ETHICALLY
ALIGNED
AI

AI Ethics: From Theory to Practice

An Ethically Aligned AI Report

Trystan S. Goetze, Ph.D.
trystan.goetze@dal.ca

April 2021

Contents

Chapter 1

Introduction

The aim of this report is to collect some of the important theoretical and practical work in AI ethics, to inform the creation of new AI ethics tools, processes, and theory. The general structure of the report moves from more theoretical to more concrete applications, in recognition of Morley et al. (2020)'s observation that AI ethics discussion has tended to focus on the "what" rather than the "how" – on the ethical problems and risks that arise when AI systems are implemented, and on the ethical principles and codes that developers should follow, rather than on the practical steps, tools, and training needed to actually follow those principles and codes in order to prevent or mitigate those problems and risks.

Accordingly, the report surveys 22 AI ethics tools that have been developed in a variety of contexts. Prior to this discussion, however, the report surveys recent sociological and philosophical research on AI ethics that forms the background of contemporary AI ethics discussions within and outside the academy. The reason for this is that there will be times when none of the available tools is apt, and no principles provide adequate guidance. In cases like these, it is important to have more general, robustly defensible ethical principles to appeal to. To aid in this, advice is also provided on how to make practical use of these theories. After this abstract discussion and before turning to the ethical tools is a discussion of recent arguments made by researchers at the Oxford Internet Institute. Their work has concentrated on how best to bridge the gap between the theoretical and the practical in AI ethics, and thus serves as a useful transition between theory and tools.

The tools discussed are grouped by their general methodology:

- Ethical Design and Review Processes
- Ethics Checklists and Frameworks
- Information Slips
- Ethics Teaching Tools
- Assessment and Auditing Tools

Each is summarized, and, where applicable, specific limitations are discussed.

Finally, it is important to note that no one off-the-shelf tool – or package of tools – should be thought of as *the* solution to any particular AI ethics case. As socio-technical systems, AI systems involve not just technological complexity, but also sociological complexity. That is to say, the *social contexts* in which AI systems are developed and deployed are just as important to understanding how to make and use them ethically as their *technical* considerations. This commitment informs the selection of tools and theories presented below.

Chapter 2

Theoretical Background

This chapter very briefly covers some of the theoretical background informing the selection of and commentary on the tools in the following sections. Some initial thoughts on how to apply these theories in practice in combination with or as an alternative to specific tools is also provided.

2.1 Socio-Technical Systems

In the academic study of technology ethics in general, and computer ethics in particular, the technologies under discussion are usually conceived as *socio-technical systems*. This theoretical approach comes from science, technology, and society studies (STS), a humanistic and sociological discipline of understanding the relationships between technology and society. The basic idea is that no technology is simply an object or artefact. All technologies depend on people and social structures to exist, many depend on these to continue to function, and many are expressions of the values of the people and structures that produce and maintain them.

On a classic STS and computer ethics website, Chuck Huff outlines how computing technology is a combination of physical, informational, human, and social components, along with the relationships between them. When describing a technology as a socio-technical system, Huff suggests that one must consider the following components, and how they are connected (Huff, 2001):

- Hardware
- Software
- Physical surroundings
- People
- Procedures
- Laws and regulations

- Data and data structures.

To this we can add the socio-political context, the moral and political values held by the people and social structures involved, and social power structures.

In their widely used computer ethics textbook, Deborah Johnson and Keith Miller summarize the insights of the socio-technical systems perspective as follows (Johnson and Miller, 2009, 13–18):

- Society and technology shape each other. It is not the case that technology develops along some pre-defined path from one advancement to another, regardless of social context, nor is it a foregone conclusion that the introduction of some technology will shape a society in wholly predictable ways.

- No technology is merely a material object. All technologies are socio-technical systems.

- Technology is not neutral. Rather, technologies are infused with values. The kinds of actions that new technologies enable or constrain reflect those values.

Thinking about AI systems as socio-technical systems is essential to an ethical analysis of them. The people who create training datasets, testing benchmarks, and machine learning models are just as influenced by their social context as any other innovator. And no AI system is deployed in a social vacuum: even systems created simply for the purpose of research will have some social impact. We as a society also have decisions to make about how we wish the adoption of AI technologies to shape us. By coordinating the relationships between the values and contexts of both technologists and stakeholders, AI systems can play a role in improving human life while minimizing collateral damage.

2.2 Bias in Computer Systems

A hot topic in 21st century AI ethics is bias in datasets and machine learning models. But the notion that computer systems could be biased is not new. In a landmark paper published in 1996, Batya Friedman and Helen Nissenbaum present three kinds of bias that can arise in computer systems (Friedman and Nissenbaum, 1996). While their examples are now somewhat dated, precisely the same patterns of bias are now observable in contemporary computer systems, and AI systems in particular. Their conceptual framework is thus a useful one to include.

Friedman and Nissenbaum define biased computer systems as “computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others” (Friedman and Nissenbaum, 1996, 332). They then detail three kinds of bias:

- **Pre-existing Bias**: This type of bias is rooted in already existing societal bias, such as structural racism or misogyny. The same bias ends up being reproduced in the computer system. Friedman and Nissenbaum illustrate with the example of “digital redlining,” where credit applications from African Americans were disproportionately rejected by computerized credit check systems. The reason for the bias was an old, explicitly racist practice of labelling neighbourhoods with higher populations of African Americans as “high-risk.” A contemporary example is predictive policing AI systems that are trained on arrest data without considering the fact that human police officers arrest African Americans at a higher rate than the rest of the population. The resulting AI system determined that African American neighbourhoods were higher-crime areas and should be patrolled more frequently (McGrory and Bedi, 2020).

- **Technical Bias**: This type of bias arises because of technical limitations of the computer system, such as display size or resolution, processing speed, or the capabilities of a machine learning model. Friedman and Nissenbaum illustrate with a case of an airfare booking service that organized tickets alphabetically by airline name, with many other choices being cut off at the bottom of the screen. The airlines need the top of the list – e.g. Air Canada or British Airways – are thus clicked and booked more frequently than airlines towards the bottom of the list, even when they offer a better deal for their customers. A contemporary example is the limitation of facial recognition algorithms trained on unrepresentative photo datasets, which cannot recognize faces with darker skin tones (Buolamwini and Gebru, 2018). While also a pre-existing bias (cf. the inability of photographic film to capture darker skin tones or Māori tattoos accurately (Weber, 2019; Caswell, 2015)), this example also shows a technical bias because it arises due to the inability of the machine learning model to extend its training to new cases.

- **Emergent Bias**: This type of bias arises when the context in which a computer system is used changes. This may occur when the testing conditions are significantly different from the conditions in which the system is actually used, or it may occur when the system moves from one context of use to another. Friedman and Nissenbaum illustrate with the example of the National Resident Matching Program (NRMP). This program matches recent medical school graduates in the USA with hospitals and clinics for their residency training, using a computer system to find matches between graduates’ preferences and the needs of medical institutions. The system worked reasonably well until the number of women in medical programs increased, leading to a larger number of couples with both partners graduating from medical school in the same year. The NRMP was set up in such a way that spouses could have both of their preferences taken into account, but the system also required one spouse to identify as the “primary” spouse, whose preferences would be weighted more strongly. As a result, the system would sometimes recommend matches that were suboptimal: if the primary spouse’s first choice of hospital was a match, then the system would find the closest match for the secondary spouse at a nearby institution, even if both spouses could get their second choice in another location and thus a more optimal result. A contemporary example comes again from the use of facial recognition systems. Many of these systems are currently used in social media platforms, such as Snapchat or Facebook, to allow for applying filters or automatic tagging of people in photos. In this context, the system

> is convenient and benign, even entertaining. But if deployed for surveillance technology – as has been seriously suggested – these same AI systems would present a significant threat to civil liberties (Stanley, 2019).

Friedman and Nissenbaum provide a number of recommendations for avoiding or mitigating these three kinds of bias. These suggestions will reappear throughout the tools surveyed below.

To mitigate pre-existing bias, technologists need to be aware of the biases that exist in society, and how those biases affect their own behaviour and cognition. Increased workplace diversity and diverse focus groups can help catch these biases, because drawing on a more diverse range of social experiences enables bias to be more easily detected.

To mitigate technical bias, technologists need to be aware of the context of use of the systems they design, and test their systems with those contexts in mind. They should be looking for technical limitations in the context of use and test for potential bias that may appear as a result of these limitations.

To mitigate emergent bias, technologists should test in a wider range of contexts of use than initially envisaged, or specify the ideal operating context and anticipated risks of using the system in a context it was not designed for. Technologists must also actively monitor the performance of their systems in their actual contexts of use and take action to correct for unanticipated biases that may emerge.

2.3 Anti-Oppression Theory and Praxis

Several important contributions to our understanding of AI ethics, data ethics, and computing ethics have come from feminist social science. These projects emerge from concerns about how the under-representation of women in technology, especially women of colour and queer women, contributes to bias and harm in the design and implementation of AI systems and other technologies. Of these I will outline three projects: Safiya Umoja Noble's criticism of Google Search, Ruha Benjamin's concept of the "New Jim Code", and Catherine D'Ignazio and Lauren Klein's manifesto for feminist data science.

2.3.1 Algorithms of Oppression

Safiya Umoja Noble's work on algorithmic bias in Google Search begins with an essay published in *Bitch* magazine (a feminist pop culture and criticism periodical) on her discovery of a troubling pattern in the way the search engine returns results for queries about racialized girls and women. She writes that she often asks her students to imagine being a Black mother looking for information on topics of interest to her daughters, or being those daughters themselves. In such a situation, a Google search for "Black girls" seems like a natural move. But, at the time, such a search yielded primarily pornographic results. While some may think it naïve of Noble to be shocked by this result, it is worth stepping back to ask why we *wouldn't* be shocked. After all, there is nothing suggesting sexual content about the search term "Black girls"

– at least, outside the context of a search engine designed without considering the interests and needs of Black women and girls. As Noble writes,

> I consider myself far from prudish. I don't care if someone types "porn" into a search engine and porn is what they get. I do care about porn turning up in the results when people are searching for support, knowledge, or answers about identity. I care that someone might type in "black girls," "Latinas," or other terms associated with women of color and instantly find porn all over their first-page results. I care that women are automatically considered "girls," and that actual girls find their identities so readily compromised by porn. (Noble, 2012, 38)

While Google has since tuned their Search algorithm to suppress pornography when one searches for "Black girls" and "African girls," this is not the case for similar searches: a private tab Google search I just did (29 March 2021, from Toronto, ON) comes up with a mix of arts and culture sites and dating sites when I searched for "Latinas," mostly sexual material when I searched for "Asian girls," and mostly photos of beauty models when I searched for "Indian girls."

Noble has since expanded this initial anecdotal research (Noble, 2018). She argues that Google Search has contributed to reinforcing racism in various ways, drawing on material from Black feminist theory and information science. She points out how the sexualized search results for "Black girls," among other search queries that contain a marker of social identity, very often line up with stereotypical images. Black women and girls are often hypersexualized, rarely associated with powerful social roles or technical skills, and more likely to appear in image searches when one searches for domestic or service work. She also discusses a case where searches containing the term "Jew" often turned up anti-Semitic content, and Google's failure to respond to calls to adjust their algorithms to direct searchers away from such harmful and offensive material. These biases can be considered what Friedman and Nissenbaum call pre-existing biases, since they replicate existing patterns of racism. But we can also see them as emergent biases: a search tool initially designed by (and for) heterosexual men works well for their purposes, but is revealed to have sexist and racist biases when the tool fails to be useful to the purposes of racialized women and girls.

Noble also contrasts the role of Google Search to that of a reference librarian. The librarian's role, she observes, is to be a source of information as part of an institution that serves the public good, working under well-established professional norms that demand that libraries supply current and accurate information that is appropriate to their client's needs. Google, on the other hand, is a for-profit private corporation, which serves its shareholders rather than the public good, and has no established norms regarding the accuracy or appropriateness of the content it serves. When combined with philosopher Michael Lynch's observation that we increasingly delegate various information retrieval and other cognitive processes to Google – a phenomenon he calls "Google-knowing" (Lynch, 2017) – Google's role in our lives becomes troubling. We all treat Google as if it were a neutral source of knowledge, but it is not, for two reasons. First, it is dependent on other people, who are not neutral in their decisions regarding what to put online. Second, Google's decision-making is ultimately driven by profit from their advertising business, to the point where they have allowed misleading and harmful information to persist in their search results. Noble suggests in order to counteract these problems, we need a public option for search that would be accountable in ways similar to reference librarians.

2.3.2 The New Jim Code

Sociologist Ruha Benjamin argues that many new technologies, particularly machine learning models employed in social and political decision-making deepen, obscure, and accelerate already existing social inequities. In a reference to the racist "Jim Crow" laws in the post-Reconstruction US South, she describes these technologies as creating a *New Jim Code* (Benjamin, 2019b). While AI systems are often toted as taking the human element out of decisions and therefore removing human bias, when these systems are developed from data produced by humans with those very biases, precisely the outcomes we sought to avoid are reproduced and, often, worsened. Coupled with what Benjamin calls the *imagined objectivity of technology*, the misconception that because a decision was made by a computational system, it must be more objective or less biased, the risk posed by these technologies is heightened and hidden.

For example, elsewhere Benjamin discusses a health care algorithm that was intended to automate decisions about which patients should receive additional resources (Benjamin, 2019a). The model was based on the assumption that patients whose prior health care was more expensive would be at higher health risk, and therefore would stand to benefit most from additional resources. However, the training data reflected a pre-existing bias in health care: health care providers are typically less willing to recommend spending additional resources on non-white patients, leading the algorithm to drastically underestimate the health risk of people of colour.

2.3.3 Data Feminism

In their extraordinarily well-received book, Catherine D'Ignazio and Lauren Klein outline a manifesto for feminist data science (D'Ignazio and Klein, 2020). Drawing on feminist epistemology and philosophy of science, feminist analyses of power, and anti-racist, LGBTQ+ rights, and women's rights activism, they argue that a failure to adopt feminist theory and praxis into data science and the technology sector more generally has led to various high-profile harms in recent years. Through a series of rich examples, they demonstrate both their critique, *and* the potential for data science and the technology sector to contribute to social justice if feminist principles are adopted in those spaces. The book is available in physical and electronic format from MIT Press, but is also available for free from `datafeminism.io`.

D'Ignazio and Klein summarize their approach to data science in the following *principles of data feminism*:

- **Examine Power**: Data feminism begins by analysing how power operates in the world.

- **Challenge Power**: Data feminism commits to challenging unequal power structures and working toward justice.

- **Elevate Emotion and Embodiment**: Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world.
- **Rethink Binaries and Hierarchies**: Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.
- **Embrace Pluralism**: Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.
- **Consider Context**: Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.
- **Make Labor Visible**: The work of data science, like all work in the world, is the work of many hands. Data feminism makes this labor visible so that is can be recognized and valued.

Each chapter of their book provides theoretical background for each of these principles, and examples of both data science that lives up to the principle being showcased, and data science that flouts those principles.

2.4 Ethical Theories

In normative ethics – the philosophical study of right and wrong, good and bad, morally speaking – it is common to distinguish between several clusters of moral theories. A detailed discussion of normative ethics is beyond the scope of this report; however, it is still useful to introduce these philosophical accounts for practical AI ethics, for several reasons. First, there will be times when the use of an ethics tool fails to produce clear guidance. Second, there will be instances where different ethics tools suggest different answers to a particular problematic situation. Third, because AI and other computing technologies develop and change rapidly, situations will arise where there are no tools applicable to the problem at hand. In each of these cases, falling back on the more general principles at the core of moral theories is important for making ethically justified decisions, and for developing robust new tools.

This section outlines some of the main theoretical divisions in contemporary normative ethics, and closes with a discussion of how to make practical use of them.

2.4.1 Consequentialism

The *consequentialist* family of ethical theories is based on the notion that when evaluating whether an action would be right or wrong, we should concentrate on the consequences that action would bring about. Different consequentialist theories emphasize different kinds of consequences as morally valuable or disvaluable. Some emphasize a single kind of valuable consequence (and its disvaluable opposite), such as John Stuart Mill's utilitarianism, according to which all moral value or disvalue can be evaluated using the following principle:

> actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure, and the absence of pain; by unhappiness, pain, and the privation of pleasure. (Mill, 1879, Ch. 2)

Other forms of consequentialism are more pluralistic, admitting a wider range of morally valuable consequences. For example, G. E. Moore's moral philosophy recognizes beauty and love in addition to the production of pleasure and elimination of pain as goods that we should strive to produce in the world (Moore, 1903).

Considering the consequences of their inventions is crucial for technologists working on AI systems. Too often, enthusiasm for new developments leads to AI systems being pushed to market before the risks of implementing them in different contexts are properly considered, with harmful results. Technologists must consider both the potential happiness and harm that their inventions may cause, and determine whether the good produced outweighs the harm. This process is essentially what utilitarians describe as the procedure to be followed when determining the right action. In addition, pluralistic accounts such as Moore's remind us that we must consider what other valuable or disvaluable consequences might result from implementing an AI system, beyond the production of happiness or harm, such as environmental destruction or even beauty.

2.4.2 Deontology

The *deontological* family of ethical theories concentrates on the notion of moral duties. The most famous exponent of this approach is Immanuel Kant. On his view, the consequences of an action are unimportant; the only thing "which can be regarded as good without qualification [is] a *good will*" (Kant, 1785, 393). And for Kant, having a good will has fairly stringent requirements: the agent must act *purely* from a sense of duty, that is, from a commitment to doing what the moral law commands. The moral law, on his account, is something human beings discover through the exercise of their reason. This process, he thinks, will lead us to, among other things, the *categorical imperative*, a moral rule that applies in all situations. The most straightforward form of the categorical imperative is known as the *formula of humanity*, which states that we must never treat other persons as mere means to our ends. While we may employ the services of others to serve our interests, we must also respect their autonomy as rational creatures to decide for themselves what is valuable and how to act. Kant's moral philosophy also goes into detail on various kinds of more specific duties we may have towards ourselves and others.

Despite the dominance of Kant in contemporary deontology, deontological thinking appears in many other areas of moral philosophy that are not explicitly tied to Kantianism. For example, it is common to invoke the notion of a having moral duty (or responsibility) to do something, or to emphasize the importance of having good will towards others, without any reference to the categorical imperative. There are also non-Kantian deontologists, such as W. D. Ross; his work, which is critical of both

Kant and utilitarianism, emphasizes duties to keep our promises, to make amends when we do wrong, to return the favour when someone does something kind for us, to promote good in the world, and to do no harm to others (Ross, 1930).

What deontology adds to a picture of AI ethics is the notion that there may be moral duties for technologists to fulfill, regardless of the consequences. The formula of humanity in particular implies that there are lines that must not be crossed. Technologists must ensure that AI systems they develop do not undermine the autonomy of stakeholders, and that when they gather data for training AI models, they do not treat their data sources merely as means to the end of training the model. Particular duties such as those suggested by Ross also imply that technologists have a responsibility to make amends when things go wrong.

2.4.3 Virtues and Vices

Whereas consequentialism and deontology begin their analyses from a focus on what makes actions and states of affairs good or bad, the point of departure for *virtue and vice theories* is a focus on the *traits* that make a person good or bad. In the West, the most influential account of virtues and vices is Aristotle's (2011; 1999), but virtue-ethical theories have also been developed across the globe by figures such as the Buddha, Confucius, and Lao Tzu. On Aristotle's view, the traits of character that are considered virtues are those which help human beings to flourish and achieve excellence in life. For him, this means a life of practical action that is guided by reason. The kinds of traits the help us to achieve such excellence include courage, generosity, truthfulness, temperance, and having a good sense of humour.

Philippa Foot adds to this account that each virtue is a corrective to tendencies we have (either from human nature or because of social pressures) towards less than ideal behaviour (Foot, 2003). Hence, as Aristotle tells us, each virtue is opposed by two vices, one of excess and one of deficiency, with the virtue occupying the appropriate place between the two. Courage, for example, lies between cowardice and rashness, but is closer to rashness because we more often have to stand up to our fears than we have to back down from perceived dangers.

The other traditions in normative ethics tend to be quite explicit that the rules and principles at their cores are meant to apply universally, for all people, in all situations. A distinctive feature of Aristotelian virtue ethics is a rejection of this universalism in favour of contextualism. What constitutes virtuous behaviour varies depending on the person and the situation; the best we can do for general principles is offer a sketch. Knowing what to do in any given case requires wisdom, which can only be acquired through practice.

Virtue ethics has been adapted to applied and professional contexts by narrowing the scope from the traits that make someone a good *person*, to the traits that make someone a good *professional* in their field. As Justin Oakley and Dean Cocking write,

> what counts as acting well in the context of a professional role is...determined by how well that role functions in serving the goals of the profession, and by how those goals are connected with...a key human

> good…which plays a crucial role in enabling us to live a humanly flourishing life. (Oakley and Cocking, 2001, 74)

Oakley and Cocking's examples are more typical professions such as medicine or law, which have obvious connections to key human goods (health and justice, respectively). If their account is right, then technologists working on AI systems should ask themselves whether their profession is contributing to a key human good, and which traits of character would promote that result. Reflecting on virtues important to the computing professions in general, Richard Volkman lists integrity, honesty, courage, and good judgement (Volkman, 2013). And from Foot's account, we could add that technologists should consider what temptations towards vices, such as greed or apathy, may exist in their line of work, and how they might resist them.

2.4.4 Moral Rights, Contracts, and Justice

The topic of *justice* has been central to ethics since antiquity, but the concept of *moral rights* and the insight that social, political, and moral relations can be understood as a kind of implicit *contract* are early modern inventions. Moral rights are entitlements and freedoms that all persons are owed, unlike *legal rights*, which are valid only within a particular jurisdiction's legal system – though charters of legal rights frequently draw upon and include universal moral rights. The UN's Universal Declaration of Human Rights is an often-cited document defining a number of commonly acknowledged moral rights, such as the right to liberty, to freedom of thought, to freedom of expression, to privacy, to work, to own private property, to health, and to education (United Nations, 1948).

Two philosophical defences of moral rights are worth outlining in this context. The first is John Locke's contractarianism (Locke, 1690). On his view, even in a *state of nature*, that is, a hypothetical situation where no social, moral, or political institutions exist, human beings would have the right to defend their lives, the right to own things that they improve with their own labour, and the right to punish wrongdoers. Each of these rights stem from Locke's fundamental insight that people *own themselves.* Locke argues that society emerges from the state of nature when people begin to make agreements in order to manage transfer of property and to protect one another from thieves and other bad actors. These agreements form the *social contract*, which specifies the rights and obligations of each member of the society.

The second philosophical defence of moral rights worth considering is John Rawls's contractualism (Rawls, 1971). Rawls defends an account a just society, starting from the insight that in order to make rules that would be fair to all, the rule-makers would have to be ignorant of the positions they would eventually occupy in that society. Rawls argues that from this position behind what he calls the *veil of ignorance*, the rules we would agree to would agree to at least the following principles. First, everyone would have an equal right to the greatest possible extent of freedom that does not begin to infringe on the freedoms of others. Second, positions in society should be open to all who qualify, without arbitrary discrimination. Third, any inequalities in the distribution of resources can only be justified if these inequalities actually make those who are the worst off in society better off than they would be if resources were shared equally. These principles then inform the rule-makers' decisions as to which rights and other social rules they will adopt.

In addition to respecting the legal rights of stakeholders, technologists must also consider whether their innovations respect stakeholders' moral rights. This additional consideration is important for several reasons. First, not all jurisdictions respect all of their citizens' moral rights. For example, 71 jurisdictions still criminalize consensual same-sex sexual activity (Human Dignity Trust, 2021). AI technology implemented in these jurisdictions could therefore be used to persecute LGBTQ+ people. Second, not all moral rights are enforced evenly. For example, Locke's argument about having a rights to own things that we improve with our labour is often invoked as a defence of intellectual property rights that is independent of any particular legal system or enforcement mechanism. If this argument works – and some argue that it does not (Nozick, 1974) – then we have moral reasons not to break copyright or other intellectual property rights in the creation or implementation of AI systems, even if enforcement is unlikely. Finally, as Rawls reminds us, no system of rights is just if some members of the society are denied equality of opportunity. Technologists must bear this in mind particularly when working with vulnerable populations who have been subject to historical injustice, and consider how their AI systems might contribute to or instantiate the kinds of unfair bias that Noble, Benjamin, and D'Ignazio and Klein discuss.

2.4.5 Care Ethics

While the contributions of feminist philosophy have been significant in all areas of normative ethics, one particularly evident result has been the establishment of the *ethics of care* as a distinctive approach to ethical theory. While initially growing out of observations that some girls and women differ in their moral reasoning from boys and men and the association of caring with femininity (Gilligan, 1982), care ethics need not posit a (problematic) essential link between care and the feminine.

On Nel Noddings's account (Noddings, 1984), ethics should be based in relationships between people rather than abstract principles. We are subject to ethical demands when we discover that someone has a need that we can fulfill. But these demands are more forceful when we have a stronger or closer relationship with the person in need; so, unlike utilitarianism or Kantianism, which emphasize impartiality as a component of ethical reasoning, Noddings's ethics of care recognizes that some of our interpersonal relationships give rise to greater moral demands than others. More recent developments in care ethics have also emphasized that an aspect of caring is taking into account how others will be affected by our actions, and that those who are particularly vulnerable to harm as a result deserve special attention in our moral deliberations (MacKay, 2020).

Care ethics is especially important to consider when AI systems are developed and implemented in contexts where some stakeholder groups are made vulnerable to the systems' behaviour. For example, clients of public services such as law enforcement, health care, education, and others have considerably less power than agents of those services. Decisions to deny access to these services, or to exercise the authority vested in these government agencies to impose penalties, can have drastic effects on the wellbeing of these clients. Because we are vulnerable to their decisions (not) to care, we hold police officers, physicians, and teachers to a higher ethical standard. When AI systems are developed for or implemented in these contexts, the same higher standard applies to those systems. By extension, this higher standard applies to the designers of those systems, since in many cases human agents will adopt an AI system as a way of delegating difficult or tedious tasks, removing a degree of human oversight in decision-making. Designers need to work closely with both stakeholders in the public service and with members of the public who will be clients of those services to determine how the AI system might make clients vulnerable, and how to mitigate or eliminate potential harms.

2.4.6 How to Use a Theory

In professional applied ethics practice, explicit use of these theories is less frequent, but similar values to those which are captured by these theories appear throughout the tools, principles, and heuristics deployed in those contexts. Understanding the basics of these theories is thus a useful basis from which to approach particular ethics tools. And, as mentioned above, an understanding of ethical theory is useful when the available tools lead to conflict, or when no apt tools exist for the problem at hand.

Plug-and-Play

When a situation arises when the more general principles and justifications of a philosophical theory are needed, there are several ways to deploy the theory. Many applied ethics courses and papers take the approach of "plugging in" the details from real life or realistic cases to see what moral evaluations result, as if the theories were functions one could call and the facts of the case were arguments passed to those functions. This is the most straightforward approach, and can be highly productive and yield useful insights. We can call this the *plug-and-play approach* to using an ethical theory.

Something the plug-and-play approach often loses sight of is how philosophical theories are not timeless and readymade. Rather, as Elizabeth Anderson argues, drawing on John Dewey, our philosophical theories should – and do – adjust in response to the results of their application (Anderson, 2014; Dewey and Tufts, 1985) Those applying philosophical theories in practice would be well advised to treat them not as scripture that is not to be questioned, only obeyed, but rather as guidance that may be questioned or even changed. Practical ethical work is an essential component of how normative ethical theories are tested and adjusted.

Foundations vs. Lenses

Writing from the perspective of biomedical ethics, Susan Sherwin makes some instructive suggestions for how to deploy philosophical theories in applied ethics (Sherwin, 1999). Against what she calls the *foundations approach*, Sherwin advocates for an approach that treats the theories as *lenses*. The default position in academic philosophy is to treat normative ethical theories as the "firm foundations" upon which applications are built. The theory itself is not open to question, and considerations from different theories cannot be imported unless they can somehow be made compatible with the chosen foundation. This approach is problematic in an applied context because bioethicists – and AI ethicists – must consider multiple different values and value systems that are held by the different stakeholders, and they do not have the luxury of investing time to solve fundamental conflicts between the different theories.

Sherwin's suggestion is to reject the foundations approach, and to think of ethical theories instead as lenses. When examining an object of interest, one might employ a number of different lenses depending on the features of the object to which one wishes to attend. The same tree, for example, could be studied using binoculars, a hand lens, or a microscope, each revealing different information. Similarly, we can examine the same ethical problem using the perspectives of different ethical theories, each revealing different information suggesting how we should evaluate the situation.

For example, consider the ethics of creating a deepfake video without the permission of the person being imitated. The utilitarian would look only at the consequences of creating the video, and sum up the happiness and harm produced – including, perhaps, the amusement of those who watch the video and the embarrassment of the person imitated. Since the pain and humiliation caused to the person imitated only counts for so much, if enough people find amusement in the deepfake, the utilitarian will always conclude that it is right to create it. The Kantian may ask, instead, if the subject of the deepfake has been used merely as a means to the ends of the deepfaker. Since they did not give their permission to be deepfaked, it seems clear that this is the case. We may then decide that in this case, the violation of the imitated person's dignity is more important, and cannot be justified by any degree of happiness so produced. Or, we may decide that because the subject of the deepfake is, say, a powerful dictator who deserves to be taken down a peg, the utilitarian analysis is more appropriate to this case.

While lacking the rigour of an academic philosophical account of the problem, the lenses approach enables applied ethicists to more quickly examine the various ethically relevant aspects of a problem from a variety of perspectives. It thus provides a more practical way of making use of ethical theories in conjunction with – or as a replacement for – the tools detailed in the following chapters.

Chapter 3

From the Theoretical to the Practical

Working with private consulting and auditing startups, the Oxford Internet Institute has been a pathfinder when it comes to putting philosophical and sociological theory into practice in AI ethics. Morley et al. (2020, 2143) identify six classes of ethical issues of particular concern with regard to machine learning and AI:

- **Inconclusive Evidence**: Even AI systems that are highly accurate can produce results that are incorrect or misleading. Human interpretation of those results before putting them into practice is essential.

- **Inscrutable Evidence**: Also called the "black-box" problem, many AI systems are difficult to interpret or understand, making monitoring and evaluation difficult.

- **Misguided Evidence**: Datasets used to train and test machine learning models can be biased in various ways. Failing to vet datasets for such bias increases the risk that the resulting models will produce unfair outcomes.

- **Unfair Outcomes**: When an AI system has a disproportionate effect on one or more groups, particularly when those groups are already subject to structural oppression or social injustice, the system is discriminatory. Such outcomes often happen when the dataset is biased (see above), but emergent biases can also occur when the context of application changes (cf. Friedman and Nissenbaum, 1996).

- **Transformative Effects**: AI systems may violate privacy or undermine autonomy when data is harvested for training, testing, or running a model.

- **Traceability**: Harms caused by an AI system can be difficult to trace to a specific agent, complicating attributions of moral responsibility and legal liability.

Each of these issues raises further topics of theoretical interest, such as the nature and value of privacy and autonomy; the varieties of bias, prejudice, and injustice; and how to attribute responsibility for the actions of an autonomous computer system. The challenge, as Morley et al. remind us, is how to move to *practical* tools and processes for addressing these issues.

A significant aspect of this endeavour is to create tools that are neither too flexible – and thus susceptible to "ethics bluewashing," where a superficial appearance of ethical practice is maintained by a tokenistic effort to enact ethics principles – or too strict – and thus susceptible to "ethics shopping," where ethical rules and principles are cherry-picked to make one's products or practices *appear* ethical, instead of a justified set of principles informing the process of creating those products or practices all the way along (Floridi, 2019). The difficulty lies in how to move from abstract theoretical principles to more concrete, actionable tools, standards, and procedures.

3.1 Principles, Codes, and Guidelines

One common way of attempting to translate abstract theories to practical guidance is the creation of *codes of ethics* – typically modelled after established professional codes such as the Hippocratic Oath (Hippocrates, 2002) or the ACM Code of Ethics and Professional Conduct (ACM, 2018) – or lists of *principles*. Examples of codes and principles in AI ethics include Floridi et al. (2018)'s AI4People framework, Floridi and the Baron Clement-Jones (2019)'s five principles for AI ethics frameworks, the EU's Ethics Guidelines for Trustworthy AI (European Commission, 2019), the IEEE's general principles of ethically aligned design (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019), the principles of the Montréal Declaration for Responsible AI Development (Abrassart et al., 2018), the Beijing Academy of Artificial Intelligence's principles (Beijing Academy of Artificial Intelligence, 2019), the OECD Council on Artificial Intelligence's recommendations (OECD, 2019), Zook et al. (2017)'s 10 rules for responsible big data research, Microsoft's AI principles (Microsoft, nd), or Google's AI principles (Google AI, nd).

The proliferation of these guidance documents over the last few years is impressive: according to Algorithm Watch, there are now over 160 such documents (Algorithm Watch, 2020). However, at this point we are arguably in a position of what we may call *guideline glut*, where there are far too many guidance documents for practitioners to assimilate. While it might be argued that having a large number of codes and sets of principles is a good thing given that many different perspectives may be represented within, in actual fact there is little fundamental variation between them. According to (Morley et al., 2020, 2146), all of these codes and sets of principles can be summarized under the following five-principle framework:

- **Beneficence**: AI should do good in the world.
- **Non-maleficence**: AI should do no harm, with particular emphasis on privacy and security.
- **Autonomy**: AI should promote greater, not lesser, autonomy for human beings.

- **Justice**: AI should promote justice and fairness for all, and not contribute to discrimination and oppression.
- **Explicability**: AI systems must be understandable to human beings, and decisions made using AI should be explainable and justifiable.

Still, it can be useful for a practitioner or firm to adopt a recognized set of principles as a commitment to ethical AI development. Any set of principles will help to distill abstract philosophical and sociological theory into more practically usable directives.

But while these codes and principles may be one step down in level of abstraction from philosophical or sociological theory, they are still relatively abstract, particularly for technologists whose past training and education did not include humanities and social sciences. As Peters (2019) observes,

> While principles are essential, many tech makers are frustrated by how little help they provide in actual practice. Principles must be sufficiently abstract to retain truth across contexts, but this abstraction also leaves them too vague to be useful for specific design decisions on their own.

Rather than explore these more specific theoretical topics or elaborate on the various codes and lists of principles, then, the remainder of this report turns to a selection of practical AI ethics tools aimed at various stakeholders, including developers, managers, executives, educators, consumers, and members of the public. Each of these tools provides a procedure to follow for addressing, mitigating, or preventing the kinds of ethical problems noted above.

3.2 AI Ethics Tools Typology

Morley et al. (2020) organize over 100 AI ethics tools according to a typology based on where in the product lifecycle the tool is meant to intervene: business and use-case development, design, training, building, testing, deployment, or monitoring (Morley et al., 2019). By contrast, the present report organizes a smaller selection of tools according to their general *methodology*, namely, whether the tool comprises a process to follow during or throughout the lifecycle of an AI product, a design process to follow, a checklist to review at different steps of the design and implementation of the product, a protocol for filling out an information slip about the product, a teaching tool for technologists or executives to learn about AI ethics, or an assessment tool for conducting internal or third-party audits. The aim is to offer a complementary survey of recent AI ethics tools of particular interest to philosophers, consultants, and auditors working in the AI ethics space.

Before moving to the discussion of these tools, however, three remarks from Morley et al. (2020) are worth noting. During their analysis of AI ethics tools, Morley et al. found three patterns: there is an overabundance of tools focused on making machine learning models explicable; the majority of tools fail to provide means of assessing the impact an AI system may have on individuals, social groups, or society; and many tools “are positioned as discourse aids, designed to facilitate and document rational decisions about trade-offs in the design process that may make [a machine learning] system more or less ethically-aligned" (Morley et al., 2020, 2157), rather than tools that can be used practically by technologists or members of the public of any level of technical or philosophical skill. These concerns about the available AI ethics tools informed the curation of tools in the following chapters. The aim, as Morley et al. write, is "to remove friction from applied ethics" (2020, 2157) as far as possible.

3.3 Ethics as Service

In a more recent paper, Morley et al. couple their earlier claims about the potential problems with using existing AI ethics tools with the observation that the proliferation of AI ethics tools has been very recent, and there is at present little empirical data to suggest that these tools are effective (Morley et al., 2021, 3). Morley et al. argue that in order to make AI ethics tools more useable in practice, more than just a lowering of abstraction is needed. In addition to being too vague and abstract, most existing AI ethics tools are diagnostic, able to identify potential issues but failing to provide practical guidance on how to resolve them, or providing technical fixes that avoid engaging with the underlying social issues at the root of the problem. Practitioners need more specific guidance on how to translate the results of ethical diagnoses into appropriate treatments.

Moreover, many tools are presented as "a 'one-off' test: something that just needs to be completed for compliance purposes (to be awarded a 'kitemark' [i.e., a safety certification badge like that awarded to consumer products by the CSA Group or BSI Group] of some description, for example) and then forgotten about' (Morley et al., 2021, 6). This perception of ethics tools undermines their entire purpose: the goal is not to introduce a "pre-flight checklist" at the end of the design process, but rather, to incorporate ethical reasoning at all stages of the design process. Morley et al. explain that ethical implications of AI and other information systems must be evaluated in three phases:

1. **Validation** (early design, pre-prototype): Is this the right algorithmic system for the use case?

2. **Verification** (mid-design, at and after prototyping): Is the algorithmic system being developed appropriately?

3. **Evaluation** (post-design, during testing, iteration, and implementation): Is the algorithmic system operating appropriately? Should it be revised? Can it be improved?

Some of the tools discussed below intervene at specific points in this process, or embed ethical reflection and evaluation throughout.

Morley et al. go on to say that some compromises have to be made in order to reach an appropriate level of abstraction and to avoid the pitfalls they have identified with many extant ethics tools. Firstly, an ethics tool has to find the appropriate middle ground between being too strict and being too flexible. In so doing, the tool will have to be sensitive to the context in which it is to be used: more flexible tools may be needed in some settings, and more strict tools in others. To do this properly, they argue that a firm must commit to the following procedures (Morley et al., 2021, 9):

- Adopt a set of ethical principles across the firm. These principles must be agreed to via a process where representatives from all levels of the organization, as well as external stakeholders, policy experts, and environmental stewards. Additionally, these policies should be reviewed periodically, ideally about once per year.

- Create a set of procedures that employ tools which translate the agreed-to principles into technical standards for each relevant unit of the firm. These procedures should be followed in the same way each time they are used.

- Put in place an oversight mechanism to monitor the use of these procedures in all stages of the design process.

One might seek to address some of the shortcomings of AI ethics tools by outsourcing to a third-party auditor. Morley et al. (2021) note that this is an important element of putting AI ethics into practice. However – and this is the second compromise – they also argue that most existing audit frameworks are insufficient. They tend to focus on specific aspects of the system (e.g., fairness or explainability), rather than comprehensive ethical evaluation of the entire system. Furthermore, audits tend to be conducted after implementation in response to harms already caused. A more responsible approach would be to perform audits proactively, so as to prevent harms from arising in the first place. Finally, audits are often limited in effectiveness or scope because of a lack of legislative backing, allowing auditees to prevent the auditing or publication of certain aspects of their products or processes, citing intellectual property and user privacy rights. Some level of compromise may be necessary between internal audit procedures and external, third-party audits.

An alternative approach proposed by Morley et al. (2021) is based on the "Platform as Service" approach to cloud computing technology. The model of *ethics as a service* takes on needs to reduce abstraction, be sensitive to context, integrate ethics into every stage of the design process, and audit the entire algorithmic system in collaboration with the technology firm. The core activities in an ethics as a service approach are as follows (Morley et al., 2021, 13):

- **Activities performed by the client firm**
  - Contextually interpret ethical principles from the ethics service provider's ethical code for the present project and situation.
  - Choose ethical tools from a pre-approved list compiled by the ethics service provider.
  - Conduct ethical review at all stages of the design cycle.

- **Activities performed by the ethics service provider**
  - Development and review of the ethical code.
  - Develop processes to ensure compliance with the code in client firms.

- Evaluate ethical tools and compile a list for use by the client firm.
- Audit AI systems proactively and reactively.

The ethics service provider should consist of a multi-disciplinary advisory board trained in a variety of specializations and from multiple demographics. In addition, practitioners at the client firm would need training in how to interpret and make use of the principles, tools, processes, and review procedures made available to them by the service provider. Morley et al. (2021) claim that a UK firm called Digital Catapult (nd) is currently trialing the ethics as a service approach as a first step to determining its feasibility.

Chapter 4

Ethical Design and Review Processes

4.1 Responsible Rendition of the Double Diamond design framework (R2D2)

The *double diamond* design framework (DD) is a widely used visualization of the design process, produced in 2004 by the Design Council (Ball, 2019, see Figure 3.1). DD divides the design process into two phases, each of which has a divergent stage and a convergent stage:

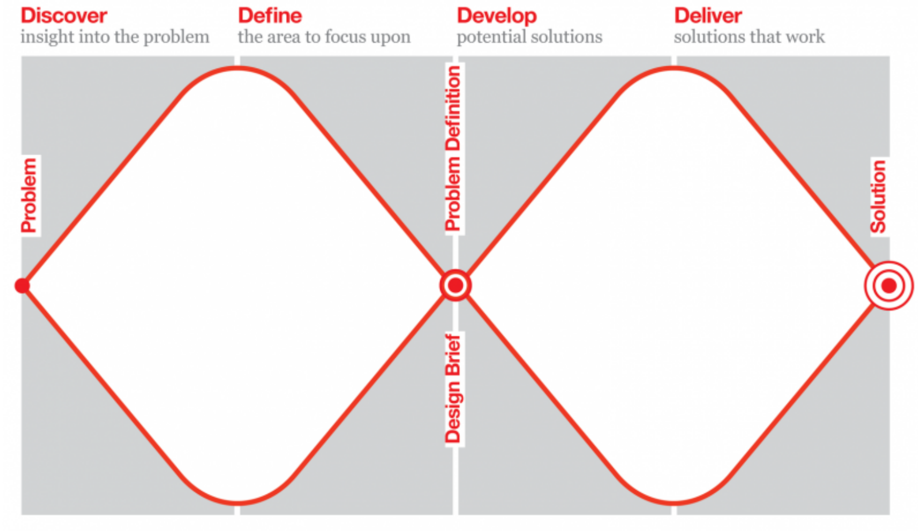
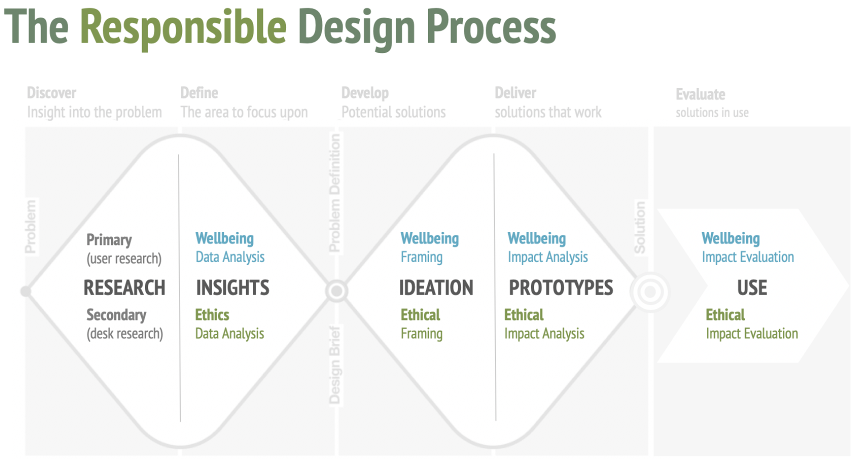


Figure 4.1: The Double Diamond framework.

- **Problem Definition**
  - *Discover*: Research potential design problems and end user needs.
  - *Define*: Specify the design problem to be addressed.
- **Solution Development**
  - *Develop*: Outline, prototype, and test various potential solutions to the design problem.
  - *Deliver*: Produce the most promising of the solutions and prepare it for implementation.

The Design Council has since expanded DD into the *Framework for Innovation* (Design Council, 2015), which includes additional elements such as design principles and a recognition that the process of design may iterate through different stages of the framework more than once.

Recognizing that it is insufficient to ask technologists to consider the ethical impacts of their innovations only after the design cycle is completed, Peters et al. (2020) introduce a modified version of DD that they call the *Responsible Rendition of the Double Diamond*, or "R2D2" (see Figure 3.2). A key advantage of R2D2 is that it makes use of a framework many technologists will already be familiar with, while encouraging them to take ethical thinking seriously as a necessary aspect of the design process.



The Responsible Design Process, v1.0 by Dorian Peters is licensed under a Creative Commons Attribution 4.0 International License.

Figure 4.2: The Responsible Rendition of the Double Diamond (R2D2) (Calvo and Peters, nd)

R2D2 introduces ethical considerations to the four stages of DD, and adds a fifth stage, *Evaluate*, after the product is released, where the product's performance is considered while it is in actual use, and unforeseen ethical issues can be addressed. In each stage, R2D2 directs designers to consider two ethical dimensions of their ongoing work: wellbeing and other ethical considerations. "Wellbeing" here corresponds roughly to the utilitarian conception of moral good, and is expected to draw from psychological research on wellbeing; however, for a richer and broader conception of wellbeing we could also substitute the Aristotelian notion of human flourishing. Among the "other" ethical considerations are: "fairness, data governance, ecosystem wellbeing, or democratic participation" (Peters et al., 2020, 37), as well as the five principles highlighted by Morley et al. (2020) – beneficence, non-maleficence, respect for autonomy, justice, and explicability.

While R2D2 has the advantage of extending an already widely known design framework, it suffers from the limitation that technologists engaged in an R2D2 design process must themselves be familiar with AI ethics principles and their theoretical background, or at least have access to consultants or internal experts who are sufficiently familiar with *both* the design process *and* the relevant ethical literature. The main reason for this limitation is that R2D2, in essence, amounts to adding an ethics layer on top of the existing DD framework, rather than describing a new design framework with ethics baked into the process.

4.2 AI Blindspot

The AI Blindspot tool is a series of flashcards – which may be digital or printed – designed to help AI developers consider ethical issues in their design, with a particular focus on data justice and algorithmic bias. (Calderon et al., 2019). The cards are grouped into phases of the design cycle, using a simplified version of the process described by Ball (2019) and Peters et al. (2020). The obverse of each card has an icon and a brief explanation of an AI ethics principle linked to fairness and equity; the reverse features a list of several specific points to consider, a brief description of a case study, a list of stakeholder and expert groups to consider engaging, and a QR code that links to the AI Blindspot website for further information.

The cards highlight several ethical issues at each stage of the simplified design process:

- **Planning**
  - *Purpose*: AI should have positive outcomes for human wellbeing.
  - *Representative Data*: Datasets should be representative of the human population.
  - *Abusability*: Security vulnerabilities and malicious uses of the AI should be considered and mitigated.
  - *Privacy*: User data processed by the AI or in training datasets should not be obtained without informed consent from the data subjects, and should be protected from unauthorized access.
- **Building**

  - *Discrimination by Proxy*: When there are correlations in data caused by social injustice or structural oppression, AI can reproduce discriminatory results even when demographic information is not explicit. These proxy discriminations should be anticipated, tested for, and mitigated.
  - *Explainability*: Recommendations produced by AI systems should be interpretable by a human being, and the reasons for their decisions should be explainable and justifiable.
  - *Optimization Criteria*: Metrics for success must be balanced against the potential harms that the AI system may cause, with particular attention to marginalized and vulnerable populations.

- **Deploying**
  - *Generalization Error*: When the context in which the AI system was created and tested is different from the context in which the AI system is deployed, unforeseen errors may arise (cf. Friedman and Nissenbaum (1996)'s notion of "emergent bias").
  - *Right to Contest*: Stakeholders should be allowed and enabled to contest the results of AI systems that affect them, particularly if they allege biased or discriminatory outcomes.

The cards also include a "joker" card that is left blank for practitioners to fill in with their own cases.

The cards are presented as a tool to help guide technologists through different kinds of ethical considerations that arise at different steps of the design cycle. For this reason, they can be thought of as a tool that supplements or implements the R2D2 framework, supplying additional practical details that are not included in Peters et al. (2020). At the same time, their flashcard design makes them perhaps better suited to the role of teaching tools (see Chapter 6), to be used in the context of a design assignment in a course or as part of a hypothetical design exercise in a workshop. Their easy to understand content and attractive presentation also makes them suitable for teaching members of the public, policymakers, or other non-specialist stakeholders. In an actual design environment, the cards are perhaps better conceived of as a document providing guidelines to follow. Another potential way to make use of the cards would be to print them at larger scale as posters to remind members of the design team of the steps to be followed.

4.3 Data Ethics Canvas

Developed by as part of a white paper by the Open Data Institute (ODI), a non-profit dedicated to improving the "data ecosystem" through consulting, outreach, and advocacy, the Data Ethics Canvas is a worksheet for teams working with datasets in the development of AI and other computer and information systems (Broad et al., 2017). The canvas asks teams to fill in a number of questions on the following aspects of their data sources:

- Data sources
- Limitations in sources (e.g., bias, gaps)
- Who the data will be shared with and under what conditions
- Relevant policies and regulations
- Who has what rights over the data
- Ethical codes and frameworks that apply
- Reasons for using the data (e.g., primary purpose, business model, socio-economic effects of success such as job loss, who benefits if anyone)
- Communicating your purpose (i.e., for informed consent from data subjects, alerting stakeholders)
- Positive effects on people of your project, how these are to be measured and increased
- Negative effects on people of your project, how these are to be measured and mitigated
- Steps to be taken to minimize harm
- Engaging with people (e.g., allowing data subjects to challenge your work, how affected people can request changes, whether you will implement a right to erasure)
- Communicating risks and issues to stakeholders
- How ethics reviews will be conducted and how they will affect the ongoing development of the product
- What actions you will take in response to the considerations highlighted by this canvas

The data ethics canvas splits the difference between the information slips approach (see §5) and assessment tools like the Harms Modeling approach (see §7.3). Like the former, the canvas provides a standard framework for answering a variety of ethically relevant questions about the data that can be shared with other stakeholders. Like the latter, the canvas is more specific about the different kinds of harms that can arise, though less comprehensive. The canvas is available as a PDF, an online Google Doc, and hard copies, and the ODI offers a half-day course in how to use the canvas effectively (Open Data Institute, 2019).

Chapter 5

AI Ethics Checklists and Frameworks

One common type of AI ethics tool is the *checklist*. These tools are formatted as a series of questions or steps for developers and consumers to walk through as they create or select an AI system. One benefit of the checklist format is that it distills sometimes complex ethical considerations into easy to understand and consistent frameworks, making them portable to various organizations and saving deliberation time for managers and executives. At the same time, they can be customized to an organization's context and priorities. A potential disadvantage of checklists is that they risk becoming formulaic, failing to provoke deeper reflection and investigation of ethical issues that are sometimes needed. Careless practitioners might also attempt to import a checklist to a context that it was not designed for, leading to some ethical or technical issues being neglected.

5.1 Framework for Ethical AI in Education

Developed by a short-term research institute at the University of Buckingham, *The Ethical Framework for AI in Education* (Institute for Ethical AI in Education, 2020b) defines a checklist "aimed at those making procurement and application decisions relevant to AI in education," (Institute for Ethical AI in Education, 2020b, 3) with the purpose of helping these decision-makers choose AI services that enable teachers and learners to benefit from the implementation of AI systems while safeguarding against the potential risks. The report draws on interviews with policymakers, academics, philosophers, industry experts, educators in a variety of educational contexts, and students. The report's annex describes in more detail the justifications for each aspect of the framework (Institute for Ethical AI in Education, 2020a). Several interim reports produced during the lifespan of the Institute are also available on their website (Institute for Ethical AI in Education, 2020c).

The framework is divided into several objectives, each of which has multiple criteria to fulfil. Each criterion itself has a short checklist of questions to consider, with a specification of which stage in the process of implementing an AI service in education those questions should be considered: pre-procurement, procurement, implementation, monitoring and evaluation, or all stages. The objectives described in the report are:

- **Achieving Educational Goals**: AI should support independently justified educational objectives.
- **Forms of Assessment**: AI should be used to evaluate a wide range of learner capabilities.
- **Administration and Workload**: AI should make organizational processes more, not less, efficient.
- **Equity**: AI should promote equity and social justice for various groups, and not contribute to discrimination or structural oppression.
- **Autonomy**: AI should increase learners' control over their education and development.
- **Privacy**: Implemented AI systems should respect learner privacy as much as possible while still collecting sufficient data to serve their intended purposes.
- **Transparency and Accountability**: Human beings should have significant oversight of AI systems and ultimate discretion over learning processes.
- **Informed Participation**: Learners, teachers, assistants, and other practitioners should understand AI systems that are implemented, as well as the consequences of such implementation.
- **Ethical Design**: Designers and developers of AI systems must understand their impact in the educational context and take ethical and justice issues into consideration during the design and development process.

A limitation of this framework – admitted by the authors of the final report (Institute for Ethical AI in Education, 2020b, 3) – is that it relies on AI designers and developers to follow their own ethical principles and procedures. The framework is aimed at stakeholders who must be able to trust that vendors and developers have done their own due diligence.

5.2 Machine Learning Reproducibility Checklist

Developed by a group of researchers at multiple academic institutions and big tech firms, the Machine Learning Reproducibility Checklist is a quick-use tool intended to promote more reproducible research in machine learning (Pineau et al., 2020). As philosophers of science have argued since the mid-20th century, reproducibility is an essential element of modern scientific research because it is a way of ensuring that results can be verified independently by other scientists, thereby ensuring that the scientific consensus is robust (Fidler and Wilcox, 2018). Other scientific fields, particularly psychology, have recently been wracked a "crisis" of reproducibility, where studies that had been popularized in mass media have been shown time and again not to be replicable by independent researchers following exactly the same procedures, throwing all work developing upon those famous studies into doubt.

A similar worry plagues the machine learning research community, where sometimes independent researchers using exactly the same code are unable to reproduce similar results to those initially reported. This checklist is intended to facilitate the work of reseachers in making studies easier to test for reproducibility of results, with the aim of improving the quality of all scientific research in this field. The checklist itself is brief, comprising a single page asking researchers to verify that their study meets certain standards, such as complete descriptions of the algorithms used, a clear and complete description of the theoretical claims, complete information about the datasets used and how they were processed, complete copies of the code (for training, evaluation, and post-training), and complete specifications for results reported.

This tool is primarily useful for machine learning researchers, academic or otherwise. The ethical issue it addresses is primarily one of scientific integrity: researchers must be clear about their own work as contributors to their epistemic community, so that the entire community can make progress on advancing understanding. The checklist provides one way to help researchers do their due diligence when writing up their research to share with others.

5.3 Digital Catapult's Ethics Framework

The ethics framework developed by consulting firm Digital Catapult, in collaboration with the Oxford Internet Institute, is intended to support firms developing AI and machine learning systems in designing ethical products and services, and is the ethical code that forms a part of their ethics as a service business model. The framework consists of seven principles, each of which has multiple questions to prompt reflection by practitioners on how to meet the requirements of each principle. The framework also provides links to frequently asked questions and a position paper on why ethics is important in the AI industry (Digital Catapult, 2021). Because the framework is mainly a list of questions, I have categorized it as a checklist – for a questionnaire is, after a fashion, a sort of more demanding checklist.

Digital Catapult's framework contains the following principles, each of which comes with many questions to work through:

1. **Clear benefits**. Most AI ethics is concerned with risks and harms, but firms also need to be clear about the potential benefits of their AI products. AI is not worth pushing to market for its own sake; some possible AI developments may *only* cause harm, without appreciable benefit. If there are no benefits, then there is no reason to develop the product.

2. **Know and manage the risks**. Risks of harm to individuals, communities, and the environment should be considered, from a variety of perspectives. Risk assessments should be done considering not just the ideal use case, but also

actual conditions in which the system will be implemented. The possibility of the system being co-opted for malicious purposes must also be considered.

3. **Use data responsibility**. Firms should comply with the relevant data protection and privacy legislation (e.g., with the GDPR or PIPEDA). However, ethical issues connected to data go further. Systems should be tested on situations that go significantly beyond scenarios in the training data. Datasets not carefully curated can be biased, and thereby produce models that are biased. Finally, data sources should be fairly compensated, whether they are microtask workers (e.g. Amazon Mechanical Turk workers) or society as a whole (e.g. public datasets).

4. **Be worthy of trust**. Elements of trustworthiness go beyond community management and responsiveness to client queries. To be trustworthy, AI systems must be reliable and consistent across use cases. They must also produce results that are explainable and decisions that are justifiable. Finally, firms must be accountable for errors and harms produced by their products.

5. **Diversity, Equality, and Inclusion**. Firms should consider the impact of their technologies on society as a whole, paying particular attention to their effects on populations that are already marginalized or oppressed. Firms should also prioritize diversity and inclusion in their hiring and stakeholder consultation practices.

6. **Transparent communication**. Firms must be clear in their communications about their products. Insofar as is possible for them without giving up their competitive edge, they should be transparent about the data and algorithms they use, their models' performance on various benchmarks, and their efforts to mitigate risk.

7. **Ethical business model**. Ethics must be fully integrated into the firm's business model, and not simply an afterthought for compliance or in reaction to mistakes. Integrity, fairness, and social responsibility must be central values.

Chapter 6

AI Information Slips

A recent development in AI ethics tools is what I refer to here as *information slips*. The basic approach is to include short documentation about the training data used to create an AI model, the context(s) in which the data were collected, the characteristics of the model, any foreseen dangers of using the model in contexts other than the original intent, and other relevant information. Interestingly, using these approaches is one of the key recommendations made in the conference paper (Bender et al., 2021) that infamously precipitated the ouster of Gebru and Mitchell from Google's AI ethics team – an incident which raises additional issues in professional ethics that go beyond the scope of this report.

Information slips are mainly intended to overcome three shortcomings of AI model and dataset development and presentation. First, the format most widely used for datasets, plain text comma-separated values or `.csv`, lacks support for the inclusion of contextual metadata, necessitating the creation of additional documentation to include information relevant to analysing the dataset. However, no widely endorsed standard yet exists for such information. Second, datasets and machine learning models are often difficult to scrutinize directly for quality and for ethical risks, and, third, most technology specialists who would be best placed to perform such analyses lack suitable procedures to follow. As a result, many organizations simply have no processes in place for conducting ethical review of datasets and models prior to implementation.

6.1 Datasheets for Datasets

Datasheets for datasets are a proposed standard for documenting information about datasets used in training and testing machine learning models (Gebru et al., 2018). The basic idea comes from the electronics industry, where every component in a device has a datasheet explaining its standard operating characteristics, recommended use, and test results. Gebru et al. recommend that datasheets for datasets contain the following information:

- **Motivation for dataset creation**, as well as information on tasks the dataset has been used for previously, and any cases for which the developers think it ought *not* to be used.

- **Dataset composition**, with particular attention to the nature of the instances (i.e., the individual items in the dataset), how many instances of each type there are, and any external dependencies.

- **Data collection process**, including information on who participated in the process, when, in what region, demongraphic information, and any known errors, sources of noise, and missing data.

- **Data preprocessing**, including information on how the data were cleaned and why.

- **Dataset distribution**, including information on where the dataset has been made available, how widely, and under what intellectual property licence or other legal restrictions.

- **Dataset maintenance** information, such as who is responsible for maintaining the data, how often it will be updated, or if the dataset is no longer being actively maintained.

- **Legal and ethical considerations** that were taken into account when creating the dataset, such as obtaining informed consent from human data subjects, legal permissions from data controllers, ethics review board approvals for the data collection, privacy legislation compliance (e.g. GDPR); as well as information on anticipated risks of harm posed by using the data, with particular attention to vulnerable or oppressed groups.

To aid in filling in each of these sections, Gebru et al. (2018, 7–8) provide a list of sample questions that model developers should ask as a starting point. They also give several examples of completed datasheets.

6.2 Model Cards

Model cards are a proposed standard format for documenting the performance characteristics of machine learning models (Mitchell et al., 2019). Whereas the information slips listed above concentrate on reporting aspects of the datasets that are fed into machine learning models, model cards provide information on the machine learning model itself. The aim is to help various stakeholders – including general users, model developers, policy makers, researchers, and people impacted by models – to understand how the model works at a high level, to compare the model to others with similar purposes, and to determine how harmful consequences of the model may have arisen and where remedies can be recommended.

Model cards are one or two pages long, and summarize the following information:

- **Basic information** about the model, such as the people responsible for creating or maintaining the model, version information, features and algorithms, links to research, and contacts.

- **Intended use** of the model, with information on both the intended purpose of the model and the intended users, as well as any uses that are explicitly outside the intended scope.

- **Factors** affecting the model's performance, such as differences in performance when applied to different social groups, differences in instruments used to gather data to be fed into the model as inputs, and differences in the environment in which the model is deployed.

- **Metrics** used for performance testing, that is, quantitative measures of performance on relevant tests, along with justification for why these metrics were used.

- **Evaluation data** information, that is, a description of the datasets used to test the model, how the data were collected and from whom, how the data were processed prior to being fed into the model, and a justification for why these data were used. Mitchell et al. recommend that this section point to a datasheet for the datasets referenced (see §5.1).

- **Training data** information, containing at least the same information as the description of evaluation data.

- Results of **quantitative analyses** performed on the model according to the metrics defined above, and categorized by the factors specified above.

- **Ethical considerations** that informed the development of the model, such as social groups that may be vulnerable to harm if the model is deployed in a particular context, whether the model requires sensitive personal data to process, whether the model is intended to inform decision making that could have a major impact on human well-being, strategies used in model development to mitigate the risk of harm, remaining risks of harm if the model is deployed, and particular use cases that are expected to be high-risk.

Mitchell et al. (2019) provide two examples of model cards to illustrate this tool's use. As noted above, they are not intended as a standalone solution, but rather, they are meant to be a stakeholder-facing tool for communicating information about machine learning models in conjunction with other types of information slips. Moreover, their use depends upon an internal process for conducting performance and ethics reviews of the model and the datasets used for training and evaluation purposes.

6.3 Data Statements

Bender and Friedman (2018) propose another type of information slip they call *data statements*. The aim is "to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software" (*id.* 587). Based on documents produced in psychological and health care research to disclose information on populations studied, data statements provide the following information:

- The **curation rationale** used to collect the data, including the original goals that led to selection decisions.
- The **language variety** (or varieties) represented within the dataset. This is especially important to natural language processing datasets, because regional and dialectical variations in a language may lead to unexpected results.
- The **speaker demographics** represented in the dataset. Even within a language variety, there may be relevant differences in vocabulary, grammar, prosody, and other linguistic features depending on various demographic factors. Bender and Friedman also recommend including quantitative information on how many speakers are represented in the demographics listed.
- The **annotator demographics**, that is, information on the person or group of people who compiled and annotated the dataset. This is important to include because the social location of the annotator will affect their interpretations of the data they are annotating.
- The **speech situation**, that is, the context in which the the linguistic items were created.
- When working with written texts, other **text characteristics**, such as genre or topic, that may influence vocabulary, structure, or other features.
- When working with audio recordings, the **recording quality** should be described, as poor quality recordings may lead to unexpected results.
- **Other** relevant information.
- When a dataset is made at least in part by combining other datasets, a **provenance appendix** should be included, with information on the different sources.

As several of the items above suggest, the data statement model was developed with natural language processing models in mind. However, the basic framework is adaptable to datasets produced for other types of machine learning models. Demographic information relevant for linguistic applications may, for example, be replaced with more relevant information on the social or cultural context of the data, depending on the contents of the dataset and its intended purposes.

The above list describes what Bender and Friedman call a *long form* data statement. They recommend that this format be used in academic publications presenting new datasets. They also suggest that publications drawing on datasets for training, testing, or tuning a system should contain *short form* data statements of 60–100 words for each dataset used. Short form data statements summarize the main points in and provide a permanent link to the corresponding long form data statements.

6.4 Dataset Nutrition Labels

Dataset nutrition labels are a format for providing at-a-glance information about a dataset for training or testing machine learning models (Chmielinski et al., 2020; Holland et al., 2018). They are similar to model cards in their brevity, and to datasheets in that they present information on datasets rather than on models. The approach is inspired by the nutrition facts labels that have been mandatory on food products sold in the USA and elsewhere since the 1990s, and a more recent proposal to develop a similar label standard for digital privacy (Kelley et al., 2009). Similar to how nutrition facts list edible ingredients and provide an indication to consumers of how "healthy" a particular foodstuff may be, dataset nutrition labels "[highlight] the 'ingredients' of a dataset to help shed light on how (or whether) the dataset is healthy for a particular algorithmic use case" (Chmielinski et al., 2020, 2).

Dataset nutrition labels are generated using various quantitative and qualitative measures and statistical models, but are presented in a standard format that uses graphic design elements similar to familiar food nutrition facts labels. The aim is to analyse datasets for quality and for ethical issues before they are used for training models, so that potential risks and harms can be anticipated and mitigated or prevented *before* the models are trained and deployed. One advantage of the dataset nutrition label standard is that it incorporates techniques from data storytelling and graphic design to present information visually for ease of reading.

Part of the motivation for the dataset nutrition labels standard was an anonymous survey of AI technologists. 15 out of 34 respondents (about 44%) indicated that their organizations had no best practices in place to review datasets prior to model training, and 34 out of 58 respondents (about 59%) indicated that they had taught themselves how to analyse datasets from a variety of academic, professional, and informal sources (Holland et al., 2018, 3). The authors' hope is to create a standard that can be used more efficiently and consistently across different organizations.

The first version of the dataset nutrition label standard is a diagnostic framework built from multiple modules. Each module represents a standalone, partial analysis of an aspect of the dataset. The intent is to enable professionals to choose which modules are most appropriate to their dataset (Holland et al., 2018). The second iteration of the label standard introduces an interactive web-based graphical user interface for ease of navigation through and interaction with the different modules and use case diagnostics (Chmielinski et al., 2020).

The modules included are:

- **Metadata**: Basic technical information on the dataset and its components, licensing, and a description of the dataset and its purpose. This is the only required module.
- **Provenance**: Information on where the data came from and whether it builds on previously presented datasets.
- **Variables**: A description of each column in the dataset.
- **Stastics**: Basic statistical information on each variable (means, medians, minima and maxima, etc.)

- **Pair Plots**: Distributions and correlations between pairs of variables.
- **Probabilistic Model**: Data generated through more complex probabilistic calculations on the dataset.
- **Ground Truth Correlations**: Correlations between a chosen variable and "ground truths" such as census data.

For the qualitative modules, the current standard of the label includes a series of questions for the dataset curator. For the quantitative modules, the label calculates correlations between variables and other statistical and probabilistic measures, presenting them as histograms and heat maps. The current standard also features "badges" to highlight specific features or items of interest, such as whether the dataset is about human beings, whether the dataset has been subject to ethical review, funding information, licensing information, and how frequently the dataset is updated, if at all. Finally, the current standard produces "alerts" based on quality issues in the dataset (e.g., the presence of duplicate items, or inconsistencies in recording quality) or ethical issues (e.g., the dataset's representation of relevant demographic groups). Alerts are colour-coded depending on whether the issue can be mitigated (yellow for 'yes', orange for 'maybe', red for 'no').

The dataset nutrition labels standard is currently maintained by a team at the MIT Media Lab and Harvard's Berkman Klein Center (The Data Nutrition Project, 2021). This site also presents several example labels using the current standard.

6.5 FactSheets for AI Services

Developed by a team at IBM Research, *FactSheets for AI Services* are intended to improve the trust that consumers and developers have in AI models. Also called a Supplier's Declaration of Conformity (SDoC), FactSheets draw upon the other information slip approaches with the aim of standardizing this type of documentation (Arnold et al., 2019; Mojsilovic, 2018; IBM Research, nda; Richards et al., 2020).

The FactSheet approach is based on IBM's "Pillars of Trusted AI" (IBM Research, ndb):

- **Fairness**: Training data and models should be free of bias.
- **Robustness**: AI systems should be secure.
- **Explainability**: Decisions and suggestions provided by AI systems should be understandable by users and developers.
- **Lineage**: Details of the development, deployment, and maintenance of AI systems should be available for audit.
- **Transparency & Accountability**: Developers must be able to measure and explain the system's performance on each of the above, and be honest about its strengths and limitations.

FactSheets are intended to improve transparency and accountability in particular, and to assist organizations with AI governance. FactSheets come in several different standard formats to enable presentation in different contexts: a longform full format, a shortform tabular format, and a slide deck presentation format. IBM has also provided a Slack community, glossary, FAQ, and various articles and videos on the FactSheets approach (IBM Research, nda).

When filling out a FactSheet, IBM researchers recommend following the following iterative steps (Richards et al., 2020):

1. Know your FactSheet consumers.
2. Know your FactSheet producer.
3. Create a FactSheet template.
4. Fill in the FactSheet template.
5. Have the producers fill in the FactSheet.
6. Evaluate the FactSheet in consultation with consumers.
7. Design additional templates and FactSheets for different audiences and purposes.

Exactly which items and how much detail should be included in a FactSheet depend on the intended audience and the nature of the AI service described. IBM researchers recommend considering questions such as (Arnold et al., 2019, 7–8):

- What is the intended use of the AI service's output?
- What algorithms or techniques does the service implement?
- Which datasets were used to train and test the service? Were they checked for bias?
- What was the testing methodology? What were its results?
- Does the service implement any bias detection and remediation procedures?
- What bias, ethical issues, or risks are known to exist with the service?
- Are the service outputs explainable or interpretable by a human user of the service?
- What is the expected performance of the service on new data?
- Was the service checked for robustness against attacks? Are they any known vulnerabilities?
- When were the models last updated? How actively is the service's architecture maintained, and by whom?

IBM Research provides several examples of completed FactSheets on their website (IBM Research, nda).

6.6 Limitations of Information Slips

Information slips are an important tool for transparency of and communication about AI models, datasets, and services to developers, consumers, and other stakeholders. There are, however, several important limitations of the information slip approach in general, each of which speaks to the importance of incorporating information slip–style documentation as a componenent of instead of a broader approach to AI and technology ethics within a firm.

One limitation is simply the newness of this approach. Because the preparation of information slips is a relatively recent innovation in AI ethics, there is a lack of standards within subfields of AI service development and domains of application. As such, there are relatively few exemplars or off-the-rack templates, meaning that most organizations will have to tailor the general approaches outlined above for their own purposes. This is valuable work, as it forces teams to reflect carefully on which aspects of their models, datasets, and services require inclusion in information slips – which may also prompt further ethical analysis – but this step also introduces additional costs in terms of time, money, and training.

In addition, most information slip approaches lack a specification of the procedures necessary to perform an adequate ethical review of a dataset, model, or service. The approaches designed by Bender, Gebru, Mitchell, and their colleagues suffer from this limitation most significantly, though so do IBM's FactSheets – the dataset nutrition labels are the exception, as they include modules designed to perform specific kinds of ethical review. Most information slip approaches thus need supplementation by sound procedures for conducting ethical review of datasets, models, and services.

Leading on from the lack of ethics review specifications in many information slip approaches, it is important to emphasize that information slips are not a replacement for an organizational environment that supports and requires ethical assessment at each step of the product lifecycle, from design to implementation. Nor are they a solution to an environment that *fails* to provide such support. As designed, information slips can be produced simply at the end of the design cycle, without any other form of ethical oversight or review. There is thus a risk that these slips will operate as an *ethical smokescreen*: a corporation could produce information slips as an afterthought in order to give the *appearance* of operating ethically, when in actuality, their information slips were reconstructed after the fact in order to comply with the standard. To be effective, information slips must be part of an ongoing process of internal and external ethical review that takes place throughout the design process, and also iterates after datasets, models, and services are released to stakeholders.

Chapter 7

AI Ethics Teaching Tools

A persistent difficulty in practical AI ethics – and technology ethics generally – is that many technologists lack training or education to prepare them to identify and address ethical issues that arise during the lifecycle of their products. On the other side, philosophers and other professionals with ethical training typically lack an understanding of AI and other technologies underlying it, and members of the public typically lack knowledge in *both* the technical and the philosophical aspects of AI ethics. To address each of these knowledge shortfalls, several practical teaching tools have been devised, which are aimed at different stakeholders in AI.

7.1 Harvard's Embedded EthiCS Modules

Since Spring term 2016, computer science courses at Harvard have included *Embedded EthiCS modules*, ethics lessons that are integrated into the technical material of the course. Developed by philosophers in consultation with computer scientists and taught by graduate students and postdoctoral fellows in philosophy, these modules cover a variety of topics from privacy to algorithmic bias to autonomous weapon systems. The programme has been well-received by both faculty and students, and the Embedded EthiCS brand is gaining popularity in other computer science programmes. A repository of previously taught modules is available online under a Creative Commons Attribution licence (Embedded EthiCS @ Harvard University, nd). Modules specifically of interest to AI ethics include:

- Verifiably ethical software systems (Grant, 2018)
- Autonomous weapons systems (Grant, 2020)
- Discrimination in Machine Learning (Navas, 2020)
- Privacy and statistical inference from data (Navas, 2019)
- Automation and the value of work (Vredenburgh, 2019)

While designed for undergraduate computer science classes, the activities in each module can stand alone and be adapted to other contexts.

7.2 Princeton's AI Ethics Case Studies

As part of an AI ethics educational initiative, researchers at Princeton University have developed six fictional, but realistic, case studies to prompt reflection and discussion (University Center for Human Values and Center for Information Technology Policy, 2021). These cases were designed following five principles:

1. **Empirical Foundations**: The examples are fictional, but are rooted in real life incidents.
2. **Broad Accessibility**: The cases are intended to be understandable by specialists in various professions and academic fields. Accordingly, they explain important concepts and background knowledge from ethics, law, and engineering.
3. **Interactiveness**: The cases come packaged with both short and long discussion questions, intended to prompt different degrees of reflective thinking on a variety of issues raised by the cases.
4. **Multiple Viewpoints**: The cases are described from multiple perspectives (e.g., technologist and stakeholder) to highlight differences in values and to discourage simplistic narratives where one agent is the "bad actor" and all others are innocent.
5. **Depth over Brevity**: In order to capture the richness and complexity of real life cases, the case studies are long (5–10 pages) and bring up multiple ethical issues.

The six case studies are based on the following topics, raising the ethical issues noted in parentheses:

- **Automated Healthcare App** (Foundations of legitimacy, Paternalism, Transparency Censorship, Inequality)
- **Dynamic Sound Identification** (Rights, Representational harms, Neutrality, Downstream responsibility)
- **Optimizing Schools** (Privacy, Autonomy, Consequentialism, Rhetoric)
- **Law Enforcement Chatbots** (Automation, Research ethics, Sovereignty)
- **Hiring By Machine** (Fairness, Irreconcilability, Diversity, Capabilities, Contextual integrity)
- **Public Sector Data Analytics** (Democracy, Secrecy, Inequality, Fallibility, Determinism)

Given their length, the case studies require a longer period of time to work through: it takes approximately 15–20 minutes to read the case and approximately 45–90 minutes to reflect on and answer the questions. The ideal format is small group discussion facilitated by one or more interdisciplinary seminar leaders. It would additionally be helpful for participants to pre-read the case study and think about at least some of the discussion questions.

7.3 Art as an AI Ethics Teaching Tool

Drawing on literature on art as a teaching and communication tool, Srinivasan and Uchino (2021) argue that art can be used as a powerful way of teaching technologists and the public about machine learning and AI. Pointing to the use of art in education and science & technology communication, they suggest that the arts have an important role to play in both raising awareness of ethical issues connected to AI, as well as enhancing technologists' moral knowledge and empathy for their products' stakeholders. For example, as part of communicating her research on racial bias in facial recognition algorithms (Buolamwini and Gebru, 2018), Buolemwini makes striking use of photographs of herself wearing a porcelain-white mask that contrasts sharply with her dark brown skin. These complement various video lectures where she demonstrates a facial recognition algorithm failing to detect her actual smiling face, but instantly recognizing the inhumanly pale and emotionless mask.

The following subsections describe a few more examples of how art can be used to teach and communicate about AI and AI ethics.

7.3.1 Apps as Art

Srinivasan and Uchino cite another case illustrating the dangers of badly trained image processing models: an app called "ImageNet Roulette." While browsing the "People" category in ImageNet, a widely used dataset of photographs compiled by researchers at Princeton and Stanford, Crawford and Paglen noticed some disturbing trends in the way the images had been categorized and tagged by Amazon Mechanical Turk workers. They note that a photograph of a child wearing sunglasses was categorized as "failure, loser, non-starter, unsuccessful person," and a smiling woman in a bikini as "slattern, slut, slovenly woman, trollop" (Crawford and Paglen, 2019). In order to study the downstream effects the dataset, they commissioned the creation of an app based on a model trained on ImageNet's "Person" category called ImageNet Roulette. The app took an uploaded photograph of a person and returned tags and classifiers that the app determined to be appropriate. As reported widely in the press, for many users, the app produced amusing miscategorizations, but for others, the results were insulting or offensive. For instance, a photo of *Guardian* reporter Julia Carrie Wong returned a racial slur (Wong, 2019). Srinivasan and Uchino identify the app as an art project that had a real impact both on public understanding of bias in image processing models and on the dataset managers themselves: since the appearance of the app (which is no longer available), ImageNet has removed hundreds of thousands of images in the "People" category.

7.3.2 Data Murals

To highlight what they mean by "data for co-liberation," D'Ignazio and Klein discuss the work of Emily and Rahul Bhargava, artists and researchers who have worked with multiple communities on art installations they call *data murals* (D'Ignazio and Klein, 2020). The Bhargavas work with lay communities to identify areas where data collection and analysis can further their goals. They assist them in performing this technical work, then facilitate events with the goal of creating murals that explain their findings to the community. For example, in Sommerville, MA, the Bhargavas worked with an urban agriculture nonprofit to involve youth in a project to promote making healthy food choices. Local young people participated in a workshop that culminated in a series of murals drawing on the group's data analysis to explain the importance of access to healthy food, and the nonprofit's efforts to reclaim urban spaces for local vegetable gardens (Bhargava and Bhargava, 2013). Community members remarked that the murals had worked as didactic tools, and associated events gave youth the opportunity to connect with their elected officials and explain issues that were important to them.

7.3.3 The Library of Missing Datasets

Another artwork discussed by D'Ignazio and Klein is Mimi Ọnụọha's work, *The Library of Missing Datasets*. The artwork consists of "a list of datasets that one might expect to already exist in the world, because they [would] help to address pressing social issues, but that in reality have never been created" (D'Ignazio and Klein, 2020). Taking the form of a physical filing cabinet stuffed with empty folders with labels such as "People excluded from public housing because of criminal records," "Total number of local and state police departments using stingray phone trackers (IMSI-catchers)," "Publicly available gun trace data," "Quantifiable effect of corruption in lean economies," and "Number of mosques surveilled by FBI agents" (Ọnụọha, 2016). A second project, *The Library of Missing Datasets 2.0*, focuses on missing datasets that would be of use in confronting social injustices faced by African Americans, highlighting that Black folks "[feature] strongly as objects of collection but rarely as subjects with agency over collection, ownership, and power" (Ọnụọha, 2018). The newer library includes folders with labels such as "Accurate birth registration data in Rwanda," "Number of Americans without bank accounts in 2008," and "Demographics of all BitCoin users."

7.3.4 Games as AI Education through Art

In a recent philosophical book, C. Thi Nguyen makes the case that games are a distinctive art form. According to him, games use human agency as their medium – similar to how painting uses paints and canvas or literature uses words as their media – to enable aesthetic experiences. Moreover, games can serve an educative function by teaching us different ways of using our agency. A game refines a very specific way of acting and narrow set of values within the context of the game, and we can take what we learn about our own ways of acting and valuing into non-game contexts. But at the same time, games can serve the function of teaching us alternative ways of relating to one another socially, by placing us in a variety of collaborative and competitive relationships structured by game rules (Nguyen, 2020). In connection with Nguyen's account of games, we can view games intended to teach players about AI or ethical issues related to AI as an instance of art as an AI ethics teaching tool.

An example from business ethics more generally is what Convercent, an ethics and compliance consulting firm, calls *compliance games*. (Near as I can tell, no AI ethics educational game has yet been developed. Searching both Morley et al.'s typology (Morley et al., 2019) and AI Global's Responsible AI Community Portal (AI Global, nd) returned zero results.) These ten games are intended to improve the experience of ethics and compliance training by making the activities more fun and engaging. These activities range from trivia games, to debates on ethical dilemmas, to having participants compete to identify ethics or compliance issues in a mock-up of a workplace scenario. Strictly speaking, not all of these activities can be considered games, and others are only minimally gamified by introducing an element of competition. Convercent notes that they had experimented with board and card games but that they had not identified a design that was effective in the long term (Read, 2018).

However, one point of caution on gamifying education that Nguyen (2020) discusses is the risk that game-players may begin to focus too much on the ends dictated by the game, and not the more valuable activities that the gamified elements are intended to encourage. For example, an AI ethics evaluation that awards the AI product a score may encourage firms to focus only on maximizing the number of points their product can earn, losing sight of the value and importance of doing a thoughtful ethical review for its own sake. The issue is similar, in this respect, to how firms will sometimes focus too much on legal compliance at the expense of deeper ethical reflection, confusing the concepts of what is permitted *by law* and what is *morally* permissible. Caution must therefore be exercised in the presentation of ethics education and evaluation activities that incorporate game-like elements.

7.3.5 The Tarot Cards of Tech

Developed by the Artefact Group, a Seattle-based strategy and design firm, the Tarot Cards of Tech are a set of 12 cards designed to resemble the major arcana of a traditional tarot deck or the face cards in a classic set of French-suited playing cards. The cards are organized into three themes, with each card's face design inspired by the technology sector, social/news media, and human-technology interaction; the reverse of each card contains discussion-prompting questions (Artefact Group, 2018):

- **Scale and Disruption**
  - The Scandal (potential for harm caused by the product and resulting bad press)
  - The Smash Hit (how the product or society might change if the product is adopted by a large number of people, say 100 million or more)
  - The Radio Star (the economic impact of the product's success, particularly in terms of which jobs or industries may cease to exist – think, "video killed the radio star")

  - Mother Nature (the environmental impact and sustainability of the product)

- **Usage**
  - The Siren (ways using the product too much be detrimental to someone's health or social life)
  - The BFFs (ways the product could enhance, damage, or change personal relationships and social roles)
  - The Superfan (how the most passionate users of the product might behave, for good or for ill)
  - The Big Bad Wolf (how the product could be exploited by abusers, criminals, and other bad actors)

- **Equity and Access**
  - The Forgotten (the groups who have been excluded from considerations of who the typical user is, and the perspectives that have not been included in the design process)
  - The Service Dog (ways the product could be designed specifically to benefit underserved users)
  - The Catalyst (how cultural mores might change the product, or be changed by the product, over time)
  - The Backstabber (how people might lose trust in the product, and how the designers plan to mitigate these events)

Also included is a prompt sheet listing topics categorized into Life, Society & Culture, and Environment, to seed discussion when participants are stuck. The cards can be accessed digitally at `tarotcardsoftech.artefactgroup.com`, or downloaded for free as a PDF suitable for printing.

The cards are, in themselves as artefacts, interesting art pieces that should provoke reflection and discussion. But the inclusion of the discussion questions on the back, which are informed by an essay on "humanity-centred design" written by the Artefact Group's CEO (Girling, 2017), add structure to that reflection and discussion, making the cards suitable for a workshop on technology ethics. The topics are somewhat higher-level than AI ethics specifically, but many of the issues considered are the same. The cards could also be adapted to other gamified discussion sessions, perhaps modeled on an actual tarot reading, which could generate greater engagement among workshop participants.

Chapter 8

Assessment and Auditing Tools

A number of tools for simplifying the assessment and auditing of AI systems have been developed.

8.1 Responsible AI Design Assistant

Designed by AI Global, a non-profit dedicated to helping firms develop responsible and ethical AI systems and to inform policymakers as AI and technology regulations are drafted, the Responsible AI Design Assistant is an online survey intended to help AI developers identify potential ethical issues in their design. The assistant has been in active development since the launch of the beta in April 2020 (AI Global, 2020).

The assistant takes a modified form of the five principles described by Morley et al. (2020) as its starting point:

1. Accountability
2. Explainability and Interpretability
3. Data Quality
4. Bias and Fairness
5. Robustness

The tool comprises the assistant itself, and a brief guide for how to use it. The assistant collects contact information, then steps through several questions including Lickert-style ratings, queries about compliance with legislation and ISO standards, questions about procedures and reviews at the developer's organization, anticipated risks, and so on, for each of the five principles listed above. At the end, the survey generates an evaluation of the project on the each of the five principles, summarized as a rating (Needs Improvement, Acceptable, or Proficient), and a radar chart visualization.

One advantage of the design assistant is that it offers a standard process for evaluating an AI system that recognizes its nature as a socio-technical system: multiple questions are about the working environment in which the AI was developed and the policies, procedures, and training exercises that are completed therein. This shows that AI Global has taken on board the view that the values of the people and organizations inform the technologies that they create, and become embedded in those technologies.

At the same time, the standard format of the design assistant's evaluations is subject to a criticism similar to Floridi (2019)'s complaint that some AI ethics frameworks are too strict. However, instead of giving rise to "ethics shopping," the risk here is of a mis-match between the tool and the context in which it is being used. Every standard must perforce abstract away from context and specificity in order to produce quantifiable and comparable data. It is probable that in the process, at least some cases of AI systems design and development will fail to be properly evaluated – and may not be evalu*able* – by the assistant.

Another potential disadvantage of the assistant is that the survey is comprehensive, and likely requires multiple team members from different organizational units to complete. The survey makes use of technical terms from ethics and engineering and refers to legal and ISO standards, each of which require subject matter specialists to interpret correctly. The tool may thus not be suitable for use in all organizational contexts.

8.2 Harms Modeling

Harms modeling is a qualitative approach for assessing the potential harms of a technology. The version developed for Microsoft's Azure cloud computing services fits into the following recommended design procedure (Cassidy et al., 2020b):

1. Define the purpose of the application.
2. Describe use cases for the application.
3. Consider stakeholders, with particular attention to marginalized groups.
4. **Assess the application's potential for harm**.
5. Make a plan for mitigating harms and maximizing positive outcomes.

The harms modeling process itself involves considering the following types of harm, each of which is subdivided into more specific harms (Cassidy et al., 2020c):

- **Risk of Injury**
  - Physical Injury
  - Emotional or Psychological Injury
- **Denial of Consequential Services**
  - Opportunity Loss

    - Economic Loss

- **Infringement on Human Rights**
    - Dignity Loss
    - Liberty Loss
    - Privacy Loss
    - Environmental Impact
- **Erosion of Social & Democratic Structures**
    - Manipulation
    - Social Detriment

For each relevant harm, developers are instructed to evaluate the harm's severity, scale, probability, and frequency.

This assessment framework is comprehensive, and well-informed by the philosophical, legal, and sociological literature. Because of the breadth of topics involved, completing a harms modeling evaluation may require coordinating between multiple members of the development team as well as other organizational units and external stakeholders.

8.3 Community Jury

The Azure documentation suggests one way to facilitate consultations with stakeholders that they call a *community jury*. The basis of this process is the *citizens jury*, as developed by the Center for New Democratic Processes (nd). Similar to a town hall discussion, a citizens jury gathers interested members of a community to discuss an issue or problem that affects them. However, in a typical town hall meeting, citizens are restricted to receiving information from and criticizing elected officials. By contrast, a citizens jury gives community members a significant amount of resources and influence. The general form is a process of gathering information from experts, deliberating on potential solutions, and making recommendations to be passed on to decision-makers. The Center for New Democratic Processes describes the procedure as follows:

1. Understand the challenge.
2. Design the process.
3. Invite the community.
4. Select participants.
5. Provide background information.
6. Facilitate deliberation.

7. Create recommendations.

8. Amplify and share.

Key to the success of a citizens jury is having a representative subset of the community with a mix of relevant expertise and diverse perspectives, supplying jurors with accurate information, and giving jurors sufficient time to work through the problem.

The community jury approach for AI systems follows a similar procedure for consultations with stakeholders on potential harms of an AI system. The Azure documentation outlines a process specific to this context, with recommendations on selection of community participants, experts to provide background knowledge, and protocol to follow to facilitate the jury discussion (Cassidy et al., 2020a). While citizens juries can take 3–6 days (or more, for complex and highly sensitive issues such as national security) to reach a consensus on recommendations, the Azure documentation suggests that a community jury can complete its discussion in a 2–3 hour session. This may suffice for narrow scope AI-driven applications, but the wider scope of AI applications, and the greater sensitivity of data processed by them, in the public sector or in other domains with wider and more complex potential for harm (e.g. finance, insurance, telecommunications) may require more in-depth and lengthy deliberation.

8.4 Aequitas Bias and Fairness Audit

Aequitas is a tool created by researchers at the Center for Data Science and Public Policy of the University of Chicago for developers to self-assess machine learning models, specifically risk assessment models such as those used by policymakers, for signs of bias and unfairness. The tool allows developers to upload their own data, or a test dataset, and generates a report noting various sources of bias. There are two main types of bias assessed by Aequitas (Saleiro et al., 2019):

1. Biased actions or interventions that are not allocated in a way that is representative of the population.

2. Biased outcomes through actions or interventions that are a result of the system being wrong about certain groups of people.

To perform an assessment, developers must supply the following kinds of data:

- Data about the general populations whose information will be processed by the model, and their demographic markers (e.g., gender, race, socio-economic status, etc.)

- Data about the specific sub-populations that the (post-training) model selects for policy interventions

- Information about the outcomes of interventions

Aequitas can be used as a webtool, as a Python library that can be imported into code, or downloaded as a tool that runs locally from the command line. After processing the input using statistical tools, the tool produces a summary report, a detailed set of statistics on fairness and bias, and an interactive dashboard presenting this information visually. The tool also allows developers to choose a type of fairness model to use to assess their machine learning model, depending on the desired kind of representation among those selected for interventions (Saleiro et al., 2018).

The Aequitas audit is an interesting proof of concept for how statistical tools can be used to review models for fairness. It is limited in that it was developed for a very specific type of model, but the open source code can be adapted to other kinds of projects. Another limitation is that it is based on statistical notions of fairness, which may not always correspond to what philosophically informed understandings of justice or ethics may suggest.

Chapter 9

Conclusion

This report has surveyed only a small portion of the hundreds of AI ethics tools that are available, with more appearing regularly. The hope is that by presenting them alongside the theoretical background that informs them, as well as the work by Morley, Floridi, and their colleagues to make more practically usable tools that are still well-grounded in research, the development of more effective AI ethics tools can be facilitated.

By way of general conclusion, it is worth noting the wide range of material that informs the tools and theories canvassed here. Work ranging from philosophical ethics, to sociological theory, to statistical methods, to technical writing standards and even nutrition labelling, have informed the tools and principles discussed above. Clearly, AI ethics requires interdisciplinary collaboration at all levels, reaching across and beyond the academy.

Bibliography

Christophe Abrassart, Yoshua Bengio, Guillaume Chicoisne, Nathalie de Marcellis-Warin, Marc-Antoine Dilhac, Sébastien Gambs, Vincent Gautrais, Martin Gibert, Lyse Langlois, François Laviolette, Pascale Lehoux, Jocelyn Maclure, Marie Martel, Joëlle Pineau, Peter Railton, Catherine Régis, Christine Tappolet, and Nathalie Voarino. 2018. *The Montréal Declaration for a Responsible Development of Artificial Intelligence.* https://www.montrealdeclaration-responsibleai.com/the-declaration

ACM. 2018. ACM Code of Ethics and Professional Conduct. https://ethics.acm.org/code-of-ethics/

AI Global. 2020. Responsible AI Design Assistant. https://oproma.github.io/rai-trustindex/

AI Global. [n.d.]. Responsible AI Community Portal. https://portal.ai-global.org/resources

Algorithm Watch. 2020. AI Ethics Guidelines Global Inventory. https://inventory.algorithmwatch.org/

Elizabeth Anderson. 2014. Social Movements, Experiments in Living, and Moral Progress: Case Studies from Britain's Abolition of Slavery. *The Lindley Lecture* 52 (2014), 28 pages. https://kuscholarworks.ku.edu/handle/1808/14787

Aristotle. 1999. *Nicomachean Ethics* (2nd ed.). Hackett, Indianapolis, IN.

Aristotle. 2011. *Eudemian Ethics.* Oxford University Press, Oxford, UK.

Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv:1808.07261 [cs.CY]

The Artefact Group. 2018. The Tarot Cards of Tech. http://tarotcardsoftech.artefactgroup.com/

Jonathan Ball. 2019. The Double Diamond: A universally accepted depiction of the design process. https://www.designcouncil.org.uk/news-opinion/double-diamond-universally-accepted-depiction-design-process

Beijing Academy of Artificial Intelligence. 2019. Beijing AI Principles. https://www.baai.ac.cn/news/beijing-ai-principles-en.html

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Canada, virtual). ACM, New York, NY, 610–623. https://doi.org/10.1145/3442188.3445922

Ruha Benjamin. 2019a. Assessing Risk, Automating Racism. *Science* 366, 6464 (October 2019), 421–422. https://doi.org/10.1126/science.aaz3873

Ruha Benjamin. 2019b. *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity, Cambridge, UK.

Rahul Bhargava and Emily Bhargava. 2013. Mural-ing Our Way to Data Literacy. https://civic.mit.edu/2013/08/06/mural-ing-our-way-to-data-literacy/

Ellen Broad, Amanda Smith, and Peter Wells. 2017. *Helping Organizations Navigate Ethical Concerns in the Data Practices.* Technical Report. The Open Data Institute, London, UK. https://www.scribd.com/document/358778144/ODI-Ethical-Data-Handling-2017-09-13

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

Ania Calderon, Dan Taber, Hong Qu, and Jeff Wen. 2019. AI BlindSpot. https://aiblindspot.media.mit.edu/

Rafael A. Calvo and Dorian Peters. [n.d.]. Responsible Design Process. http://www.positivecomputing.org/p/process.html

Doug Cassidy, Alex Buck, Dhanashri Kshirsagar, and Harmony Mabrey. 2020a. Community Jury. https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/

Doug Cassidy, Alex Buck, Dhanashri Kshirsagar, and Harmony Mabrey. 2020b. Harms Modeling. https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

Doug Cassidy, Alex Buck, Dhanashri Kshirsagar, Page Writer, David Coulter, Cory Fowler, Harmony Mabrey, and Adam Boeglin. 2020c. Types of harm. https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/type-of-harm

Estelle Caswell. 2015. Color film was built for white people. Here's what it did to dark skin. https://www.vox.com/2015/9/18/9348821/photography-race-bias

Center for New Democratic Processes. [n.d.]. How We Work | Citizens Juries. https://jefferson-center.org/about-us/how-we-work/

Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2020. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. In *NeurIPS 2020 Workshop on Dataset Curation and Security*. Neural Information Processing Systems Foundation, San Diego, CA, 7 pages.

Kate Crawford and Trevor Paglen. 2019. Excavating AI. https://excavating.ai/

The Design Council. 2015. What is the framework for innovation? Design Council's evolved Double Diamond. https://www.designcouncil.org.uk/news-opinion/what-framework-innovation-design-councils-evolved-double-diamond

John Dewey and James Hayden Tufts. 1985. *Ethics*, revised edition. In *The Later Works of John Dewey, 1925–1953*, Jo Ann Boydston (Ed.), Vol. 7. Southern Illinois University Press, Carbondale, IL, 1–463.

Digital Catapult. 2021. Ethics Framework. https://www.migarage.ai/ethics/ethics-framework/

Digital Catapult. [n.d.]. Digital Catapult. https://www.digicatapult.org.uk/

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* MIT Press, Cambridge, MA. https://data-feminism.mitpress.mit.edu/

Embedded EthiCS @ Harvard University. [n.d.]. Module Repository. https://embeddedethics.seas.harvard.edu/module.html

European Commission. 2019. Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation

Fiona Fidler and John Wilcox. 2018. Reproducibility of Scientific Results. In *The Stanford Encyclopedia of Philosophy* (fall 2018 ed.), Edward N. Zalta (Ed.). Center for the Study of Language and Information, Stanford University, Stanford, CA, 60 pages. https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/

Luciano Floridi. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology* 32 (2019), 185–193. https://doi.org/10.1007%2Fs13347-019-00354-x

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelinand Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28 (2018), 689–707. https://doi.org/10.1007%2Fs11023-018-9482-5

Luciano Floridi and Timothy F. the Baron Clement-Jones. 2019. The five principles key to any ethical framework for AI. https://tech.newstatesman.com/policy/ai-ethics-framework

Philippa Foot. 2003. *Virtues and Vices, and other essays in moral philosophy.* Oxford University Press, Oxford, UK.

Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems* 14, 3 (1996), 330–347. https://doi.org/10.1145/230538.230561

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. In *Proceedings of the5thWorkshop on Fairness, Accountability, and Transparency in Machine Learning* (Stockholm, Sweden). ACM, New York, NY, 24 pages. https://arxiv.org/pdf/1803.09010.pdf

Carol Gilligan. 1982. *In a Different Voice: Psychological Theory and Women's Development.* Harvard University Press, Cambridge, MA.

Rob Girling. 2017. What's next for design: Towards humanity-centered design. https://www.artefactgroup.com/ideas/towards-humanity-centered-design/

Google AI. [n.d.]. Artificial Intelligence at Google: Our Principles. https://ai.google/principles

David Gray Grant. 2018. Verifiably ethical software systems. https://embeddedethics.seas.harvard.edu/classes/cs-152-2018-spring

Lyndal Grant. 2020. Autonomous weapons systems. https://embeddedethics.seas.harvard.edu/classes/cs-189-2020-spring

Hippocrates. 2002. The Hippocratic Oath. https://www.nlm.nih.gov/hmd/greek/greek_oath.html

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 [cs.DB] https://arxiv.org/abs/1805.03677

Chuck Huff. 2001. Why a Socio-technical System? http://computingcases.org/general_tools/sia/socio_tech_system.html

Human Dignity Trust. 2021. Map of Countries that Criminalize LGBT People. https://www.humandignitytrust.org/lgbt-the-law/map-of-criminalisation/

IBM Research. [n.d.]a. AI FactSheets 360. https://aifs360.mybluemix.net/

IBM Research. [n.d.]b. Trusting AI. https://www.research.ibm.com/artificial-intelligence/trusted-ai/

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

The Institute for Ethical AI in Education. 2020a. *Annex: Developing The Ethical Framework for AI in Education.* Technical Report. The University of Buckingham. https://fb77c667c4d6e21c1e06.b-cdn.net/wp-content/uploads/2021/03/Annex-Developing-the-Ethical-Framework-for-AI-in-Education-IEAIED.pdf

The Institute for Ethical AI in Education. 2020b. *The Ethical Framework for AI in Education.* Technical Report. The University of Buckingham. https://fb77c667c4d6e21c1e06.b-cdn.net/wp-content/uploads/2021/03/The-Ethical-Framework-for-AI-in-Education-Institute-for-Ethical-AI-in-Education-Final-Report.pdf

The Institute for Ethical AI in Education. 2020c. The Institute for Ethical AI in Education. https://www.buckingham.ac.uk/research-the-institute-for-ethical-ai-in-education/

Deborah G. Johnson and Keith W. Miller. 2009. *Computer Ethics: Analyzing Information Technology* (4th ed.). Prentice Hall, Upper Saddle River, NJ, and Columbus, OH.

Immanuel Kant. 1993/1785. *Grounding for the Metaphysics of Morals.* Hackett, Indianapolis, IN.

Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "Nutrition Label" for Privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (Mountain View, California, USA) *(SOUPS '09).* Association for Computing Machinery, New York, NY, USA, Article 4, 12 pages. https://doi.org/10.1145/1572532.1572538

John Locke. 1690. *Second Treatise of Government.* A. Millar, H. Woodfall, J. Whiston, B. White, et al., London, UK.

Michael P. Lynch. 2017. *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data.* Liveright, New York, NY.

Kathryn MacKay. 2020. Feminism and Feminist Ethics. In *Introduction to Philosophy: Ethics* (1.3.3 ed.), George Matthews (Ed.). Rebus Community, Web, 63–73. https://press.rebus.community/intro-to-phil-ethics/chapter/feminism-and-feminist-ethics/

Kathleen McGrory and Neil Bedi. 2020. Targeted: Pasco's sheriff created a futuristic program to stop crime before it happens. It monitors and harasses families across the county. https://projects.tampabay.com/projects/2020/investigations/police-pasco-sheriff-targeted/intelligence-led-policing/

Microsoft. [n.d.]. Microsoft AI principles. https://www.microsoft.com/en-us/ai/responsible-ai

John Stuart Mill. 1879. *Utilitarianism.* Longmans, Green, and Co., London, UK.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19).* Association for Computing Machinery, New York, NY, 220–229. https://doi.org/10.1145/3287560.3287596

Aleksandra Mojsilovic. 2018. Factsheets for AI Services. https://www.ibm.com/blogs/research/2018/08/factsheets-ai/

G. E. Moore. 1903. *Principia Ethica.* Cambridge University Press, Cambridge, UK.

Jessica Morley, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mokander, and Luciano Floridi. 2021. Ethics as a service: a pragmatic operationalisation of AI Ethics. arXiv:2102.09364 [cs.CY] https://arxiv.org/abs/2102.09364

Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2019. Applied AI Ethics Typology. http://tinyurl.com/appliedaiethics

Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26 (2020), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Diana Acosta Navas. 2019. Privacy and statistical inference from data. https://embeddedethics.seas.harvard.edu/classes/cs-265-2019-spring

Diana Acosta Navas. 2020. Discrimination in Machine Learning. https://embeddedethics.seas.harvard.edu/classes/cs-181-2020-spring

C. Thi Nguyen. 2020. *Games: Agency as Art.* Oxford University Press, New York, NY.

Safiya Umoja Noble. 2012. Missed Connections: What Search Engines Say About Women. *Bitch* 54 (2012), 36–41.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press, New York, NY.

Nel Noddings. 1984. *Caring: A Feminine Approach to Ethics and Moral Education.* University of California Press, Berkeley, CA, and Los Angeles, CA.

Robert Nozick. 1974. *Anarchy, state, and utopia.* Basic Books, New York, NY.

Justin Oakley and Dean Cocking. 2001. *Virtue Ethics and Professional Roles.* Cambridge University Press, Cambridge, UK.

OECD. 2019. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

The Open Data Institute. 2019. The Data Ethics Canvas. https://theodi.org/article/data-ethics-canvas

Dorian Peters. 2019. Beyond Principles: A Process for Responsible Tech. https://medium.com/ethics-of-digital-experience/beyond-principles-a-process-for-responsible-tech-aefc921f7317

D. Peters, K. Vold, D. Robinson, and R. A. Calvo. 2020. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society* 1, 1 (2020), 34–47. https://doi.org/10.1109/TTS.2020.2974991

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2020. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). arXiv:2003.12206 [cs.LG] https://arxiv.org/abs/2003.12206

John Rawls. 1971. *A Theory of Justice.* Harvard University Press, Cambridge, MA.

Keith Read. 2018. Increasing Employee Awareness, Participation, & Engagement through Compliance Training Games. https://www.convercent.com/blog/compliance-games

John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. 2020. A Methodology for Creating AI FactSheets. arXiv:2006.13796 [cs.HC] https://arxiv.org/abs/2006.13796

W. D. Ross. 1930. *The Right and the Good.* Oxford University Press, Oxford, UK.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2019. Aequitas: A Bias and Fairness Audit Toolkit. arXiv:1811.05577 [cs.LG] https://arxiv.org/abs/1811.05577

Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani. 2018. Aequitas. http://www.datasciencepublicpolicy.org/projects/aequitas/

Susan Sherwin. 1999. Foundations, frameworks, lenses: the role of theories in bioethics. *Bioethics* 13, 3–4 (1999), 198–205. https://doi.org/10.1111/1467-8519.00147

Ramya Srinivasan and Kanji Uchino. 2021. The Role of Arts in Shaping AI Ethics. In *AAAI 2021 Workshop - Reframing Diversity in AI: Representation, Inclusion and Power.* IBM Research, USA, 6 pages. http://ceur-ws.org/Vol-2812/RDAI-2021_paper_3.pdf

Jay Stanley. 2019. An Army of Robot Surveillance Guards Is Coming. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/army-robot-surveillance-guards-coming

The Data Nutrition Project. 2021. The Data Nutrition Project. https://datanutrition.org/

United Nations. 1948. Universal Declaration of Human Rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights

University Center for Human Values and Center for Information Technology Policy. 2021. Princeton Dialogues on AI and Ethics Case Studies. https://aiethics.princeton.edu/case-studies/

Richard Volkman. 2013. Being a good computer professional: The advantages of virtue ethics in computing. In *Professionalism in the Information and Communication Technology Industry*, John Weckert and Richard Lucas (Eds.). Australian National University E Press, Canberra, ACT, 109–126.

Kate Vredenburgh. 2019. Automation and the value of work. https://embeddedethics.seas.harvard.edu/classes/cs-189-2019-spring

Jasmine Weber. 2019. How a 19th-Century Photographic Technique Erased a Māori Tradition. https://hyperallergic.com/499222/how-a-19th-century-photographic-technique-erased-a-maori-tradition/

Julia Carrie Wong. 2019. The viral selfie app ImageNet Roulette seemed fun – until it called me a racist slur. https://www.theguardian.com/technology/2019/sep/17/imagenet-roulette-asian-racist-slur-selfie

Matthew Zook, Solon Barocas, danah boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson, and Frank Pasquale. 2017. Ten simple rules for responsible big data research. *PLOS Computational Biology* 13, 3 (2017), e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Mimi Ọnụọha. 2016. The Library of Missing Datasets. https://github.com/MimiOnuoha/missing-datasets

Mimi Ọnụọha. 2018. The Library of Missing Datasets 2.0. https://mimionuoha.com/the-library-of-missing-datasets-v-20