The Heart of the Matter: Patient Autonomy as a Model for the Wellbeing of Technology Users

Emanuelle Burton
Dept. of Computer Science
University of Illinois at Chicago

Kristel Clayville
Philosophy and Religion Dept.
Eureka College

Judy Goldsmith
Dept. of Computer Science
University of Kentucky

Nicholas Mattei
Dept. of Computer Science
Tulane University

Abstract

We draw on concepts in medical ethics to consider how computer science, and AI in particular, can develop critical tools for thinking concretely about technology's impact on the wellbeing of the people who use it. We focus on patient autonomy—the ability to set the terms of one's encounter with medicine—and on the mediating concepts of informed consent and decisional capacity, which enable doctors to honor patients' autonomy in messy and non-ideal circumstances. This comparative study is organized around a fictional case study of a heart patient with cardiac implants. Using this case study, we identify points of overlap and of difference between medical ethics and technology ethics, and leverage a discussion of that intertwined scenario to offer initial practical suggestions about how we can adapt the concepts of decisional capacity and informed consent to the discussion of technology design.

Introduction

Algorithms, robots, and other cyber-physical systems play an increasing role in our everyday lives. These systems already make important decisions that affect our everyday lives: who deserves parole (pro), who is approved for a home loan(Prez-Martn, Prez-Torregrosa, and Vaca 2018), and who is in need of medical care.[1] In the near future, such decision-making systems will be even more deeply integrated into individual and social experience, including driving vehicles (Bonnefon, Shariff, and Rahwan 2016), coordinating the rescue of disaster victims (Imran et al. 2014), and providing care to the elderly (Sabanovic et al. 2013). In ways both large and small, current and in-development applications of technology and artificial intelligence (AI) are altering the basic conditions of human experience.

All of these AI-driven decisions are necessarily predicated on comparative value judgments about human worth and human goods: the importance of children's lives vs. seniors' lives in a natural disaster, or the value of students' security vs. their personal dignity and privacy at a high-risk high school,[2] or the appropriate course of medical care for a terminally ill patient who is physically and emotionally suffering (Avati et al. 2017). These are the same value judgments that transplant teams make every time they prepare to operate, or other scarce medical interventions are allocated (Persad, Wertheimer, and Emanuel 2009). These ideas have received some attention in matching and allocation markets that are mediated by AI applications (McElfresh, Conitzer, and Dickerson 2018; Li 2017). Whether the values are pre-determined by developers or companies through a "top-down" approach or are learned by example through a "bottom-up" machine learning approach (Anderson and Anderson 2011), these automated decisions—and the comparative values that are encoded in the decision making algorithms—will have a profound impact on people's lives and wellbeing.

But what exactly makes a human life valuable and distinctive? What qualities of internal self or external environment need to be in place for a person to be able to live and act as a person? How do particular changes to their environment enhance, or circumscribe, their ability to be a version of themselves that they recognize and prefer? These are highly abstract questions that require careful thought and consideration of many points of view. Even with extensive training and experience in philosophy, it is difficult to formulate an answer that does not rely only on an individual's moral intuition.[3] Relying on intuition can often be a reliable guide in familiar spheres of life, but can be highly unreliable when one is designing and deploying technology for groups they are less familiar with, i.e., young programmers making devices for old or infirm individuals. In addition, technology can and does create new conditions for human experience where no intuition exists. Hence, it is imperative to develop terminology that is clear and useful for non-philosophically-trained technologists, to enable them to realize their goals for bettering human life.

The proliferation of AI in daily life makes it vital that technologists who are designing, building, and deploying these systems are well trained not only to build the technology but also to consider the impacts (Burton, Goldsmith, and Mattei 2018; Burton et al. 2017; Narayanan and Vallor 2014). In this article we argue that technologists also need to be able to think specifically about those aspects of the person that make them recognizable and distinct as people, and furthermore how those human qualities are amenable to improvement, or vulnerable to harm, through specific changes in the conditions of daily life. It is imperative that AI ethics develop its own conceptual tools that can account for the particular ways in which AI can impact the conditions of daily life that affect *personhood*. Equipped with these tools, technologists will be able to discuss both the parameters and significance of the interventions that their designs are mak-

[1] https://www.ibm.com/watson/health/imaging/

[2] https://www.wired.com/story/algorithms-monitor-student-social-media-posts/

[3] However, many philosophers continue to argue that fundamental moral principles are grasped by intuition rather than reason or observation (Stratton-Lake 2016).

ing, and to think more concretely about how design and programming choices can protect and enhance the lives of individuals and societies.

Using technology to enhance, rather than diminish, human lives is made particularly difficult by the knowledge gap between those who build and maintain the technologies and those who use them without the technological expertise to understand how they work. Normal, i.e., non-specialist, users face several disadvantages when confronted with even "easy to use" technology. For example, these users are less likely to be aware of potential security breaches, the signs of such breaches, or the steps they might take to prevent them; these users are also far less likely to be aware of any customization tools that would enable them to fine-tune their experience for their own personal comfort and convenience. Thus, *even at the level of everyday personal technology use, there exists a significant power imbalance between technology experts and non-experts*. The depth and scope of that power imbalance grows exponentially if one also considers those experts' professional work designing, building and maintaining the systems that other users rely on but lack the expertise to understand.

This expertise-based power imbalance, while particularly pressing in technology ethics, is not unique. A similar power imbalance has long existed in medicine, a field whose practitioners need extensive specialist knowledge even as they serve a user base of patients who mostly lack that knowledge. Because of the power imbalance implicit in the vast majority of patient-practitioner relationships, patients are often prevented from making choices about their own care even when doctors or nurses are at pains to leave the choice in the patient's hands (Henderson 2003). To mitigate this problem, medical ethics has developed a family of concepts and practices to help its expert practitioners to navigate the inevitable imbalance in power and knowledge (Quill and Brody 1996). We argue that these concepts can be usefully imported, with some significant revision, into technology ethics (Johnson 2009). Adopting these outlooks can then enable technology developers to identify specific technology design practices that preserve non-expert technology users' capacity for self-determination.

Contribution. In this paper, we described the concept of patient autonomy from medical ethics, as well as the corollary concepts of informed consent and decisional capacity. We use a fictional case study to highlight both the points of intersection and points of divergence between the concerns of traditional medical ethics and technology ethics. We then propose working definitions of informed consent and decisional capacity that are attuned to the central problems facing technology ethics. Finally, we offer some concrete examples of how some current projects in AI and technology are working to support human autonomy and how they could be adapted to support it further.

Autonomy in Medical Ethics

Most western medical practitioners would identify **autonomy** as the central tenet of medical ethics. Autonomy is the principle that mandates **respect for persons**, meaning that individuals have free exercise with regard to whether and what kind of treatment to receive, and honoring this independence is central to contemporary medical ethics (Jonsen, Siegler, and Winslade 2015) . Patient autonomy as a governing concept in medical ethics is relatively recent; the shift toward it and away from medical paternalism was fueled both by broader social movements that sought to empower the individual and by the development of a more consumerist model of medicine as physicians sought to protect themselves from malpractice (Billings and Krakauer 2011).

In practical medical ethics, the term autonomy has two distinct uses, which are related but which also operate independently of each other. The first usage is to affirm that the patient deserves autonomy, the power to exert influence over what happens to them; the second usage concerns the question of whether the patient is able to exercise that autonomy. Because people frequently seek medical care at a moment when they are mentally or physically compromised, *it is not enough to affirm that a patient deserves autonomy. It is necessary for medical providers to take deliberate steps in order to protect the patient's autonomy, and ensure that the patient is able and empowered to make decisions that reflect their wishes, and that their wishes are respected even when they are not capable of asserting them.*

Neither dimension of autonomy—autonomy-as-recognition or autonomy-as-exercise—simply exists as a given. Because of the systemic power imbalance between expert care providers and their non-expert patients, two key constraints have been put in place to ensure that the patient's autonomy is honored in practice as well as in principle. They are **informed consent** and **decisional capacity**.

In the United States, when a patient undergoes a medical procedure, that patient must consent to it, and that consent must follow a conversation in which the doctor explains the procedure's risks, benefits to the patient, as well as other treatment options. After this conversation has happened, the patient signs a document acknowledging that this conversation took place, and the patient is thereby giving *informed consent* to the procedure. Because informed consent documents a conversation, *it is approached as a process rather than a one-time event*. Patients can change their minds at any point leading up to or during the procedure.

No medical procedures or treatments should be undertaken without informed consent, but only patients who have *decisional capacity* can give informed consent. In general, adult patients are presumed to have decisional capacity, but there are categories of patients who lack it. Patients can lack decisional capacity due to age (children), medical status (dementia patients), temporary states (sedated), or institutional status (prisoners). But this absence of decisional capacity is not permanent; children will age into being decisional and able to give informed consent, sedated patients may wake up, and prisoners may be freed, thus enabling them to make decisions free of coercion.

Ideally, a patient who knows they may be non-decisional in the future will prepare an advance directive, a document such as a living will or power of attorney form that either describes their preferences for care or appoints someone else to make those decisions, in the event that they are unfit to make decisions for themselves. In this way, a patient can protect herself from any decisions she might want to make when her ability to decide for herself is compromised.

Paradoxically—or so it seems at first—these limits on a patient's decision-making were instituted precisely to preserve the patient's autonomy, because they place limits on a doctor's ability to manipulate patients into undergoing treatments. The constraints were developed in response to abuses of paternalism, and were designed to constrict doctors' ability to take advantage of patients who were, for whatever reason, unable to exercise their own autonomy.

As medical culture has evolved toward being more patient-centered, the language and conceptual framework of autonomy have likewise been enhanced to focus more on how patients can exercise autonomy, rather than on the constriction of the doctor's. Patients can, in fact, prepare for a future in which they are non-decisional, by creating legal documents that spell out their wishes, should they be incapacitated. They can also cede decision-making power to specified others, for such an eventuality. In the absence of such explicit and legally binding instructions, it is assumed in most societies that a surrogate decision maker from the family can speak for the patient's wishes.

As we will argue, the concept of patient autonomy—and the related concepts of informed consent and decisional capacity—offer a useful model for technology ethics in thinking about how to preserve and enhance the wellbeing of technology users. As the above discussion illustrates, however, the core problems in medicine are not identical to those in technology. *In order for these imported concepts to be useful to technology ethics, they need to be adapted, but in a way that preserves the elements that make them useful.* We use the following fictional case study to illustrate points of overlap and divergence.

Case Study

Joe is a cardiology patient who has two implants: a pacemaker, which regulates his heartbeat, and an implantable cardioverter defibrillator (ICD), which can restart his heart if it stops. This is a common case in the US with over 947 heart related implants per million people (Mond and Proclemer 2011). Some years ago, in consultation with his doctor (as is legally required), Joe requested and was granted Do Not Resuscitate (DNR) status.

Joe is suffering from the early stage dementia. At a recent cardiologist's visit, Joe was told that restarting his heart would be very painful, and in such an event, his heart would likely fail and need to be restarted repeatedly. It was unclear to Joe's doctor, and to his wife, whether Joe understood this information, but at the end of the appointment, Joe asked that his ICD be turned off on his next visit. Then at home, he changed his mind.

Joe's case raises a set of questions that are common to many medical ethics case studies, most of which center around autonomy.

1. How much of his current situation does Joe need to understand in order to participate in his own care?
2. Should Joe's cardiologist remind Joe about his request to turn off the ICD, or wait for Joe to bring it up?
3. Given Joe's (early-stage) dementia, can he be trusted to make decisions about his own long-term wellbeing?
4. Given Joe's dementia, how should caregivers weigh earlier decisions (e.g. his DNR) against what he says now?

The framing of these questions presumes the concept of autonomy: that Joe deserves the right to determine what happens to him, and this right to self-determination must be preserved in balance with medicine's broad imperative to preserve and extend life whenever possible. Joe's right to refuse treatment is recognized, but so is the fact that the very conditions of his treatment may mean that he is not decisional, and thus not fit to make decisions that may harm his person.

But as technologists and those thinking about technology ethics will immediately recognize, this slate of questions excludes some important issues, including issues that might be understood in terms of autonomy. Other questions should be raised pertaining to the security of Joe's personal information and self-direction that are directly influenced by the specific technologies that are now part of his body.

1. Who or what should make the decisions about the use of Joe's devices? In other words, should the defibrillator itself, a control system, or a human monitor (or some combination) be able to decide to not resuscitate Joe?
2. What means, if any, are available to Joe for altering his directives, once they have been entered? How complex or demanding is the process for making changes?
3. Who is responsible for the maintenance of Joe's machines and for the security of Joe's cardiac data? What if Joe doesn't want his data online?

Like the medically-oriented questions, these technological questions also recognizably concern Joe's autonomy as a patient/technology user. The underlying premises of the technology ethicist's questions recognize Joe as an entity deserving of the same sort of autonomy accorded to him by the medical ethics list. But there are two key differences between them. The first is that these questions expand the sphere of Joe's autonomy (in the autonomy-as-recognition sense) to include concerns about his personal information and to consider a wider range of possible agents who might impact Joe's wellbeing. The second difference is that, while these questions broaden the scope of Joe's autonomy as something for professionals to worry about, they constrict its actual exercise by the patient himself (in the autonomy-as-exercise sense). In focusing—appropriately and necessarily—on systems-level concerns such as information security and encryption of medical data, *these questions leave little room for Joe's ability to make decisions for himself, or even to understand what is at stake in the decisions he might make.* Although the questions are about the sphere of Joe's autonomy, they do not create or identify an opportunity for him to exercise it.

The contrast between these sets of questions highlights both how medical ethics could refine its notion of autonomy in conversation with technology ethics, and how technology ethics could benefit by importing the notion of autonomy from medical ethics. With respect to the first dimension of autonomy—recognizing what the patient deserves as a person—technology ethics usefully broadens the sphere of Joe's autonomy by broadening the scope of what counts as Joe's self. In an age when medicine relies heavily on networked technology and data, medical ethics needs to learn from technology ethics' reconfiguration of autonomy.

Yet technology ethics is less well equipped than medical ethics to attend to the second aspect of autonomy, the patient's right to determine what happens to him. A concern for Joe's right to exercise his own particular preferences might lead to questions such as the following: Does Joe understand the capabilities and risks (either to his body or his data) of the devices that have been implanted within him, to a degree that he can make an informed decision about them? Is he aware of the experiences of other patients with similar implantations? Does he feel able to ask his doctors to shut off the implanted devices, to opt out after opting in?

These are the sorts of questions technologists need to be asking when designing and creating the ways in which Joe will interact with the world. The goal of designing many technologies is to provide a one-size-fits-all or lightly customizable system that works for as many people as possible. This goal can come into conflict with notions of autonomy: what if a user does not want any of their data uploaded over the internet? In many ways it can be *more* difficult to respect patient autonomy and/or self-direction in designing and deploying technology than in medicine where there is a responsible professional who implements treatments. Therefore, it is not helpful for technology ethics to simply adopt the concept of autonomy from medical ethics unmodified. And yet, if the human wellbeing of technology users—technology ethics' equivalent of patients—is not to fade from view, it is crucial to identify and clarify a notion of autonomy that technologists *can* use, a definition that is analogous to that in medical ethics but more closely keyed to the problems faced in technology ethics. As technology increasingly sets the conditions for human life, this sort of working definition will prove crucial for technologists who wish to preserve a space for the exercise of human autonomy.

Reframing Autonomy for Technology

As our case study indicates, the high level idea of patient/user autonomy is relevant for both technology and medicine, though the precise definitions will be different. As human lives are increasingly managed at both the macro- and micro-level by smart technologies—and as medical technology itself advances—it becomes pressing for technologists to consider how to enhance, or at least to preserve, users' autonomy. To do so, technologists must consider not only users' right to make decisions for themselves (the first aspect of autonomy), but the conditions that enable them to exercise that autonomy (the second aspect).

In addition, technology ethics also faces particular hurdles in incorporating user autonomy into existing frameworks of inquiry and development. When we compare the two sets of questions in our case study, it becomes clear that individual autonomy can be easily overlooked by products or features designed to make users' lives more easy or efficient. One comparatively low-stakes example is Gmail's new suggested-reply feature. Although this feature does not coerce users into relying on its standardized responses, it rewards them with its convenience, and thus provides a disincentive for users to take the time to craft responses that reflect their individual voice or include non-urgent concerns. Sometimes these designs are conceived to "solve" the idiosyncrasies of how individual users, such as smart home security systems that detect authorized users by gait or biometrics and can therefore fail to recognize an authorized user whose stride or biometric indicators have been altered by mood or physical condition (Rapoport 2013). These hurdles are particularly difficult to overcome in the case of AI, which outsources both large- and small-scale decision-making to embedded and unmodifiable algorithms and models.

A further challenge faced by technology ethics is that there is rarely an appointed human mediator between the user and a particular device. While these pieces of technology are designed to be used easily, they are nevertheless designed to be used independently and by many different people. However, in medicine, complex decisions and processes are implemented and overseen by one or more professionals. No matter how personal technology gets, there will not be a human intermediary in every interaction. Hence, medical ethics is structured around the relationship between patient and care provider, which invests the individual care provider with particular duties and responsibilities. Any useful adaptation of patient/user autonomy must assign responsibility in a manner that is both ethically and ethical and practicable.

The concept of user autonomy can be rendered more manageable when we approach it by way of of informed consent and decisional capacity. These two concepts were developed in medical ethics as a means to preserve the patient's autonomy when her capacity to exercise that autonomy is in some way compromised. Informed consent and decisional capacity make sure that the patient/user's autonomy is maintained even in the presence of disruptive or distorting factors.

Informed Consent

In a medical context, *informed consent* helps to preserve the patient/user's autonomy by requiring the doctor to keep the patient apprised of relevant information, and permitting the patient to rescind consent at any point. Informed consent presumes a user who never develops expertise of her own, and is not penalized for it; the burden remains on the expert-provider to communicate clearly and consistently with the user, to ensure she understands and that her wishes are being honored. While this is not the norm in technology, we are starting to see ideas like this appear. For example, the Android operating system's reliance on *permissions* for every individual application which can be granted or revoked from an easy-to-find screen (Andriotis, Sasse, and Stringhini 2016). Indeed, research in both HCI (Abdul et al. 2018) and technology law (Pasquale 2017) are starting to emphasize the requirement of explainability and transparency in delegated permissions within technology products.

Informed consent presents deep challenges to the basic design principles of technology, because it is deliberately inefficient and resistant to closure. *Whenever technological efficiency is achieved by eliminating the need for the user's input, there is a real risk that the user's autonomy could be compromised.* These challenges arise from informed consent for two main reasons: first, informed consent prioritizes certainty that the patient/user understands over the efficient delivery of information. Second, by allowing the patient/user to opt out at any point, it mandates a structure in which processes are begun but never completed, both because patient/users sometimes withdraw consent partway and because even consenting patients/users retain the option to withdraw consent.

But the inefficiency imposed by informed consent is crucial if the patient/user's autonomy is to be preserved. Because efficiency requires that certain decisions or functions take place en masse for a group of entities without stopping to consult each one, some kinds of efficiency cannot coexist with informed consent (Frischmann 2018). The smarter and more seamless a technology becomes, the more deliberate the technology designer needs to be about maintaining space for this sort of inefficiency. For example, a massive push update to a high-tech medical implant will be much easier to accomplish if the manufacturers assume that the patient/users have already consented simply by having the device implanted. If, however, a patient's condition or wishes have changed, she might not want her implant to be updated.

It is important to note that not all kinds of efficiency are necessarily at odds with informed consent. Many forms of automation increase the efficiency with which the user's goals are achieved without eclipsing her ability to revise her goals or judgments. There is no need for a given technology to build in opportunities for ongoing consent when that technology executes tasks the users already understand and intend to perform, such as washing dishes or taking depth or temperature measurements.

By looking back to medical ethics' notion of informed consent, we can usefully specify *what kinds* of inefficiency are important for maintaining use autonomy. In medicine, the deliberate inefficiency of informed consent affords the patient time to consider (and reconsider) her options in terms of her values and goals. It also forces the care provider to support the patient in this process, rather than imposing decisions upon her. Because the patient's goals or preferences might shift over time or due to changes in her circumstances, the efficient option—taking the patient's initial goals and decisions as a presumptive guide to the future—would undermine her autonomy. Such changes in goals or preferences can be understood as "human" inefficiencies: inefficient or unpredictable changes of character or goals that are essential to a person's autonomy and crucial to preserving their well-being. Medical informed thus consent protects the patient's autonomy by preserving ongoing ability to express her preferences, even when it renders her overall program of care more inefficient. In other words, the efficiency of the treatment process is valuable as long as it preserves or enhances the autonomy of the patient/user, and is potentially damaging to her autonomy insofar as it imposes efficiency on the messy and inefficient processes of self-determination.

Therefore, a usable concept of informed consent for technology ethics is one that enables technologists to consider the specific ways in which a given technology creates efficiency. Does it smooth the user's path to a goal she understands and wants? Does it equip her to understand which sort of determinations are being made for her by automated processes, and to single out the determinations that matter to her for further scrutiny and input? Does it create space for her to revise her engagement with it, should her goals or preferences change? With such questions in mind, a technologist is better prepared to evaluate which kinds of efficiency might categorically interfere with a user's autonomy, which ones require ongoing user input of some kind, and which functions can best serve the user in silent efficiency.

Decisional Capacity

Like informed consent, the notion of *decisional capacity*—the recognition that autonomous users are sometimes not in a state to exercise their own autonomy—can be adapted to technology ethics as a means to preserve and enhance user autonomy. As noted above, medical doctors use a range of criteria to determine whether a patient is decisional, but those criteria have two common denominators: they expect the decisional patient/user to make choices in a manner consistent with their previous character and preferences, and they expect any departures from that prior consistency to be "reasonable"— in line with socially-determined ideas.

Decisional capacity in medical settings is typically binary in nature, because the patient/user's role in the relevant medical process is widely understood to be one of consent, rather than execution. (See, for instance, (Jeste et al. 2007).) If heart patient Joe decides that he wants his ICD turned off, his decisional capacity depends only on whether he is currently capable of making the decision: a medical expert (either Joe's doctor or an ICD specialist) will implement the decision. Joe will be the one to live with the consequences of his choice—which is why he must be decisional in order to make the choice—but his capacity to execute that decision is not a relevant factor. if Joe's judgment is sufficiently consistent with himself, and/or with what is "reasonable," to make what his doctor deems to be a clear-headed decision, then his decision is medically legitimate.

Technology complicates this notion of decisionality because, in most cases, users are also in charge of implementing their decisions. In some cases end-users may be unable to understand the technology of the interface of the technology in a way that enables them to actually implement their decisions. Additionally, it can require some deftness of body and mind to use technology well, which is difficult when in an impaired state: emails typed while drunk, or social media posts made in the heat of anger, can reflect the user's long-standing intent but still fail to realize her goals because her impairment limits her ability to act effectively. Like medicine, technology is a sphere that can magnify the consequences of a given decision; but unlike medicine, technology empowers users to act *without* the mediation of an expert practitioner who can clarify the scope or stakes of the user's action, or handle the niceties of implementation. Furthermore, as Vallor argues, the "sticky" enticements of a range of newly-available virtual goods can cloud a user's ability to distinguish what is most valuable to her, especially in the short term (Vallor 2016).

In most cases, the fact that technology extends the scope of its users' ability to act is the primary virtue. The fact that users are able to take these actions instantly, or near-instantly, is further evidence of the quality of a piece of technology. But these same qualities make users particularly vulnerable to undertaking actions whose technologically-augmented scope exceeds the user's capacity to assess the consequences in the moment of decision. It therefore seems not only helpful but necessary to adapt the notion of decisional capacity for use in technology ethics.

In order to be optimally useful for technology ethics, the notion of decisional capacity needs to be expanded to account for the user's role in implementing their own decisions. It can be helpfully recast for technology ethics as *decisional-executive capacity*, incorporating a second layer that raises the question of whether the user is fit, in a given moment, to undertake an action in a manner that they will be happy with later. Examples of at least checking for this include automatic tone alerts for angry emails and Slack warnings before a message is sent to everyone at the workplace.

Decisional capacity creates an opportunity for AI to enhance the autonomy of technology users and medical patients. As noted above, decisional capacity is quite imperfectly realized in a medical context, as doctors are far more likely to deem a patient decisional if the patient agrees with them. An AI, however, is less likely to succumb to this bias (Hurst 2004). While a doctor's ingrained biases can compromise her assessment of whether her patient is decisional, the doctor-patient relationship is nonetheless a useful model for the AI-user relationship in one key respect. While consistency (the first criterion for determining decisionality) is best judged only with respect to the patient himself, the reasonableness of his wishes (the second criterion) is more broadly culturally determined; what seems like a good reason in one society may seem bizarre in another. Because the human doctor will be influenced by the same broad cultural norms, she is well-positioned to assess whether the patient's expressed wishes fit within those cultural norms, though she is also less likely to be sympathetic to reasons that do not fit those norms. In contrast, an AI that determines decisionality could be structured on universal terms. The ideal approach might call for an AI to learn primarily from local data in order to better assess the reasonableness of expressed wishes.

Proposed Redefinitions and Their Application

As noted above, autonomy has two necessary but distinct dimensions: the first recognizes the patient or user as an agent who deserves the right to exert control over what happens to them, and the second recognizes that the patient/user's autonomy must be actively maintained as well as acknowledged. Here we offer compact formulations of the two concepts crucial for maintaining autonomy.

Informed consent is the process of ensuring that a user understands the terms on which she is engaging with a given technology and is comfortable with those terms. The process is never definitively complete, because users' goals and preferences can always change with time and circumstance. It is also necessarily inefficient, to a degree, because automated decision-making gains much of its efficiency from the presumption that users' expressions of preference are definitive and do not need to be revisited. While few (if any) users would want to trade away efficiency for a thoroughgoing reevaluation of all aspects of their technology use, many if not most users harbor concerns or preferences with regard to some particular sphere of technology use.

Applications of informed consent: In order to balance the imperative toward efficiency against the demands of informed consent, we advocate a tiered model of explanation and consent when users are installing or using an app. Installation wizards that give the user the option of either standard or custom installation processes are a good basic example of this approach. A more fully-realized version would offer users a breakdown of all the component elements of a user agreement or of settings, and offer the option of default or customized setting for each separate component. The user's informed consent could be further enhanced by offering pop-up reminders to revisit each set of agreements or settings at a time increment of her choosing.

Decisional-executive capacity is a concept that can be used to protect a technology user's autonomy, by assessing whether they are in a suitable state to make and implement a decision that matches the larger pattern of their goals for themselves. This term expands the medical ethics notion of "decisional capacity," on the grounds that technology users (unlike medical patients) are typically responsible for implementing the action they have decided is worth pursuing. Therefore, technology users need to be able to do more than select a course of action that aligns with their goals: they also need to be able to implement it in a manner that matches those goals. Incidental circumstances such as being angry, depressed, or inebriated can all interfere with a person's ability to make decisions that they would later affirm; those conditions further compromise a person's ability to execute their aims in a manner that matches their long term goals.

Applications of decisional-executive capacity: One application is to build "advance directives" into the technology in our lives. This would allow decisional users to plan for times when they are not acting in their own best interests. This could be as simple as 15-minute delays on emails or text, to help protect ourselves from angry or drunken message-sending. It could be restrictions we place on where their self-driving cars will take us (e.g., not to circle our ex's block), or it could pertain to our medical care.

Incorporating these concepts presents real challenges: we need to program universal norms for decisionality, but also program systems to learn cultural and societal norms (Conitzer et al. 2017; Noothigattu et al. 2018); that systems can learn both the inherent desires and preferences for an individual, and their patterns of deviance (chemically or age-induced); that we can foresee which technology *needs* the ability to reason about decisionality; and finally, we can program systems to do the necessary reasoning.

Conclusion

There is a growing societal anxiety about artificial intelligence that ranges from fears of loss of jobs for humans to terror that we will be displaced entirely by self-aware, higher-functioning AIs. One strain of this anxiety is that the machines will be programmed with more concern for efficiency than for the wellbeing of the humans they are designed to serve. But these things are not determined yet. What is necessary to balance the drive toward efficiency is a focus on how AI can support the distinctively human qualities of its users. We believe that engineers and computer scientists can learn from medical ethicists, and provide a vital viewpoint to the field of medical ethics itself. Through this, and broader communication throughout the industries and domains where AI is applied, we can ensure that AI can live up to the potential envisioned by its boosters, and become a vital part of the architecture of a better human future.

References

Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B. Y.; and Kankanhalli, M. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

Anderson, M., and Anderson, S. L. 2011. *Machine Ethics*. Cambridge University Press.

Andriotis, P.; Sasse, M. A.; and Stringhini, G. 2016. Permissions snapshots: Assessing users' adaptation to the android runtime permission model. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6.

Avati, A.; Jung, K.; Harman, S.; Downing, L.; Ng, A.; and Shah, N. H. 2017. Improving palliative care with deep learning. In *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*, 311–316. IEEE.

Billings, J. A., and Krakauer, E. L. 2011. On patient autonomy and physician responsibility in end-of-life care. *Arch Intern Med* 171(9):849–853.

Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576.

Burton, E.; Goldsmith, J.; Koenig, S.; Kuipers, B.; Mattei, N.; and Walsh, T. 2017. Ethical considerations in artificial intelligence courses. *AI Magazine* 38(2).

Burton, E.; Goldsmith, J.; and Mattei, N. 2018. How to teach computer ethics with science fiction. *Communications of the ACM* 8(5).

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. In *AAAI*, 4831–4835.

Frischmann, B. 2018. Here's why tech companies abuse our data: because we let them. *The Guardian, US Edition* April 10.

Henderson, S. 2003. Power imbalance between nurses and patients: a potential inhibitor of partnership in care. *Journal of Clinical Nursing* 12(4):501–508.

Hurst, S. A. 2004. When patients refuse assessment of decision-making capacity: How should clinicians respond? *ARCH Internal Medicine* 164(16):1757–1760.

Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; and Vieweg, S. 2014. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, 159–162. New York, NY, USA: ACM.

Jeste, D.; Palmer, B.; Appelbaum, P.; Golshan, S.; Glorioso, D.; Dunn, L. B.; Kim, K.; Meeks, T.; and Kraemer, H. C. 2007. A new brief instrument for assessing decisional capacity for clinical research. *Archives of General Psychiatry* 64(8):966–974.

Johnson, D. G. 2009. *Computer Ethics*. Pearson, 4th edition.

Jonsen, A. R.; Siegler, M.; and Winslade, W. J. 2015. *Clinical Ethics: A Practical Approach to Ethical Decisions in Clinical Medicine. 8th ed.* McGraw Hill Education.

Li, S. 2017. Ethics and market design. *Oxford Review of Economic Policy* 33(4):705–720.

McElfresh, D. C.; Conitzer, V.; and Dickerson, J. P. 2018. Ethics and mechanism design in kidney exchange. Working paper.

Mond, H. G., and Proclemer, A. 2011. The 11th World Survey of Cardiac Pacing and Implantable Cardioverter-Defibrillators: Calendar Year 2009–A World Society of Arrhythmia's Project. *Pacing and Clinical Electrophysiology* 34(8):1013–1027.

Narayanan, A., and Vallor, S. 2014. Why software engineering courses should include ethics coverage. *Communications of the ACM* 57(3):23–25.

Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. D. 2018. A voting-based system for ethical decision making. In *Proceedings of Autonomous Agents and Artificial Intelligence Conference (AAAI)*.

Pasquale, F. 2017. Toward a fourth law of robotics: Preserving attribution, responsibility, and explainability in an algorithmic society. *Ohio St. LJ* 78:1243.

Persad, G.; Wertheimer, A.; and Emanuel, E. J. 2009. Principles for allocation of scarce medical interventions. *The Lancet* 373(9661):423–431.

Prez-Martn, A.; Prez-Torregrosa, A.; and Vaca, M. 2018. Big data techniques to measure credit banking risk in home equity loans. *Journal of Business Research* 89:448 – 454.

Quill, T., and Brody, H. 1996. Physician recommendations and patient autonomy: Finding a balance between physician power and patient choice. *Annals of Internal Medicine* 125(9):763–769.

Rapoport, M. 2013. Being a body or having one: automated domestic technologies and corporeality. *AI & Society* 28(2):209–218.

Sabanovic, S.; Bennett, C. C.; Chang, W.-L.; and Huber, L. 2013. PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. In *Rehabilitation Robotics (ICORR), 2013 IEEE International Conference on*, 1–6. IEEE.

Stratton-Lake, P. 2016. Intuitionism in ethics. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/intuitionism-ethics/.

Vallor, S. 2016. *Vallor2016*. Oxford University Press.