15

Ethics Principles

The fourth attribute of trustworthiness, introduced in Chapter 1, includes low self-orientation, motivation to serve others' interests as well as own interests, benevolence, and an aligned purpose. This chapter focuses on this fourth attribute and kicks off the sixth and final part of the book (remember the organization of the book illustrated in Figure 15.1). Introduced in Chapter 14, value alignment is composed of two halves: technical and normative; this chapter deals with the normative part. Unlike earlier chapters, this chapter is not presented through a fictional use case.

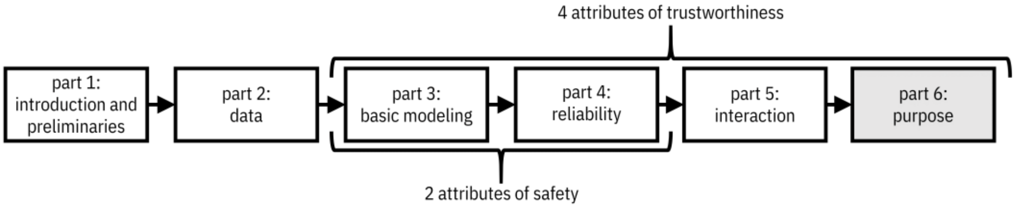

Figure 15.1. *Organization of the book. The sixth part focuses on the fourth attribute of trustworthiness, purpose, which maps to the use of machine learning that is uplifting.* Accessible caption. A flow diagram from left to right with six boxes: part 1: introduction and preliminaries; part 2: data; part 3: basic modeling; part 4: reliability; part 5: interaction; part 6: purpose. Part 6 is highlighted. Parts 3–4 are labeled as attributes of safety. Parts 3–6 are labeled as attributes of trustworthiness.

Benevolence implies the application of machine learning for good purposes. From a consequentionalist perspective (defined in chapter 14), we should broadly be aiming for good outcomes for all people. But a single sociotechnical system surely cannot do that. So we must ask: whose good? Whose interests will machine learning serve? Who can machine learning empower to achieve their goals?

The values encoded into machine learning systems are an ultimate expression of power. The most powerful can push for their version of 'good.' However, for machine learning systems to be worthy of trust, the final values cannot only be those that serve the powerful, but must also include the values of the most vulnerable. Chapter 14 explains technical approaches for bringing diverse voices into the value alignment process; here we try to understand what those voices have to say.

But before getting there, let's take a step back and think again about the governance of machine learning as a control system. What do we have to do to make it selfless and empowering for all? As shown in Figure 15.2, which extends Figure 14.5, there is a *paradigm*—a normative theory of how things should be done—that yields principles out of which values arise. The values then influence modeling.

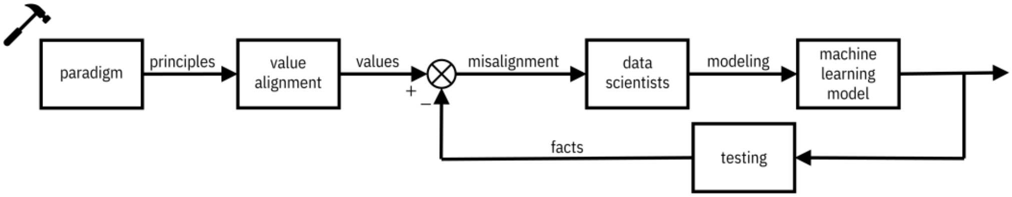

Figure 15.2. *A paradigm determines the principles by which values are constructed. The paradigm is one of the most effective points in the system to intervene to change its behavior.* Accessible caption. A block diagram that starts with a paradigm block with output principles. Principles are input to a value alignment block with output values. Facts are subtracted from values to yield misalignment. Misalignment is input to a data scientists block with modeling as output. Modeling is input to a machine learning model with output that is fed into a testing block. The output of testing is the same facts that were subtracted from values, creating a feedback loop. Paradigm is intervened upon, shown using a hammer.

There are many *leverage points* in such a complex system to influence how it behaves.[1] Twiddling with parameters in the machine learning model is a leverage point that may have some small effect. Computing facts quickly and bringing them back to data scientists is a leverage point that may have some slightly larger effect. But the most effective leverage point to intervene on is the paradigm producing the principles.[2] Therefore, in this chapter, we focus on different paradigms and the principles, codes, and guidelines that come from them.

15.1 Landscape of Principles

Over the last several years, different groups from different sectors and different parts of the world have created ethics principles for machine learning (and artificial intelligence more broadly) that espouse their paradigms. Organizations from private industry, government, and civil society (the third sector that is neither industry nor government, and includes non-governmental organizations (NGOs)) have produced normative documents at similar rates. Importantly, however, organizations in more

[1] Donella H. Meadows. *Thinking in Systems: A Primer*. White River Junction, Vermont, USA: Chelsea Green Publishing, 2008.

[2] More philosophically, Meadows provides an even more effective leverage point: completely transcending the idea of paradigms through enlightenment.

economically-developed countries have been more active than those in less economically-developed countries, which may exacerbate power imbalances. Moreover, the entire framing of ethics principles for machine learning is based on Western philosophy rather than alternative conceptions of ethics.[3] There are many similarities across the different sets of principles, but also key differences.[4]

First, let's look at the similarities. At a coarse-grained level, five principles commonly occur in ethics guidelines from different organizations:

1. privacy,
2. fairness and justice,
3. safety and reliability,
4. transparency (which usually includes interpretability and explainability), and
5. social responsibility and beneficence.

This list is not dissimilar to the attributes of trustworthiness that have guided the progression of the book. Some topics are routinely omitted from ethics principles, such as artificial general intelligence and existential threats (machines taking over the world), and the psychological impacts of machine learning systems.

Differences manifest when looking across sectors: governments, NGOs, and private corporations. Compared to the private sector, governments and NGOs take a more participatory approach to coming up with their principles. They also have longer lists of ethical principles beyond the five core ones listed above. Furthermore, the documents espousing their principles contain greater depth.

The topics of emphasis are different across the three sectors. Governments emphasize macroeconomic concerns of the adoption of machine learning, such as implications on employment and economic growth. NGOs emphasize possible misuse of machine learning. Private companies emphasize trust, transparency, and social responsibility. The remainder of the chapter drills down into these high-level patterns.

15.2 Governments

What is the purpose of government? Some of the basics are law and order, defense of the country from external threats, and general welfare, which includes health, well-being, safety, and morality of the people. Countries often create national development plans that lay out actions toward improving general welfare. In 2015, the member countries of the United Nations ratified a set of 17 sustainable development goals to achieve by 2030 that harmonize a unified purpose for national development. These global goals are:

1. end poverty in all its forms everywhere,

[3] Abeba Birhane. "Algorithmic Injustice: A Relational Ethics Approach." In: *Patterns* 2.2 (Feb. 2021), p. 100205. Ezinne Nwankwo and Belona Sonna. "Africa's Social Contract with AI." In: *ACM XRDS Magazine* 26.2 (Winter 2019), pp. 44–48.

[4] Anna Jobin, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." In: *Nature Machine Intelligence* 1 (Sep. 2019), pp. 389–399. Daniel Schiff, Jason Borenstein, Justin Biddle, and Kelly Laas. "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection." In *IEEE Transactions on Technology and Society* 2.1 (Mar. 2021), pp. 31–42.

2. end hunger, achieve food security and improved nutrition and promote sustainable agriculture,
3. ensure healthy lives and promote well-being for all at all ages,
4. ensure inclusive and equitable quality education and promote lifelong learning opportunities for all,
5. achieve gender equality and empower all women and girls,
6. ensure availability and sustainable management of water and sanitation for all,
7. ensure access to affordable, reliable, sustainable and modern energy for all,
8. promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all,
9. build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation,
10. reduce inequality within and among countries,
11. make cities and human settlements inclusive, safe, resilient and sustainable,
12. ensure sustainable consumption and production patterns,
13. take urgent action to combat climate change and its impacts,
14. conserve and sustainably use the oceans, seas and marine resources for sustainable development,
15. protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss,
16. promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels,
17. strengthen the means of implementation and revitalize the global partnership for sustainable development.

Toward satisfying the purpose of government, governmental AI ethics principles are grounded in the kinds of concerns stated in the sustainable development goals. Fairness and justice are a part of many of the goals, including goals five, ten, and sixteen, and also appear as a core tenet of ethics principles. Several other goals relate to social responsibility and beneficence.

Economic growth and productive employment are main aspects of goal eight and play a role in goals nine and twelve. Governments have an overriding fear that machine learning technologies will eliminate jobs through automation without creating others in their place. Therefore, as mentioned in the previous section, the economic direction is played up in governmental AI ethics guidelines and not so much in those of other sectors.

> “Our current trajectory automates work to an excessive degree while refusing to invest in human productivity; further advances will displace workers and fail to create new opportunities (and, in the process, miss out on AI’s full potential to enhance productivity).”
>
> —Daron Acemoglu, economist at Massachusetts Institute of Technology

As part of this goal, there are increasing calls for a paradigm shift towards AI systems that complement or augment human intelligence instead of imitating it.[5]

Furthermore, towards both economic competitiveness and defense from external threats, some countries have now started engaging in a so-called arms race. Viewing the development of machine learning as a race may encourage taking shortcuts in safety and governance, which is cautioned against throughout this book.[6]

15.3 Private Industry

What is the purpose of a corporation? Throughout much of the last fifty years, the stated purpose of corporations (with some exceptions) has been to single-mindedly return profits to investors, also known as maximizing shareholder value.

> "There is one and only one social responsibility of business: to engage in activities designed to increase its profits."
>
> —Milton Friedman, economist at the University of Chicago

In 2019, however, the Business Roundtable, an association of the chief executives of 184 large companies headquartered in the United States, stated a broader purpose for corporations:

1. Delivering value to our customers. We will further the tradition of American companies leading the way in meeting or exceeding customer expectations.
2. Investing in our employees. This starts with compensating them fairly and providing important benefits. It also includes supporting them through training and education that help develop new skills for a rapidly changing world. We foster diversity and inclusion, dignity and respect.
3. Dealing fairly and ethically with our suppliers. We are dedicated to serving as good partners to the other companies, large and small, that help us meet our missions.
4. Supporting the communities in which we work. We respect the people in our communities and protect the environment by embracing sustainable practices across our businesses.
5. Generating long-term value for shareholders, who provide the capital that allows companies to invest, grow and innovate. We are committed to transparency and effective engagement with shareholders.

Shareholder value is listed only in the last item. Other items deal with fairness, transparency and sustainable development. AI ethics principles coming from corporations are congruent with this broadening purpose of the corporation itself, and are also focused on fairness, transparency and sustainable development.[7]

[5] Daron Acemoglu, Michael I. Jordan, and E. Glen Weyl. "The Turing Test is Bad for Business." In: *Wired* (Nov. 2021).

[6] Stephen Cave and Seán S. ÓhÉigeartaigh. "An AI Race for Strategic Advantage: Rhetoric and Risks." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New Orleans, Louisiana, USA, Feb. 2018, pp. 36–40.

[7] In January 2022, the Business Roundtable came out with 10 AI ethics principles of their own: (1) innovate with and for diversity, (2) mitigate the potential for unfair bias, (3) design for and implement transparency, explainability and interpretability,

> "I think we're in the third era, which is the age of integrated impact where we have created social impact that is part of the core value and function of the company overall."
>
> —Erin Reilly, chief social impact officer at Twilio

The 2019 statement by the Business Roundtable is not without criticism. Some argue that it is simply a public relations effort without accompanying actions that could lead to a paradigm change. Others argue it is a way for chief executives to lessen their accountability to investors.[8] AI ethics principles by corporations, especially those by companies developing machine learning technologies, face a similar criticism known as *ethics washing*—creating a façade of developing ethical or responsible machine learning that hides efforts that are actually very shallow.[9] An extreme criticism is that technology companies actively mislead the world about their true purpose and intentions with machine learning.[10]

15.4 Non-Governmental Organizations

NGOs are not homogeneous, but their purpose is usually to advance the political or social goals of their members. The purpose of an individual NGO is captured in its *theory of change*, which could include promoting human rights, improving the welfare of vulnerable groups and individuals, or protecting the environment. As the third sector (civil society), NGOs serve as a watchdog and counterbalance to governments and corporations by taking on roles that neither of the two are able or willing to fulfill. By filling this niche, they lead the criticism of governments and private industry either implicitly or explicitly. Activists in NGOs often try to shift power to the unprivileged.

Critical theory is the study of societal values with the purpose of revealing and challenging power structures; it is the foundation for several NGO theories of change. It includes subfields such as *critical race theory*, *feminism*, *postcolonialism*, and *critical disability theory*. Critical race theory challenges power structures related to race and ethnicity, with a particular focus on white supremacism and racism against blacks in the United States. Feminism is focused on power structures related to gender and challenging male supremacy. Postcolonialism challenges the legacy of (typically European) imperialism that continues to extract human and natural resources for the benefit of colonizers. Critical disability theory challenges ableism. The combinations of these different dimensions and others, known as *intersectionality* (first introduced in Chapter 10), are a key component of critical theory as well.

(4) invest in a future-ready AI workforce, (5) evaluate and monitor model fitness and impact, (6) manage data collection and data use responsibly, (7) design and deploy secure AI systems, (8) encourage a company-wide culture of responsible AI, (9) adapt existing governance structures to account for AI, and (10) operationalize AI governance throughout the whole organization.

[8] Lucian A. Bebchuk and Roberto Tallarita. "The Illusory Promise of Stakeholder Governance." In: *Cornell Law Review* 106 (2020), pp. 91–178.

[9] Elettra Bietti. "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain, Jan. 2020, pp. 210–219.

[10] Mohamed Abdalla and Moustafa Abdalla. "The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Jul. 2021, pp. 287–297.

From the perspective of critical theory, machine learning systems tend to be instruments that reinforce hegemony (power exerted by a dominant group).[11] They extract data from vulnerable groups and at the same time, deliver harm to those same and other vulnerable groups. Therefore, the AI ethics principles coming from civil society often call for a disruption of the entrenched balance of power, particularly by centering the contexts of the most vulnerable and empowering them to pursue their goals.

> "A truly ethical stance on AI requires us to focus on augmentation, localized context and inclusion, three goals that are antithetical to the values justified by late-stage capitalism."
>
> —danah boyd, president of Data & Society Research Institute

As an example, the AI principles stated by an NGO that supports the giving of humanitarian relief to vulnerable populations are the following:[12]

1. weigh the benefits versus the risks: avoid AI if possible,
2. use AI systems that are contextually-based,
3. empower and include local communities in AI initiatives,
4. implement algorithmic auditing systems.

15.5 From Principles to Practice

The ethics principles from government, business, and civil society represent three different paradigms of normative values that may be encoded using the technical aspects of value alignment (described in Chapter 14) to specify the behavior of trustworthy machine learning systems. However, such specification will only happen when there is enough will, incentives, and devoted resources within an organization to make things happen. Intervening on the system's paradigm is an effective starting point, but cannot be the only leverage point that is intervened upon. Putting principles into practice involves several other leverage points as well.

The theory and methods for trustworthy machine learning start from algorithmic research. The incentives for typical machine learning researchers are centered on performance, generalization, efficiency, researcher understanding, novelty, and building on previous work.[13] Since there is now a

[11] Shakir Mohamed, Marie-Therese Png, and William Isaac. "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence" In: *Philosophy & Technology* 33 (Jul. 2020), pp. 659–684. Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. "Towards a Critical Race Methodology in Algorithmic Fairness." In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain, Jan. 2020, pp. 501–512. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Mar. 2021, pp. 610–623.

[12] Jasmine Wright and Andrej Verity. "Artificial Intelligence Principles for Vulnerable Populations in Humanitarian Contexts." Digital Humanitarian Network, Jan. 2020.

[13] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, Michelle Bao. "The Values Encoded in Machine Learning Research." arXiv:2106.15590, 2021.

growing body of research literature on fairness, explainability, and robustness (they are ‘hot topics’), the incentives for researchers are starting to align with the pursuit of research in technical trustworthy machine learning algorithms. Several open-source and commercial software tools have also been created in recent years to make the algorithms from research labs accessible to data scientists. But having algorithmic tools is also only one leverage point for putting ethical AI principles into practice. Practitioners also need the knowhow for affecting change in their organizations and managing various stakeholders. One approach for achieving organizational change is a checklist of harms co-developed with stakeholders.[14] Research is needed to further develop more playbooks for organization change.

Putting principles to practice is a process that has its own lifecycle.[15] The first step is a series of small efforts such as ad hoc risk assessments initiated by *tempered radicals* (people within the organization who believe in the change and continually take small steps toward achieving it). The second step uses the small efforts to demonstrate the importance of trustworthy machine learning and obtain the buy-in of executives to agree to ethics principles. The executives then invest in educating the entire organization on the principles and also start valuing the work of individuals who contribute to trustworthy machine learning practices in their organization. The impetus for executives may also come from external forces such as the news media, brand reputation, third-party audits, and regulations. The third step is the insertion of fact flow tooling (remember this was a way to automatically capture facts for transparency in Chapter 13) and fairness/robustness/explainability algorithms throughout the lifecycle of the common development infrastructure that the organization uses. The fourth step is instituting the requirement that diverse stakeholders be included in problem specification (value alignment) and evaluation of machine learning systems with veto power to modify or stop the deployment of the system. Simultaneously, this fourth step includes the budgeting of resources to pursue trustworthy machine learning in all model development throughout the organization.

15.6 Summary

- The purpose of trustworthy machine learning systems is to do good, but there is no single definition of good.
- Different definitions are expressed in ethics principles from organizations across the government, private, and social sectors.
- Common themes are privacy, fairness, reliability, transparency, and beneficence.
- Governments emphasize the economic implications of the adoption of machine learning.
- Companies stick primarily to the common themes.
- NGOs emphasize the centering and empowerment of vulnerable groups.

[14] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. “Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI.” In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu, Hawaii, USA, Apr. 2020, p. 318.

[15] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. “Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices.” In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (Apr. 2021), p. 7. Kathy Baxter. “AI Ethics Maturity Model.” Sep. 2021.

- A series of small actions can push an organization to adopt AI ethics paradigms and principles. The adoption of principles is an effective start for an organization to adopt trustworthy machine learning as standard practice, but not the only intervention required.
- Going from principles to practice also requires organization-wide education, tooling for trustworthy machine learning throughout the organization's development lifecycle, budgeting of resources to put trustworthy machine learning checks and mitigations into *all* models, and veto power for diverse stakeholders at the problem specification and evaluation stages of the lifecycle.