Digital Ethics as Translational Ethics

David Danks
University of California, San Diego; La Jolla, California, USA

ABSTRACT

There are growing calls for more digital ethics, largely in response to the many problems that have occurred with digital technologies. However, there has been less clarity about exactly what this might mean. This chapter argues first that ethical decisions and considerations are ubiquitous within the creation of digital technology. Ethical analyses cannot be treated as a secondary or optional aspect of technology creation. This argument does not specify the content of digital ethics, though, and so further research is needed. This chapter then argues that this research must take the form of translational ethics: a robust, multi-disciplinary effort to translate the abstract results of ethical research into practical guidance for technology creators. Examples are provided of this kind of translation from principles to different types of practices.

Keywords: Ethical Analyses; Digital Technology; Ethical Principles; Ethical Practices; Datasheets; Model Cards

INTRODUCTION

As our world becomes increasingly digital, we must ensure that we do not lose our ethical compass. Algorithms are now frequently used to determine the allocation of critical resources, in some cases even making literal life-and-death decisions. Our lives are measured, collected, analyzed, and stored, thereby reducing us in some cases to simply a set of numbers. We carry digital devices that track our every location, volunteer information to improve our online experiences, and extend our minds, families, and communities through digital means. But much, perhaps all, of these transformations have been designed, implemented, and dictated primarily by technological and economic demands. We risk the creation of digital technologies that serve the values and interests of the few, rather than the values and interests of the many. The recent calls for digital ethics are, at their heart, an effort to return our focus to humans and our values. How do we have the right digital technology for us, and how can we achieve it in practice?

One barrier to the development and implementation of digital ethics has been uncertainty or misunderstanding about the scope and nature of ethics itself. For some people, ‘ethics’ refers to their personal or cultural beliefs and principles. For others, ‘ethics’ is a matter of law, regulation, and compliance. Or perhaps ‘ethics’ is understood as an entirely relativistic domain where there truly are no right and wrong answers, or even better and worse answers. Along a different dimension, some people view ‘ethics’ as a purely negative enterprise that provides only restrictions or constraints (e.g., “killing someone is not morally acceptable except in certain extreme circumstances”). Alternately, one could conceive of ‘ethics’ in terms of positive principles (e.g., “you ought to have freedom of expression”). Of course, the reality of ethics is more complex than either of these possibilities, even if only because many positive principles presuppose restraint or constraint on others. For example, if I have freedom of expression, then others are prevented from stopping my speech (unless there are compelling counter-reasons).

In light of these disagreements about the exact nature of ethics, one might wonder whether any progress is possible at all, let alone in the specific case of digital ethics. Perhaps surprisingly, though, it can sometimes be easier to gain ethical insights in specific domains, rather than always trying to operate at an extremely high level of generality. As we briefly discuss below, 'digital ethics' is not properly understood as the "mere" application of well-established ethical principles to the particular case of digital technologies. Rather, we need to develop ethical principles, guidelines, and practices that are specific to the digital ecosystem. Even with this narrowed focus, though, we might still wonder "what is ethics?" For the purposes of this chapter, I will adopt a relatively simple characterization, with the full awareness of its limitations when trying to do highly abstract and general ethics.

I will understand 'ethics' as primarily concerned with two questions: (1) what values ought we have? and (2) given our values, how ought we act? Neither of these questions will have a unique answer. Ethical reasoning and analysis almost never determine our values; some of my interests and needs are specific to the peculiarities of my situation and life. And even if we know the values and interests of all of the relevant people, there will typically be multiple actions that are morally acceptable. Moreover, I acknowledge that these specific questions lack the subtlety that we normally expect and require from ethics, but they are nonetheless valuable in focusing our attention in productive directions. In particular, they place the focus of ethics squarely where it should reside—on the people who are impacted by the digital systems, technologies, and actions.

One of the background themes of this chapter is that digital ethics is intimately connected with the practices of digital technology creation. We need to ensure that our digital ethics is grounded in the actual challenges that we face in digital spaces, rather than proceeding at an unhelpfully abstract level. The next section thus examines the locations of digital ethics: where do ethical (and societal) questions arise in the digital technology pipeline? The following section then considers how we can translate our insights about people's values and interests into novel and improved practices. In that regard, this chapter can be understood as setting the stage for many of the other chapters in this volume. The practices of digital ethics should not be "mere" implementations of abstract ethical principles. Rather, they should be effective, grounded, focused practices that ensure our values and interests are realized through digital technologies, even if 'ethics' is not a term that appears in many of those practices.

UBIQUITY OF DIGITAL ETHICS

One common (though incorrect) claim is that (digital) technologies are not themselves the subject of ethical analysis. More specifically, this claim—sometimes called the "neutrality thesis" (e.g., Dotan, 2020; Friedman & Hendry, 2019; Simon, 2017)—says that the proper subject of ethical analysis is the *use* of technology, not the technology itself. For example, it seems quite odd to ask about the ethical values of a hammer, though we can surely inquire about the ethics of *using* a hammer in various ways. More generally, we do not normally think or talk about the ethics of particular objects or artifacts, but only the ethics of different uses of those objects or artifacts (though see, e.g., Winner, 1980). One might reasonably think that the same should hold for digital technologies. However, the neutrality thesis cannot hold for digital technologies for at least two distinct reasons.

The first problem with the neutrality thesis is that, in contrast with relatively inert physical artifacts, digital technologies sometimes make ethical decisions themselves. Autonomous and semi-autonomous technologies are increasingly being used across society, and these systems typically have the capability to plan, decide, and act in the world. These systems make ethical choices throughout their operation; in some cases, these choices are literally matters of life-and-death. The most obvious examples are robotic systems such as self-driving vehicles or autonomous weapons systems, but disembodied systems can also make ethically meaningful, entirely autonomous decisions (e.g., loan approvals, medical diagnoses). For these (semi-)autonomous systems, we cannot assign ethical responsibility to the human user of the technology, precisely because there is no immediate human user. Of course, we could try to save the neutrality thesis by ensuring that there is a human involved in these decisions (though that move risks converting the human into a "moral crumple zone" who exists solely to bear moral responsibility; Elish, 2019). If a human were always involved, then we might hope that ethical analysis could remain focused on uses, rather than the technology itself. Unfortunately, this hope is empirically implausible: there are too many contexts and uses that require (semi-)autonomous digital technologies, so we must engage with the ethics of the decisions made by such systems. Unsurprisingly, there is a large (and rapidly growing) literature about ethical analyses, challenges, and principles for behaviors of autonomous systems (e.g., Anderson & Anderson, 2015; Lin, 2016; Millar et al., 2017; Sparrow, 2016; or many papers at the AI, Ethics, & Society conferences). At the same time, though, autonomous systems are still relatively uncommon; most digital technologies involve human decisions at various points. We might thereby hope that the neutrality thesis could be mostly saved, at least if the sphere of autonomous decisions is sufficiently small. Unfortunately, there is a challenge that is discussed much less frequently, but impacts all digital technologies, not just autonomous systems.

The second problem with the neutrality thesis for digital technologies is that the creation of digital technologies involves an enormous number of ethical choices that thereby embed or implement values in the technology itself. Most obviously, digital technologies are almost always designed or optimized for success at a specific task, but 'success' is an ethically substantive term. For example, consider a loan approval algorithm: should it be optimized to maximize societal benefit through credit, or profit for the lender, or empowerment of underserved communities, or some other outcome? We can readily develop algorithms for any of these goals (assuming that we have appropriate data and measures), but almost certainly, we cannot develop an algorithm that maximizes all of these goals simultaneously. That is, the decision about what to optimize is not a technical one, but rather is an ethical one about which problems are more important. And by making one decision rather than another, we have produced a digital technology that prioritizes one ethical value rather than another, regardless of how the algorithm is actually used in the future. This particular example shows how ethical choices during technology development can imbue digital technologies with values, but the challenge for the neutrality thesis arises throughout the technology "pipeline."

There are many different ways of describing how we move from idea to digital technology. For simplicity, a high-level caricature of the technology "pipeline" is provided below. Importantly, nothing significant depends on this particular caricature; any other framework for understanding technology development would lead to the same conclusions about the role of ethics. And although this caricature (including the language of a 'pipeline') might suggest a unidirectional flow from one stage to the next, matters are rarely so simple when building a digital technology. Invariably, insights and developments at a later stage will require us to revisit decisions and choices at earlier stages. With those caveats in mind, consider these six stages in the technology pipeline:

1. *Identify*: What is the problem (or problems) to be solved, and in what contexts, with this digital technology?
2. *Design*: What constraints—technological (including data), financial, legal, regulatory, performance, societal, ethical—do we face in developing this technology? How strong is each constraint, and how do they interact with one another?
3. *Develop*: What digital technology best satisfies these various constraints?
4. *Deploy*: Who has access to this technology (including where and when)?
5. *Use*: How is the digital technology actually employed (and by whom) to solve the problem(s) in real-world contexts?
6. *Revise*: In light of what we have learned, how should the digital technology be adjusted (including potential changes to the contexts of use)?

These stages are obviously not perfectly separable, and as noted above, technology creation rarely moves through them in a unidirectional manner. Nonetheless, they provide a useful framework for recognizing the ubiquity of ethical questions throughout technology creation, including the ways that answers to those question imbue the technology with values. While a complete list of such questions would be far too long to be useful, a briefer survey can help to demonstrate the ubiquity of ethical choices and decisions in digital technology creation.

In the Identify stage, the focus is on the problem(s) that we are addressing with our digital technology. The question "What problems are we trying to solve?" is necessarily an ethical one, as the decision to address problem *A* (rather than problem *B*) implies a value judgment that *A* is more important than *B*. In addition, we need to ask whether a solution to problem *A* would potentially create new problems or challenges. If so, then we need to again make an ethical decision about which problems (including the potential new ones) are most critical. Importantly, these questions and decisions cannot be avoided; we cannot build technology without identifying its intended functionality (including problems that it will address), and that identification necessarily involves values, interests, and other ethical commitments.

The Design phase focuses on the constraints that we face in technology creation, as well as the creative development of design solutions that address those constraints in a satisfactory manner. Ethics and values play key roles in the articulation of a set of design constraints. For example, if we include financial constraints, then we are making the ethical decision that monetary value is important to this technology. The resolution of potential tradeoffs or conflicts between design constraints provides another set of ethical questions. If we decide, for example, that a certain level of performance is less important than avoiding intrusive data collection, then we have thereby made ethical decisions about the relative importance of functional versus privacy constraints.

The Develop stage is typically the primary focus of technology creators, as this stage involves the actual implementation of our digital technology. For example, if we are trying to solve a problem using machine learning, then this stage is when we actually find the best-fitting parameters for the algorithm or model, given our available data. As a result, this phase might seem to be purely technical; the problems, constraints, and designs from the previous two stages required ethical choices, but perhaps development can be the value-free implementation of those designs. However, ethical decisions arise even here. There will typically be multiple ways to satisfy our various constraints, and so development requires us to choose between them. Our constraints rarely specify absolutely every element of the technology, and the remaining decisions can involve an ethical dimension. As a concrete example, our constraints might not dictate a particular color of a button in an interface, but the choice of red versus green could be meaningful in terms of setting an expectation that pressing the button would stop versus start something.

In the Deploy stage, the questions and challenges are principally about access, which inevitably has a significant values component. We are rarely able to provide equal access to the digital technology to all individuals, particularly if we broaden our understanding of 'access' to include full, appropriate access to all capabilities of the technology. But if deployment benefits only some individuals or groups (or benefits them more than others), then we are thereby making a value judgment that some people have greater need for the technology than others. This ethical decision is particularly important to recognize when deployment is market-based. Those individuals who are able to pay do not necessarily have the most ethically important need, yet market-based deployment prioritizes those individuals over others.

Given access to some digital technology, we must then examine how Use occurs with it. Outcomes are a critical part of ethical evaluation of some technology; at the very least, we need to determine whether the technology actually helps to solve the intended problem(s) in the real world. One also needs to ask about potentially unethical variations or differences that result from use, as different contexts or knowledge can lead to radically different outcomes for people with access to the same digital technology. The Use stage also raises questions of oversight and monitoring, particularly for relatively new or under-tested technologies. In all of these cases, values, interests, and other ethical considerations should play a key role in our real-world decisions about how to translate a digital technology from the lab to the real world.

Finally, the development and use of any new technology will inevitably result in problems or mistakes, and the Revise stage provides an opportunity to address those issues. However, that opportunity requires answers to questions about which errors should be fixed, how they should be fixed, whose needs or problems remain unsolved by this technology, and so forth. Each of those questions requires ethical commitments that further shape the digital technology itself.

At every stage of technology creation, we are thus forced to confront ethical issues and questions. We might have hoped that new technologies would not require us to prioritize certain values or resolve conflicts between interests, but that hope is always in vain. Ethical questions are simply ubiquitous in technology creation. We should not pretend that technology could be value-neutral, but instead embrace the ethical questions by explicitly and openly asking them. The challenge then becomes: how do we appropriately *answer* these questions? We turn now to this issue of translating ethical considerations into practical guidance for technology creators, developers, users, regulators, and more.

TRANSLATING ETHICS

There are two intuitively plausible types of strategies to develop practices for ethical creation and use of digital technology. Unfortunately, only one of them is likely to succeed, but it is informative to first consider the problems with the other. In particular, the previous section described the many ethical and value-centric questions that arise at every stage of the technology creation pipeline. At the same time, moral philosophers have developed many normative ethical theories over the past centuries, so we might hope that we could simply translate those theories into practices. For example, we might hope that questions of tradeoffs could be directly translated into the language of our preferred normative ethical theory, and thereby answered by technology creators. Or we could change deployment practices to include explicit, formal evaluation of the ethical permissibility of various possible strategies.

This hope has been most prominently expressed with regards to autonomous technologies such as robots, where many people have hoped that normative ethical theories could be literally written into the system's code (e.g., Arkin et al., 2011; and many others). One might hope, for instance, that a self-driving car could be coded to make explicit consequentialist calculations whenever the system must make an ethical choice, or to decide in accordance with a set of deontological principles. Unfortunately, this hope will not be feasible for most autonomous technologies, as they do not understand the world in the ways that we do. A self-driving car, for example, almost certainly does not use the same concepts as we do in our normative ethical theories. They are simply not programmed in ways that enable the explicit construction or implementation of some normative ethical theory. Thankfully, though, we do not actually need to be able to explicitly code ethics into an autonomous system. If the *human* creators—designers, developers, deployers, users—make ethical decisions in the creation of the autonomous technology, then the resulting system should itself make ethically defensible decisions (all else being equal). So could normative ethical theories play a central role in the human decision-making?

Unfortunately, there are two reasons to doubt the usefulness of explicit, conscious application of normative ethical theories, at least most of the theories that have historically been developed by ethicists and moral philosophers. First, digital technology creation almost always involves substantive values such as my personal interest in connecting with others (e.g., via social networking systems). In contrast, most philosophical work on normative ethics has focused on very high-level and universally applicable values and interests, such as universal human rights. And even when normative ethical theorizing has engaged with more substantive values, it has usually left open the exact weighting or tradeoffs between those values. As such, the normative ethical theories will not provide much guidance to human technology creators. Second, as outlined in the previous section, the technology creation pipeline is quite complex, involving many stages and multiple feedback loops. There are thus many potential places of intervention and decision-making. In practice, there will rarely, if ever, be a single normative ethical theory that is appropriate for all such decisions. The complexity of technology creation precludes the possibility of coordination on a single, universal ethical theory to guide all relevant decisions. We cannot explicitly, consciously use normative ethical theories to make (ethically) good decisions throughout technology creation.

We must find a different way to implement or operationalize our values into good decisions. More specifically, we need a "translational science" of digital ethics that discovers better and best practices to support and enable ethical decision-making at all stages of technology creation. This type of translational ethics is not the simplistic application of existing "basic" research to specific problems. The ethical challenges in technology creation involve value-, situation-, and technology-specific constraints that must be incorporated into our practices. Rather than starting with high-level normative ethical theories, digital ethics in the real world requires the multidisciplinary integration of insights from research in ethics, cognitive science, organizational behavior, sociology, legal studies, and much more. Conversely, we cannot stop with the articulation of high-level "principles for ethical AI," but instead must convert them into real-world, practical guidance. We must draw from a range of different disciplines to develop practices at each pipeline stage that make good decisions more likely. For convenience, we can think about these changes as falling into four broad (and overlapping) categories: people, processes, policies, and partnerships.

First, we need to think about the *People* who are making these varied decisions. In practice, these individuals rarely have training in ethical decision-making. Algorithm developers, for example, usually have a background in computer science or statistics, while technology regulators typically come from policy backgrounds. When ethical questions are raised during technology creation, a common reply is "we know how to build tech, not answer those kinds of questions."[1] Of course, difficulty answering a question does not thereby make the question irrelevant. Ethically impactful decisions are still being made throughout technology creation, albeit implicitly in the decisions to pursue one technological option rather than another. We thus must ensure that the people in the technology pipeline are appropriately trained and empowered.

One pathway to support people is through explicit education of those individuals who are already part of the technology creation pipeline. Many organizations have begun to produce educational materials around broad topics such as "Responsible AI" as well as focused topics such as "Obtaining user consent." Some of these educational materials are suitable for a range of sectors, while others are highly sector-specific (e.g., for highly regulated sectors such as finance or healthcare). These products will undoubtedly continue to grow and improve over time. At the same time, we should recognize the potential limitations of explicit instruction in ethical reasoning and decision-making. There has been a large amount of pedagogical research on teaching engineering ethics and bioethics in a range of contexts. That work has shown that standalone courses can have a positive impact on subsequent ethical reasoning and decision-making, but their impact seems to be notably less than when ethical issues are taught as part of "normal practice" (e.g., Davis, 1993, 2006; Corple et al., 2020; though more studies are needed, see Hess & Fore, 2018). Ethics-specific educational materials developed specifically for technology creators are likely to be

[1] A closely related reply is "we will implement ethics in our systems as soon as the ethicists tell us what to code." As the previous paragraphs showed, though, this reply is misguided in multiple ways.

less effective than integration of ethical considerations throughout their original technology-centric training.[2]

A different pathway to support people is through expansion of teams to include individuals with appropriate ethical (and other) training. Nowadays, user experience (UX) designers are a completely standard part of the development team for a new digital technology, but this was not always the case. Practices have shifted over the past few decades so that any serious development team will have access to people with specialized UX design skills and training. The same shift has not happened for ethics (though some organizations have started down this path), but we can envision a future in which standard practice is to ensure that any technology creation team has access to people with appropriate ethical, psychological, and sociological training to help answer the value-centric questions that are ubiquitous in technology creation. Of course, the viability of this pathway depends on a supply of people with ethical training who also understand the processes of technology creation. Thankfully, many universities are actively developing and deploying educational programs that provide exactly this type of training. Significant open questions remain about how best to integrate these individuals into technology creation teams, but those issues are being actively addressed by researchers in disciplines such as organizational design and industrial psychology.

Second, we must translate ethical considerations into new *Processes* for digital technology creation. In many cases, unethical digital technologies—more precisely, technologies that fail to implement or realize the values that we want—occur because of relatively simple, avoidable errors and decisions in various stages of the pipeline. For example, suppose a developer creates a system that optimizes prediction of student dropout (at a university), but then the user of that system (elsewhere in the university administration) believes that the system predicts which students will have low grades. While dropout and poor scores are correlated, they are obviously not identical. And as a result of this (avoidable) miscommunication between the developer and user, the algorithm could be used in ways that are systematically biased against particular communities (Fazelpour & Danks, 2021). Or consider a decision to measure the performance of a social network platform in terms of mean user engagement rather than median user engagement. This seemingly technical decision can significantly change the performance of the platform, potentially towards being more unethical (i.e., failing to support people's values and interests). In practice, these decisions are often made by a relatively low-level employee who is thinking only about technical considerations, not ethical or societal ones, and so our processes need to shift.

As these two examples suggest, miscommunication or lack of awareness within our pipeline processes can often lead towards unethical technology. In many cases, we do not necessarily need to adjust our processes in deep ways as much as we need to ensure that all of the relevant actors understand the nature and implications of the decisions that they are making. Consider a decision such as frequency of querying a server to find out if anything has changed. This decision might seem unrelated to any ethical concerns, but that will depend on the particular domain of use and application. We cannot know in isolation whether that query frequency will support people's values and interests; knowledge of the role of that choice in the larger system is required. In general, many "standard practices" in digital technology development—modularization of code, changing variable names to abstract codes, etc.—may allow for increased efficiency, but do so at the cost of reduced understanding by the decision-makers (again, perhaps relatively low-level individuals).

[2] Of course, companies cannot travel back in time to change their employees' original education! Standalone ethics-centric training might be the best that they can do at the moment. Nonetheless, organizations should be aware of the limitations of this kind of training, and also the importance of hiring technology-centric individuals whose technology training involved a substantial ethical component.

A natural worry about these arguments is that they risk inducing paralysis within the technology creation pipeline. We obviously cannot spend two weeks analyzing every little decision, just in case it might happen to have some potential (no matter how small) for ethical impact. Rather, we need to find ways to adjust our processes so that they appropriately balance our many different values and interests, including ones like "need to ship our product soon." Thankfully, these kinds of process adjustments are quickly being developed, and many are available for implementation.

One example focuses on a persistent challenge in technology creation: data collection often proceeds separately from data analysis, and similarly for model creation and model use. Different skills are required for the different tasks, and division of cognitive labor is important for organizations. Moreover, there may even be legal barriers to having multiple steps done by the same person, or open communication between the individuals working on each step. Datasheets (Gebru et al., 2018) and model cards (Mitchell et al., 2019) are two process innovations that aim to reduce or eliminate problems that can arise from this kind of distributed work. Both instruments were developed to encode the key information that will be needed for someone else in the pipeline, but without requiring full access to, or disclosures about, the component itself. With a model card, for example, I can know the contexts in which this model is likely to be useful, even if I do not know anything about the inner workings of the model. And while it does take slightly more time to complete a datasheet or model card, the information is typically already known and available to the person completing the card. No additional research or investigation is required, just a few minutes to write up what was collected or analyzed.

A different kind of process adjustment focuses directly on identification of perhaps-unnoticed decisions. If people are unaware of the potential ethical and societal ramifications of a particular choice, then they could easily contribute to problematic digital technologies from ignorance rather than malice. One natural place for this type of problem is in the Identify stage of the pipeline, where one decides that a particular problem is worth addressing with technology. Identification is often done by a domain expert (perhaps in collaboration with a technologist or data scientist) who may not have any particular training in how to recognize or analyze potential ethical impacts (particularly since those impacts might be very removed from the current context). "Ethical triage" processes (Montague et al., 2021) can enable such individuals to quickly determine whether they should talk to someone with ethical (or social scientific) training before starting to Design or Develop. And of course, similar such tools could readily be developed for other stages in the pipeline. These kinds of process adjustments do not tell the person what they ought to do for the technology creation, but instead help them to know when they need to collaborate with others to bend the pipeline towards creation of ethical digital technologies.

Third, our *Policies* must be reconsidered by every actor in this sphere, including technology creation companies and regulators (whether government or private). While our focus in this chapter has largely been on individual- or perhaps company-level decisions, digital technology creation occurs within larger societal structures. Laws, regulations, internal company policies, industry-specific standards, international agreements, and much more all fall into the broad category of policies that can shape the constraints and decisions that are made during technology creation. Moreover, most of these fall under the scope of 'ethics' as outlined earlier, though the values, decisions, and interests are at the group level rather than individual. If we were to ban certain digital technologies, for example, then we presumably would avoid creating unethical versions that fail to support people's values.[3] Policies usually only change slowly, and so they are sometimes thought to be a poor mechanism to influence technology creation. However, exactly because they change slowly, policies can provide a stable context for creation: all of the relevant actors can assume that the relevant policies probably will not change significantly over the relevant timeframes. As a result, policies can have a far-reaching and long-lasting impact.

[3] Of course, we also would miss the potential benefits—ethical and other—that the technology might provide.

One core challenge, however, is that current policies often are a poor fit for digital technologies, for at least two different reasons. First, many policies about technologies were designed for systems that cannot be easily duplicated or replicated, such as buildings or airplanes. The policies assume that there are a limited number of items to be governed, and that there will be time or financial costs to significantly increasing that number. In contrast, digital technologies—particularly those that are "disembodied" such as software or algorithms—can usually be duplicated in a low-cost or zero-cost manner. The creation challenge is to build the first instance, not to create additional instances. As a result, policies about, for example, surveillance systems do not necessarily apply to online spaces since those governance rules assume that there are pre-existing barriers to surveillance expansion, whether physical, logistical, or financial. Online surveillance is extremely low-cost by comparison, particularly when filtered by other digital systems. As a result, the relevant power and value relationships can shift radically compared to what is assumed by the policy. Many current political efforts, including the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA), aim to rebalance these power relations.

Second, many policies assume (perhaps implicitly) that the system will be deployed only in known environments. For example, most automotive standards for safe operation assume that the vehicle is not being driven underwater. As this example shows, the relevant context might be quite broad, but it is nonetheless assumed to be known. In contrast, many digital technologies are developed without meaningful, known constraints on the contexts of deployment and use. In some cases, this lack of constraint is an important design feature for the technology, as it enables it to be (in theory) used in arbitrary new environments. In other cases, this lack of knowledge is actually the reason to look to a digital technology, as when autonomous technologies are developed to make intelligent decisions in unforeseen situations. In either case, we need to consider different kinds of regulatory and policy frameworks since much less is known ahead of time. For example, we can draw inspiration from regulation of pharmaceutical interventions (e.g., drugs), where regulators typically use a dynamic, staged, regulatory process precisely so that they can learn and adapt the use of the pharmaceutical as real-world feedback is obtained. While some adjustments need to be made for digital technologies, the case of pharmaceuticals provides an important proof-of-concept, as well as an analogue, for dynamic regulation of digital technologies to help ensure that they support people's values (London & Danks, 2018). We can translate our ethical needs and interests into society-level changes, not just local practices.

Fourth and finally, the translation of ethical considerations into practice must acknowledge the fact that technology is no longer created by a single individual or company, but rather involves tightly connected *Partnerships* between many different organizations. And just as our Processes must ensure appropriate awareness and communication during digital technology creation, inter-organizational relationships must similarly be adjusted so that the resulting technology use appropriately supports people's values. Moreover, the challenges of awareness and communication are even harder to overcome when working across organizational or institutional boundaries. While different parts of a company might sometimes fail to talk with one another, they nonetheless are (usually) permitted to do so. In many cases, though, the relevant people at different companies are contractually forbidden to speak to one another about relevant details of the digital technology. Hence, there need to be formal mechanisms that enable transfer of key information across these boundaries while preserving intellectual property and trade secrets.

This challenge is now manifesting in a particular issue that we can call *ethical interoperability*. Different organizations are increasingly developing, promulgating, and using particular ethical principles for their digital technology. Of course, as we noted above, these principles need to be translated into practices, but they nonetheless can have significant impact on an organization. For example, the United States Department of Defense has endorsed a principle that all its AI must be "Equitable," thereby incurring an obligation to develop and deploy AI technology in particular ways (and not others). The Partnership challenge of ethical interoperability is: How can we reconcile different sets of ethical technology (use) principles? If organization A develops digital technology that satisfies its principles, then when can organization B use that same technology despite B's different principles? There are many cases where the same technology will satisfy multiple sets of ethical (use) principles, so we need some way to assess ethical interoperability. And it would be natural to try to address ethical interoperability by close examination of the practices to see if they might satisfy multiple, different principles. Currently, however, such assessment methods are essentially completely unknown.

FUTURE RESEARCH DIRECTIONS

This chapter has aimed to provide a framework for thinking about digital ethics as a practical endeavor, rather than a purely theoretical one. This framework also enables us to quickly recognize the many open questions and challenges that we face. Digital ethics can only become a reality if we provide robust, validated practices and interventions for all stages of the technology development pipeline, and across all four Ps (people, processes, policies, and partnerships). Some examples of better practices were provided in the previous section, but those are merely the start. There are many gaps in our knowledge about how to have ethical digital technologies. As just one example, we do not currently have effective, low-cost ways of training people about ethical revision of digital technology. Such training could surely be developed and empirically validated, but that will require future research. More generally, we lack implemented and tested solutions for almost all stage-category combination; one goal for this chapter is to essentially lay out a roadmap for research questions to be asked, all of the form "How can CATEGORY be changed so that practices in STAGE better support people's values and interests?"

Relatedly, we need to work that we identify the best "owner" for these various innovations in practice. Some are best handled by corporate executives, or by developers working directly with data, or by end-users, or by policy-makers, or by some other group entirely. Very few best practices apply universally to all stakeholders, so we need to ensure that our changes are targeted at the right individuals. Many of these "owners" will be defined by their goals and interests, not necessarily their institutional roles, and so we should take additional care to describe the owners in ways that do not presuppose a particular title. For example, a small non-profit organization might not have a "Chief Data Officer," but nonetheless should have someone who focuses on privacy and other concerns about the data that they collect. Or consider the ethics review committees that are increasingly appearing in technology companies (e.g., Microsoft's Aether Committee): these groups can provide important guidance, but we should not assume that every digital technology developer will have access to such a committee. And of course, some of these innovations are likely to have larger impact than others, so we should work to identify those that are more important in the short-term versus longer-term priorities.

As we collectively move towards better practices, we should also aim to be pluralists about our approaches and efforts. Almost certainly, there will not be any single discipline or approach that can solve all of our research challenges. We will need to draw from a range of disciplinary perspectives and methods while still insisting on rigor and clarity. For example, some questions may require the tools of social psychology, while others might draw on methods from analytic philosophy. We should be open-minded about our approaches to digital ethics, aiming to use whatever methods, concepts, and frameworks are appropriate. Digital ethics is necessarily and inevitably a collaborative effort that depends on a range of experiences, disciplines, paradigms, and approaches. As a result, we should strive (as a community) towards a "big tent" attitude that recognizes that contributions and advances can come from many different sources. Diversity of all types will be absolutely critical as we strive towards a robust, applicable digital ethics.

CONCLUSION

The language of 'digital ethics', along with calls for its importance, has become increasingly present in discussions about our digital technologies. There is a growing awareness that *something* is wrong with the ways that we design, develop, and deploy those technologies. And while the exact diagnosis of the problems differs between thinkers, they largely agree that ethics must, in some way, be an important part of the solution. Of course, the exact form and content of that solution is often under-specified, or articulated only at the highly abstract level of ethical principles. A meaningful digital ethics that actually leads to technology that better supports our values involves the conversion of these principles into useful, tangible practices. However, this conversion requires substantive research to translate the abstract insights into useful practices. Just as translational medicine converts biomedical research into clinically useful guidance, we need a translational digital ethics to yield useful changes to our current technology creation pipeline.

This vision of digital ethics is quite different from the most common paradigms currently used in technology creation. Most notably, the vision presented here starts with the recognition that ethical decisions are made throughout technology creation, so ethical analyses cannot be treated as optional or as one last checkbox on the way to deployment. The standard practices for many present-day technologists involve an exclusive focus on the (seemingly) "purely technical" problems to create the technology, followed by a consideration of the ways that it might go wrong. Guardrails and guidelines are added only at the end, and only if there is appropriate time to reflect on the potential ethical and societal impacts. In this paradigm, ethical analyses are comparable to commenting your code: it's a nice thing to do if you have time, but it happens only at the end and very rarely changes anything important. In contrast, we need a digital ethics that is fully integrated into the practices of digital technology creation. A better parallel would thus be user interfaces: everyone now realizes that interfaces are built for essentially all software (even if you wish that you did not have to build one), and so we should employ our best science and theories to build good and usable interfaces. Similarly, digital ethics must be incorporated throughout all technology creation practices.

This type of translational ethics is still quite new in the digital technology space. We currently lack answers to many critical questions, such as how to identify stakeholders in a principled manner or how to determine whether a detailed ethical analysis is required. Thankfully, these kinds of questions are now being asked, and so the translational (digital) ethics is starting to be built. This research will inevitably require inter- and multi-disciplinary collaborations, precisely because digital technologies are, like (ethical) values, now ubiquitous and influential in many areas of our lives. Technologists do not need to know all of the answers, as that would require a skillset spanning many disciplines. But they do need to be able to collaborate closely with people from other disciplines to translate ethical insights into actionable, practical changes in our people, processes, policies, and partnerships.

ACKNOWLEDGMENT

Significant parts of this chapter were written while the author was Professor of Philosophy & Psychology at Carnegie Mellon University.

REFERENCES

Anderson, M., & Anderson, S. L. (2015). Toward ensuring ethical behavior from autonomous systems: A case-supported principle-based paradigm. *Industrial Robot: An International Journal*, *42*(4), 324-331. https://doi.org/10.1108/IR-12-2014-0434

Arkin, R. C., Ulam, P., & Wagner, A. R. (2011). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, *100*(3), 571-589. https://doi.org/10.1109/JPROC.2011.2173265

Corple, D. J., Zoltowski, C. B., Kenny Feister, M., & Buzzanell, P. M. (2020). Understanding ethical decision-making in design. *Journal of Engineering Education*, *109*(2), 262-280. https://doi.org/10.1002/jee.20312

Davis, M. (1993). Ethics across the curriculum: Teaching professional responsibility in technical courses. *Teaching Philosophy*, *16*(3), 205-235. https://doi.org/10.5840/teachphil199316344

Davis, M. (2006). Integrating ethics into technical courses: Micro-insertion. *Science and Engineering Ethics*, *12*(4), 717-730. https://doi.org/10.1007/s11948-006-0066-z

Dotan, R. (2020). Theory choice, non-epistemic values, and machine learning. *Synthese*. Advance online publication. https://doi.org/10.1007/s11229-020-02773-2

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40-60. https://doi.org/10.17351/ests2019.260

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*. Advance online publication. https://doi.org/10.1111/phc3.12760

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for datasets*. arXiv. https://arxiv.org/abs/1803.09010v7

Hess, J. L., & Fore, G. (2018). A systematic literature review of US engineering ethics interventions. *Science and Engineering Ethics*, *24*(2), 551-583. https://doi.org/10.1007/s11948-017-9910-6

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz & H. Winner (Eds.), *Autonomous driving: Technical, legal and social aspects* (pp. 69-85). Springer. https://doi.org/10.1007/978-3-662-48847-8_4

London, A. J., & Danks, D. (2018). Regulating autonomous vehicles: A policy proposal. In *Proceedings of the 2018 AAAI/ACM Conference on artificial intelligence, ethics, and society* (pp. 216-221). Association for Computing Machinery. https://doi.org/10.1145/3278721.3278763

Millar, J., Lin, P., Abney, K., & Bekey, G. (2017). Ethics settings for autonomous vehicles. In L. Patrick, R. Jenkins & K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 20-34). Oxford University Press.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the 2019 Conference on fairness, accountability, and transparency* (pp. 220-229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Montague, E., Day, T. E., Barry, D., Brumm, M., McAdie, A., Cooper, A. B., Wignall, J., Erdman, S., Núñez, D., Diekema, D., & Danks, D. (2021). The case for information fiduciaries: The implementation of a data ethics checklist at Seattle Children's Hospital. *Journal of the American Medical Informatics Association*, *28*(3), 650-652. https://doi.org/10.1093/jamia/ocaa307

Simon, J. (2017). Value-sensitive design and responsible research and innovation. In S. O. Hansson (Ed.), *The ethics of technology* (pp. 219-236). Rowman & Littlefield International.

Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs*, *30*(1), 93-116. https://doi.org/10.1017/S0892679415000647

Winner, L. (1980). Do artifacts have politics?. *Daedalus*, *109*(1), 121-136.

ADDITIONAL READINGS

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*. https://doi.org/10.1177/2053951720949566

Fjeld, J., & Nagy, A. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center for Internet & Society. https://cyber.harvard.edu/publication/2020/principled-ai

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389-399.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (pp. 59-68). New York: Association for Computing Machinery.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review, 41*, 105567.

KEY TERMS AND DEFINITIONS

Ethical Analyses: Systematic descriptions of the ethical risks, benefits, challenges, and opportunities of a particular technology.

Digital Technology: A product or artifact, perhaps non-physical, that manipulates digital representations to accomplish specific tasks.

Ethical Principles: Commitments or beliefs, often quite abstract, that guide ethical decision-making across a range of domains.

Ethical Practices: Patterns of behavior that lead to more ethical decisions, actions, and outcomes.

Datasheets: A framework and tool for representing and encoding key features of data so that others can use those data responsibly and ethically.

Model Cards: A framework and tool for representing and encoding key features of AI models so that others can use those models responsibly and ethically.