Chapter 7

A Principlist Approach to Ethical Considerations

In this chapter, we describe how the ethical framework we introduced in Chapter 3, based on the Belmont Principles of respect for persons, beneficence, and justice, can be applied in the context of data science. This principlist approach to ethics attempts to provide a shared analytic framework and vocabulary to help communities and teams resolve difficult questions. Principles are most useful when broad enough to be comprehensive and capture, rather than ignore, the tensions that make questions of "right" and "wrong" so difficult.

As the Belmont Report states: "the objective is to provide an analytical framework that will guide the resolution of ethical problems," however, "these principles cannot always be applied so as to resolve beyond dispute particular ethical problems." That is, our goal is not to provide a universal yes-or-no algorithm for ethics. Rather, it is to guide ethical decision-making so it provides practitioners and stakeholders a shared understanding of a decision-making process and logic.

To illustrate, we chose five of Chapter 6's use cases to explain with respect to the Belmont Principles: criminal sentencing, newsfeed recommendation, vaccine distribution, mobility reporting, and insurance underwriting. The three principles are not ranked in importance, and each example may not have concerns related to all three. As in the context of Belmont's original deliberations, "beneficence" includes not only individual harms and benefits but also those of society at large.

The observation that a data science application may not satisfy each Belmont Principle does not necessarily mean we advocate discontinuing the application. Vaccine mandates, for example, would prioritize public good over individual autonomy. The complexity of this balance is reflected in US law: "Since Jacobson v Massachusetts (1905), the judiciary has consistently upheld vaccination mandates" (Gostin et al., 2021), while there are COVID-19 vaccination decisions that show nuance. The original Belmont Report similarly notes of principles that: "at times they come into conflict" and "cannot always be applied so as to resolve

https://doi.org/10.1017/9781009272230.009 Published online by Cambridge University Press

beyond dispute particular ethical problems." Instead, they are meant to "provide an analytical framework that will guide the resolution of ethical problems."

Here are applications of the Belmont framework's three principles to some of the examples in Chapter 6.

7.1 Criminal Sentencing and Parole Decision-Making

As discussed in Section 6.6, algorithms for criminal pre-trial, sentencing, and parole decisions are fraught with ethical challenges.

- **Respect for persons.** All stakeholders' autonomy would be challenged if such algorithms were opaque: defending attorneys lack understanding of how decisions are made now, and defendants who may become incarcerated lack understanding of how their actions may be scored in such an automated decision system. The transparency paradox discussed in Chapter 3 thwarts complete information. Instead of over-explaining technical subjects or under-explaining (which can lead to deceptive or unfair practices), proponents of increased use of automated decision systems can adopt a "tiered" approach (Bunnik et al., 2013). This would involve explaining the basic concepts in plain language while providing extra detail to those who want increased transparency. Another concern is the likely event that algorithmic approaches do not provide interpretability. The answer to "why" a decision is rendered is both important and perhaps not answerable.
- **Beneficence.** The claimed benefits of using such algorithms, such as efficiency and uniformity, must be explicitly evaluated with respect to their impact on many parties and performance versus existing human approaches. These include defendants, potential parolees and other stakeholders, including the judiciary, the criminal justice system, and society writ large. For example, algorithms may be trained to minimize expected errors against a test set (e.g., prior decisions by human judges). However, this training may not minimize the total number of future crimes, the total expense of the justice system, or other more beneficial, societal goals.
- **Justice.** Biases resulting from deploying algorithms for criminal sentencing and parole decision support – e.g., different model outputs for demographically different defendants with similar criminal data – is the subject of ample research and journalistic inquiry. Many of these inquiries argue, via statistics as well as case studies and interviews, that some are benefiting relative to others. Also, perfectly accurate models trained on biased data perfectly reproduce these biases. As discussed in Section 6.6, the use of data science in algorithmic sentencing and parole decisions catalyzed research discussions on the multiple ways to quantify fairness (Pleiss et al., 2017). We discuss this in more detail in Section 12.3.

https://doi.org/10.1017/9781009272230.009 Published online by Cambridge University Press

In short, this example includes a variety of data science and societal challenges. The ethical considerations are clear, with reduced accountability, serious risk of harm to individuals and society, and opportunity for amplifying unfairness and injustice as applied to individuals. However, there could be enhanced judicial uniformity and possibly other benefits. Aside from ethical considerations, different stakeholders (defendants, lawmakers, members of the judiciary) might have extremely different notions of what objective an algorithm is to optimize.

7.2 News Feed Recommendations

As discussed briefly in Section 4.3 and Section 6.2, news feed recommendations are considerably more challenging than music recommendations. News corpora vary in size and quality, as well as the motivations of those who add to them. News also has a much greater impact on individuals and society.

- **Respect for persons.** Informed consent is challenged when news recommendation algorithms are opaque. The transparency paradox is exacerbated given their complexity, which prevents even their designers from fully understanding the resulting technical systems. A widely discussed example of informed consent was Facebook's 2012 "Emotional Contagion" automated news feed experiment, published in the *Proceedings of the National Academy of Sciences* in 2014 (Verma, 2014). In this experiment, Facebook users' news feeds were experimentally manipulated to amplify posts of positive or negative sentiment. Subsequent posts by those users were then used to quantify whether the algorithmic changes caused subsequent increased or decreased happiness. Section 12.4 has more detail on this.
- **Beneficence.** These algorithms can indeed inform and delight when delivering content that optimizes engagement through joy and surprise. However, they can also create filter bubbles (Section 12.4.2) or amplify fear or hate, leading to a radicalizing "rabbit hole" (Alfano et al., 2018). The complexity of content ranking algorithms, and the unpredictability of their impact on users' well-being, is driving companies to spend more time investigating unanticipated ways content recommendation may lead to harm. Of course, they also have to develop ways to mitigate these effects (Wells et al., 2021). We need long-term studies, where users' behavior is observed for several weeks or months, for developers to know the impact of a news feed's content recommendation algorithm.
- **Justice.** Without question, algorithms affect different societal groups in different ways, leading to many considerations of fairness. Some impacts are benign (perhaps, a propensity for a subgroup to get sports entertainment recommendations), while others reinforce societal problems.

https://doi.org/10.1017/9781009272230.009 Published online by Cambridge University Press

We have chosen this example because there has been increasing societal reliance on algorithmic news feeds. This has led to increased public scrutiny. The complexity of the algorithms prevents designers and readers alike from understanding what content and world view is being amplified. As to harms, the attribution of benefit and risk to these algorithms is debated daily in the press, by researchers, by companies doing news recommendations (including their own researchers), and by lawmakers and regulatory agencies.

7.3 Vaccine Distribution Optimization

Distributing a vaccine with supply, logistical, and vaccine hesitancy constraints is a truly complex problem, as briefly discussed in Section 6.3.

- **Respect for persons.** This is a concern, as societies increasingly pressure and even mandate that vaccine skeptics be vaccinated. If vaccines were outright forced, vaccination would violate the principle of informed consent. Society-level rationing also reduces individual autonomy.
- **Beneficence.** Beneficence's role in vaccine distribution, particularly since both the supply and distribution channels may be limited, is extremely difficult to know in advance. The coupling of optimization and health policies requires complex trade-offs, e.g., between supporting opening schools or reducing disease in prisons or assisted living facilities. Here, a commitment to beneficence includes adjusting distribution policy as supplies and health policies change and as the effects of a distribution policy become evident.
- **Justice.** At the individual distributor level, whether state or private, appointment booking mechanisms have varied usability and interface complexity. The technology divide may contribute to a "vaccine divide" among those with and without the technology access to secure vaccine appointments. The digital divide may correlate with demographic divides and could result in unfair outcomes.

We have included this example because of its widespread importance and serious ethical complexity. Even though data-driven models are needed to optimize vaccine distribution, ethics are of paramount importance in this example, and balancing objectives is particularly complex.

7.4 Mobility Reporting

The Google team that showed aggregate regional movement trends was cognizant of the need to preserve individual location privacy and effectively used differential privacy techniques to protect that data. (See Section 10.1 for more on differential

https://doi.org/10.1017/9781009272230.009 Published online by Cambridge University Press

privacy.) However, even if individual privacy is preserved, using geographical data gives rise to ethical issues worthy of consideration.

- **Respect for persons.** Neither any individual nor Google would have contemplated in advance that anonymized location data would be used for this purpose. However, Google's anonymization policy is explicitly written to allow Google wide latitude in the use of anonymized data (Google, n.d.-b). On the one hand, users might be surprised, given how few users read the policy. On the other hand, opt-in would have greatly reduced the likely effectiveness of this application.
- **Beneficence.** Correlations between who is mobile and how disease progresses offer societal benefit for informing health policy. However, there may be implications to individuals, even if identities cannot be inferred from published mobility data. For example, mobility data could be correlated with widely available demographic data and reinforce stereotypes or create societal divides. This same effect could occur in many other applications that aggregate anonymized individual data.
- **Justice.** Since mobility data is often gathered via smartphones, it risks being skewed towards those users. This requires a careful analysis of the policy's efficacy and impact based on such data. Sampling bias is discussed more in multiple places in Part III.

We used this example to show there are ethical concerns beyond the most obvious one, which is privacy. It shows the subtle issues a design team must navigate even when the primary goal is to produce an information system intended to benefit public health and policy.

7.5 Underwriting/Pricing of Property/Casualty Insurance

Ascertaining risk to enable better selection pricing of insurance policies is a traditional application of data, and it is significant given how important it is to people. As discussed in Section 6.5, data can be applied to many aspects of the problem space.

- **Respect for persons.** Availability and pricing of insurance should be based on the specific risks of an individual application, not exogenous factors which may not be related. Opaque algorithms which set loan policies and insurance rates for individuals challenge the concept of informed consent, as neither applicants nor insurance regulators may be able to determine the rationale for an underwriting decision. Some who particularly need insurance (e.g., those in fragile economic circumstances) have diminished autonomy, possibly meriting increased protections against hard-to-understand or deceptive terms and conditions.

https://doi.org/10.1017/9781009272230.009 Published online by Cambridge University Press

- **Beneficence.** Increased use of more personal data can itself affect the risk-taking behavior of individuals or groups. For example, high penetration of insurance and low reimbursements could drive practitioners out of a medical subfield, causing societal harm. Insurance's increased use of personal health information could also motivate individuals to avoid useful health diagnostic tests, thus causing societal harm due to a lack of preventative testing or even increased disease transmission.
- **Justice.** Such algorithms can reinforce societal bias, e.g., if they are accurately trained to reproduce biased human insurance underwriting decisions, they would constitute a form of "digital redlining." Data could facilitate the creation of new, finer-grained risk pools (e.g., assigning people with genetic predisposition to disease), thereby increasing differentials in insurance costs. This unequally distributed harm illustrates the Belmont Report's multiple meanings of justice, "in the sense of 'fairness in distribution' or 'what is deserved,'" to quote the original.

We include this example to illustrate how a mechanism that pre-dates digital computation can, by including far more data and complex algorithms, risk amplification of already present harms and injustice.

To close, we refer back to Gottenbarn and Wolfe, who state, ". . . every decision requires us to identify a broader range of stakeholders and consider how to satisfy our obligations to them. A primary function of the Code is to help computing professionals identify potential impacts and promote positive outcomes in their systems" (Gotterbarn et al., 2018). While they are talking about computing and the ACM Code, their quote is also consistent with our principlist approach to ethics. It underlies our view, as demonstrated with this section's five examples, that ethics must be considered as many types of decisions are made. We believe that doing such analyses against a set of principles, like the Belmont Principles, (1) reminds data scientists to think about difficult challenges, (2) acts as a check on significant errors, and (3) motivates practical improvements.

https://doi.org/10.1017/9781009272230.009 Published online by Cambridge University Press