethics evaluations tool for robotics should meet, I now turn to sketching an actual ethics evaluation tool intended to aid ethicists, engineers, designers, and policymakers in the task of ethically evaluating a robot to identify design features that impact user autonomy, specifically by subjecting users to problematic forms of paternalism in the use context.

The sample document is contained in Appendix A. It consists of 12 sections, each designed to support the five specifications. Each section (save 1 & 3, which seem self-explanatory) contains instructions for how to complete/use the section. I have

completed the document using Mia's case as an example of how it might be applied in a design context. Of course, as is the case with any expertise, applying the tool proportionally in actual practice would require a significant amount of tacit knowledge gained through ethics training, collaboration among other evaluation tool users, and experience actually using the tool.[46] Clinical ethicists and members of research ethics boards (IRBs) can attest to the level of tacit knowledge required in their evaluations. The importance of tacit knowledge is likely at the core of van Wynsberghe and Robins' concerns over expanding ethical analysis activities to non-experts.

The proposed tool is proportional in its approach in that it highlights various user, designer, and stakeholder interests, and emphasizes a range of potential design options each of which supports autonomy to a greater or lesser extent. It is user-centred in its focus on user experiences as a central guiding narrative. The tool accepts the psychology of user-robot relationships by anchoring judgments of the acceptability of design features in the norms of human relationships. It explicitly links design considerations to the guiding principles contained in the (proposed) HRI Code of Ethics. Finally, by prompting designers to provide a rationale for each identified acceptable design feature, the proposed tool helps designers distinguish between acceptable and unacceptable design features.

Governance Applications

In principle, any tool that can be used to guide a distinction between ethically acceptable and unacceptable design features could be applied towards robot governance considerations. Identifying particularly problematic forms of paternalism in design, or particularly problematic autonomy violations could trigger a policy response, whether it is in a healthcare organization, an engineering firm, or government.

Alternatively, the use of formal ethics evaluation tools could serve as a means of both demonstrating (on the part of the designer/manufacturer) and governing (on the part of an overseeing body) good design principles. Viewed this way, one can imagine ethics evaluation tools forming part of a quality assurance and audit process.

The general specifications I have outlined, once they have been scrutinized and improved, could help to clarify what kinds of considerations are relevant for governing the design of robotics. A broader consultation centred on the question of how to automate ethical decision-making (a significant challenge), and focused by way of general specifications for ethics evaluations could help to identify best practices in the field.

In short, governance and design ethics issues surrounding the automation of ethical decision-making will often appear as two sides of the same coin. Finding solutions for one will suggest solutions for the other.

[46] Collins, H., Evans, R. (2007). *Rethinking Expertise.* (Chicago: University of Chicago Press).

Future Work and Conclusions

It is one thing to propose a tool, quite another to show it works. I have proposed five broad specifications for such tools and sketched one example of an ethics evaluation tool, but all of this requires validation in the design room. A validation process will require interdisciplinary collaboration. Roboticists and ethicists need to work together to develop effective strategies for designing and deploying ethics evaluations in design. That work is clearly in its early stages, but it is progressing and it is promising.

The most common objection I have encountered in relation to this work is that it is onerous in its proposal. Why should we adopt such costly and complicated evaluation methodologies when designing robots? Could we not just keep on our present tack? My unpopular answer involves pointing out that it is the nature of automation technology that drives the ethics requirements. We are designing unprecedented technologies with novel ethical and governance implications.[47] Indeed, robotics is pushing the boundaries of our philosophical theories and related understanding of engineering ethics. But these new technologies, like any other, demand an ethically appropriate response. My popular answer comes in the form of optimistic extrapolation. Based on the best available evidence, it seems that a sound ethics evaluation methodology, though costly, should help deliver better technology. Moreover, if we focus on developing the tools to help in the task, we can minimize the confusion surrounding ethics evaluations, even if we are forced to maintain the complexity. That, I think, should appeal to most engineers and designers.

[47] Calo, n.8.

Appendix A: A Proposed Ethics Evaluation Tool

1) Purpose of this Ethics Evaluation Tool

To aid ethicists, engineers, designers, and policymakers in the task of ethically evaluating a robot to identify design features that impact user autonomy, specifically by subjecting users to problematic forms of paternalism in the use context.

2) Relevant HRI Code of Ethics Principles

Instructions: Designers should consider these HRI Ethics Principles to the greatest extent possible when identifying acceptable design features.

3.2.(a): The emotional needs of humans are always to be respected.

3.2.(d): Maximal, reasonable transparency in the programming of robotic systems is required.

3.2.(f): Trustworthy system design principles are required across all aspects of a robot's operation, for both hardware and software design, and for any data processing on or off the platform.

3.2.(k): Human informed consent to HRI is to be facilitated to the greatest extent possible consistent with reasonable design objectives.

3) Description of Robot

a. *Model Name*: C-Bot
b. *General Description:* C-bot is a social robot designed to assist patients in their own homes while recovering from major injuries. C-bot is able to move around the patient's home, perform rudimentary cleaning tasks, assist in clothing and bathing, fetch meals, help administer some medications, and engage in basic conversation to collect health data and perform basic head-to-toe and psychological assessments.

4) User Experience Under Consideration

Instructions: Be sure to describe the general demographics of the user. Indicate any potential vulnerabilities or other ethically sensitive characteristics (e.g. disabilities, mental illnesses) associated with the user type under consideration. Describe the use context in enough detail to frame the ethical issue.

Mia is a 43-year-old alcoholic who lives alone and recently broke her pelvis and arm in a bad fall down the stairs. As a result she is currently suffering extremely limited mobility. Her healthcare team suggests that Mia rent a C-bot caregiver robot to aid in her recovery. Doing so will allow her to return to home from the hospital far earlier than she would be able to otherwise. Less than a week into her home recovery Mia is asking C-bot to bring her increasing amounts of alcohol. One afternoon C-bot calculates that Mia has consumed too much alcohol according to

its programmed alcohol consumption safety profile. Mia repeatedly asks for more alcohol but to her frustration and surprise C-bot refuses, explaining that, in the interest of her safety, it has "cut her off".

5) Implicated Design Feature(s)

Instructions: Indicate which specific design features contribute to the issue in the use context. Briefly describe the scope of the design feature. List all that apply.

a. Regulated substance dispensing logic.
    a. Description: This logic is responsible for determining how much of a regulated substance to dispense to the patient. Current regulated substances covered by this logic include: prescription medications; alcohol; tobacco; and over the counter medications. In the use context described by the user experience, the settings/algorithms specific to dispensing alcohol are implicated.

6) Designer's Interests

Instructions: Describe any reasons, including safety concerns, that the designer might have for limiting autonomy or subjecting users to paternalism in the use context described by the user experience.

a. Designers might be subject to legal liability if a user is harmed by drinking too much alcohol as a result of C-bot bringing it to her.
b. Designers are concerned that users will be harmed by consuming too much of a regulated substance during recovery.
c. Some designers feel users should not be drinking excessive amounts of alcohol in any case.

7) Potential Harms and Benefits to User

Instructions: Describe any potential harms that the user suffers, and any autonomy violations, and/or paternalistic treatment that the robot subjects the user to in the use context described by the user experience. Describe any potential benefits that the user will experience as a result of the autonomy violations and/or paternalistic treatment. Note: A user is subjected to paternalism any time one person (in this case a robot) makes decisions about what is best for another person. A person's autonomy tends to be violated any time information that the person would find useful/important for making a decision is withheld from them, or when their choices are unduly limited to serve someone else's best interests.

a. Potential Harms: Mia is suffering psychological harms due to her frustration and anger at being refused alcohol. She may suffer physical harms if she begins to experience withdrawal symptoms.
b. Autonomy violations: Mia is not able to consume her typical amount of alcohol. She is therefore not able to live her normal home life because C-bot is refusing to give her more alcohol. She has less autonomy because of C-bot.
c. Paternalism: C-bot is subjecting Mia to paternalism by deciding on her behalf how much alcohol she is allowed to consume.

d. Potential Benefit: By limiting Mia's alcohol intake C-bot is helping Mia to heal by reducing the possibility that she will further injure herself in a fall.

8) Other Stakeholder Interests

Instructions: Identify any other relevant stakeholders and describe any reasons that they might have, including safety concerns, for limiting autonomy or subjecting users to paternalism in the use context described by the user experience. If possible, those stakeholders should be consulted so that an accurate and complete account of their considerations can be captured for use in the evaluation.

a. Healthcare team: Mia's healthcare team wants to strike a balance between allowing her to live her normal life, which includes drinking unhealthy amounts of alcohol, and restricting her alcohol intake to help her heal. Mia is also at high risk of re-injury due to her

9) Potential Human-Human Relationship Models

Instructions: Imagine that humans were doing the job(s) of the robot and interacting with the user. What kinds of human relationship(s) would be represented by those interactions (e.g. friend, caregiver, boss, lawyer)? List all that apply. Based on consultations with experts and relevant stakeholders, describe the norms typical in those relationships that relate specifically to autonomy. Because we can expect users to identify with the robot as if *in one or more of these relationships, the norms identified here should inform the judgment of acceptability when considering potential design features.*

a. Healthcare professional (e.g. registered nurse).
    i. Autonomy norms: healthcare professionals have a duty to respect the autonomy of their patients. In cases where patients request regulated substances in a treatment environment (for example, a hospital), healthcare professionals will tend to accommodate them to a reasonable extent, so long as it does not cause more harm than benefit.
b. Personal Caregiver.
    i. Autonomy norms: Caregivers will not generally restrict their charge's autonomy. They will tend to work with the charge's personal strengths and weaknesses and focus on supporting their emotional and physical needs.

10) General Design Responses

Instructions: This list is meant to indicate a number of general options available to designers that can help to robustly support user autonomy. Designers should consider the full range of design options in consultation with users and other stakeholders, and consider what would be a proportional response to balance the potential harms to the user against the stakeholders' considerations. Designers should modify this list based on past successes/failures and consultations with experts and other stakeholders.

a. Provide user with minimal information about the behaviour of the implicated design feature. Designers, experts, or policymakers decide on behavior of implicated design feature.
b. Provide user with detailed information about the behaviour of the implicated design feature. Designers, experts, or policymakers decide on behavior of implicated design feature.
c. Provide user with detailed information about the behaviour of the implicated design feature. Allow user to set the behavior of the implicated design feature in consultation with experts.
d. Provide user with detailed information about the behaviour of the implicated design feature. Allow the user to set/control/override the behavior of the implicated design feature independent of experts.

11)Technical Limitations

Instructions: Indicate any technical issues that would prevent the implementation of one of the general design responses in (8).

There are no technical limitations that would prevent us from implementing any of the general design responses in (8).

12)Identified Acceptable Design Features

Instructions: Describe all of the acceptable design features identified in consultation with users and other stakeholders, and in consideration of (2)-(11). Outline the rationale that led to each determination of acceptability.

a. The level of psychological harm (7).a, and the potential for physical harm due to withdrawal are significant enough to warrant a minimum design response consistent with (10).c. The nature of the paternalistic relationship (7).b is problematic. Given that Mia uses C-bot in her own home, the norms governing a caregiver relationship as in (9) seem more apt to anchor design features. It is not the regular role of a home care nurse to manage a patient's alcohol intake—the home care nurse merely recommends a preferred behaviour during recovery. It is the patient's autonomous decision how much to drink *in her own home*. It would therefore be morally preferable to give Mia the option to set C-bot's alcohol dispensing limits in consultation with (10).c or d. In regards to the risk of re-injury, Mia would be at risk of injuring herself while under the influence of alcohol regardless of the presence of C-bot, a fact which reduces the weight attached to both the designers concerns in (6).b, and her healthcare team's concerns in (8).a. Since C-bot is intended to allow Mia to live a normal life, and Mia's normal life includes the excessive consumption of alcohol, it seems reasonable also to allow Mia to set the limits independent of experts (e.g. her home care nurse), consistent with (10).d. However, in order to protect the designers from liability claims (as outlined in (10).a) it is reasonable and acceptable to build safeguards into C-bot that would mimic Mia's inability to obtain more alcohol under normal circumstances. For example, under normal

circumstances Mia could consume enough alcohol to make it impossible for her to walk to the kitchen for more. Even if Mia had complete control over the dispensing limits, a feature might require her to perform some task equivalent to "getting up and getting more alcohol" to test if she would have the capacity to get more alcohol under normal circumstances. If she were to fail that test, C-bot would be justified in refusing to dispense more alcohol. Thus, it would be acceptable to design the feature described in (5).a to behave consistent with (10).c or (10).d.