Approaches to Ethical AI

Erich Prem

Abstract This chapter provides an overview of existing proposals to address ethical issues of AI systems with a focus on ethical frameworks. A large number of such frameworks have been proposed with the aim to ensure the development of AI systems aligned with human values and morals. The frameworks list key ethical values that an AI system should follow. For the most part, they can be regarded as instances of philosophical principlism. This paper provides an overview of such frameworks and their general form and intended way of working. It lists some of the main principles that are proposed in the frameworks and critically assesses the practicality of the various approaches. It also describes current trends, tools, and approaches to ensure the ethicality of AI systems.

1 Introduction

Increasingly, digital systems confront us with machine-based "decisions and actions." Although we should avoid unnecessary anthropomorphization and remind ourselves that such "actions and decisions" are based on human choices and algorithms, they increasingly appear to be those of *machines*.[1] From self-driving cars to systems recommending products or categorizing our creditworthiness, humans have become subject to algorithmic action and decision-making. Such decisions can have significant effects on people's lives including detrimental ones. Just consider the case where a person is denied a loan based on a creditworthiness decision of an algorithm or, perhaps worse, denied a transplant organ as a result of a medical AI system's recommendation. Consequently, scholars and policymakers have started to develop an interest in how to ensure that an AI system's actions and

This chapter includes material that was previously published in Prem (2023).

[1] Many thanks to J. Nida-Rümelin for suggesting this important qualification.

E. Prem (✉)
University of Vienna, Institute of Philosophy, Vienna, Austria
e-mail: erich.prem@univie.ac.at

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_15

decisions do not conflict with human values, laws, or reasonable expectations about how systems *should* behave. Traditionally, these are questions that—when asked about humans—have been addressed in the scholarly discipline of *ethics*.

Ethics—the philosophical study of morality—investigates human behavior in terms of its moral value, which the American moral philosopher Bernard Gert defined as an "informal public system applying to all rational persons, governing behavior that affects others, and includes what are commonly known as the moral rules, ideals, and virtues and has the lessening of evil and harm as its goal (Gert, 2005)." As a scientific discipline, ethics does not necessarily provide clear answers regarding what *should be done* in a specific situation. It may be better to regard it as an effort to understand the consequences and the whole embedding of ethical decision-making. Often, ethical theories state certain principles that we should follow and then study the consequences within those theories and where its boundaries lie.

As an example, consider a so-called consequentialist ethical approach. It focuses on the outcomes of an action and suggests that in deciding upon what to do, we should always take the action that achieves the best outcome. This can be very different from a more rule-based, in philosophical terms *deontological*, perspective where the idea is that the action itself is considered good or bad. This means to state clear rules, similar to laws, that determine the ethical quality of our actions. Consider as an example the classification of actions following the Ten Commandments. Or, thirdly, we might propose that the best approach toward a morally good action is to always act like a *virtuous* person, i.e., similar to someone who has proven to be considerate, benevolent, helpful, friendly, courageous, etc. These are just a few ways of organizing our thinking about morality, and there are many others, e.g., contractualism, intuitionism, emotivism, etc.

The three abovementioned approaches correspond to a consequentialist, a deontological, or a virtue-based ethics. It is easy to see that these three high-level approaches to deciding what to do from a moral perspective will often lead to different choices of actions and, hence, different outcomes. Ethicists often study and debate various ethical theories and their justifications, implications, and shortcomings, but will usually steer away from the question of what should be done. The latter is a question that includes weighing the pros and cons and will often imply the necessity to have a societal debate. What is morally preferable will in many cases also imply a political question about which behavior to support and which actions to put under punishment or social despise.

2 Ethical AI

The idea to develop guidelines that ensure that actions and decisions of digital systems are aligned with our moral values is perhaps even older than the actual existence of such systems. Early science fiction authors have addressed moral decision-making of machines or machine-like creatures, as, for example, Mary

Shelley in her science fiction novel about the artificial creature of Frankenstein. Later, science fiction authors such as Isaac Asimov posed ethical rules for robots, the famous robot laws (Asimov, 1950). For example, Asimov's first law states that a robot may not injure a human being or, through inaction, allow a human being to come to harm. While science fiction authors often present an anthropomorphic image of machines (Weidenfeld, 2022), humans today are becoming the subject of algorithmic (AI-based) decisions. The question of how to ensure that machines take actions that are aligned with human moral values thus has become a center of investigation in AI research and in the philosophy of technology. In parallel, policymakers have started to investigate possible rules and regulations to ensure not only the physical safety of humans when interacting with machines but also that such machines treat people in a morally correct manner.

2.1 Can AI Be Ethical?

This poses the question whether machines can act morally. Note that Bernard Gert's definition above talks about "rational persons." Assuming that robots are not included in the category of persons, it is needed to replace this term with something like "machines that give the impression to deliberate reasons."[2] Otherwise, it becomes necessary to argue that AI algorithms should be regarded as rational persons. Also, Gert considers morality as an informal system. We may therefore expect non-trivial challenges when formalizing morality for the sake of implementation on a computing device. For many people, morality also connotes an element of conscious consideration and conscience. However, for all we know, no machine feels ashamed for its possible wrongdoing, nor does a potentially bad outcome lead to a machine's bad conscience. In addition, robotic devices cannot normally become the subject of legal procedures. All of this renders the term "ethical AI" philosophically problematic, and there is an ongoing debate about the degree to which machines can or indeed should be considered ethical agents (cf. Cave et al., 2019). To simplify the issue for our purpose here, we take the notion of *ethical AI* simply as an abbreviation for an AI system that performs actions that when taken by a human would be considered ethical.[3]

[2] Again, thanks to J. Nida-Rümelin for insisting on precision in this formulation.

[3] Note that in many situations, it may be preferable to generalize to an *ethical artifact* as the clear definition of AI remains a challenge.

Table 1 Example for the main components of an ethical AI framework for the principle of "fairness"

Concept	Bias
Concern	Not treating people fairly, e.g., taking decisions that are influenced by a person's gender or social background
Principle	**Fairness**
Remedy	Testing systems for bias and using unbiased data sets, etc.

3 Approaches to Ethical AI

Let us now take a closer look at various efforts to ensure that AI systems make ethical decisions. One specific approach to safeguarding ethical decision-making in machines that many scholars investigated is the creation of so-called AI frameworks. Such frameworks provide a set of principles that an AI system should follow to ensure that its actions or decisions are ethical. There are, of course, other approaches. For example, the European Commission has proposed *regulation* to create legal boundaries for AI systems (EC AI, 2021). There are also *standards* about how to address ethical concerns during system design (IEEE, 2021), and there are proposals for *labels* to inform us about the qualities of an AI system (Mitchell et al., 2019). (See also the chapter by Neppel on "Governance for Digital Humanism.") The next section looks at various frameworks for ethical AI.

3.1 Ethical AI Frameworks

Typically, ethical AI frameworks consist of *concepts, concerns, principles,* and *remedies* see Table 1. Concepts are specific notions to describe the ethical issues or potential shortcoming (the concern). For example, the concept of *bias* is used to explain a specific *concern* about AI classifiers. The potential (ethical) shortcoming is that they may take unfair, and, hence, unethical, decisions. *Principles* are used to describe desirable properties of an AI system or its actions and decisions. For example, the *fairness principle* could be used to demand that AI systems should not discriminate against people of different gender or social background. *Remedies* can take many forms, e.g., recommendations about how to ensure that an AI system fulfills a given principle. Note that concepts, concerns, and principles are not always clearly separated. For example, *fairness* can appear as a concept and a principle.

An early ethical framework in computing was proposed by Richard O. Mason in the context of the "information age" (Mason, 1986). With the aim that IT should help to "enhance the dignity of mankind," he suggested that an AI system should fulfill four ethical principles, namely, *privacy, accuracy, property,* and *accessibility.* Mason suggested that IT systems should not unduly invade people's privacy, be accurate in what they are doing, respect intellectual property rights, and be as

accessible as possible. In many cases, these principles will still be relevant today. Modern ethical frameworks for AI, such as those expressed in the European Commission White Paper *Ethics Guidelines for Trustworthy AI*, may include more principles including some that are especially relevant for AI systems (EC, 2019). In particular, the EC High-Level Expert Group proposed the following principles:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination, and fairness
- Societal and environmental well-being
- Accountability

The White Paper also includes a *checklist* (see below) for practical use by companies to evaluate the ethicality of the systems they may develop. The White Paper is just one example of several proposed ethical frameworks. In fact, so many frameworks have been proposed that they have become the subject of systematic analysis (Floridi & Cowls, 2021). These analyses often conclude that there are strong similarities such as common principles between the different frameworks. For example, Jobin et al. (2019) argue that many guidelines focus on *transparency, fairness, non-maleficence, responsibility,* and *privacy;* Floridi and Cowls (2021) list *beneficence, non-maleficence, autonomy, justice,* and *explicability* as common themes (see below regarding their origin in bio-ethics).

Some important principles recurring in various frameworks in one form or another and their possible interpretations for ethical AI systems include the following:

– *Beneficence:* The principle of doing good to others, of mercy, and of kindness. For an AI system, it may imply to ensure people's well-being and support sustainable development and inclusive growth. It could also include the protection of fundamental rights.
– *Non-maleficence* states that there is an obligation not to inflict harm on others. Often, this is formulated as "first do no harm." Obviously, also an AI system should prevent harm. Note that non-maleficence and beneficence are not the same as *beneficence* prompts to take an action, while *non-maleficence* may often prompt not to take an action. Beneficence is sometimes considered secondary to non-maleficence (hence, "*first* do no harm").
– *Autonomy:* The principle states that humans should be granted the right to decide on their own. This entails being informed and free to decide. For an AI system, it means to ensure and further the ability of humans to make decisions on their own. It is often interpreted to also mean human agency and oversight by humans.
– *Justice:* As a principle, justice means fairness in decisions, but also accessibility without unfair discrimination. It can entail further aspects such as the availability of redress. In AI, a system should be compatible with what is considered fair.

– *Explicability:* For an AI system, this means that its actions should be understandable for humans. It may include traceability (the ability to verify the history or logical chain) and interpretability (of results).

Some of these principles (e.g., beneficence or autonomy) are not specific to AI systems. Rather, they are often also applied in non-engineering scenarios and are used in ethical frameworks of research institutes or medical institutions. The principles of non-maleficence and beneficence are central to many types of ethics. The principle of autonomy is a key component in guidelines for scientific experiments with human subjects and to medical decision-making in general. Only a few ethical principles are specific to computational systems or are of specific meaning and importance in computational contexts. These include the following:

- *Explicability* refers to the principle that decisions of an AI system should be explainable and understandable for humans, i.e., especially for the subject of an AI-based decision. This could mean, for example, to provide reasons why a bank's classification system for creditworthiness excludes a person as a reliable borrower for a loan. Similarly, an AI-based x-ray system should provide reasons why it categorizes a specific image as that of a patient with cancer. Given that many AI systems tune thousands of parameters using large amounts of data and statistical algorithms, such explanations have often proven very difficult to provide. In addition, it is not easy to explain what precisely constitutes a valid and truthful explanation (see below).
- *Privacy* concerns the fact that many algorithmic systems including AI systems are extremely data-hungry and therefore may require or carry large amounts of personal information. This may also include information that should be particularly well safeguarded, i.e., sensitive data such as personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, genetic data, biometric data processed solely to identify a human being, health-related data, and data concerning a person's sex life or sexual orientation (cf. Art. 4 and 9 of the GDPR) (GDPR, n.d.). As a principle, AI systems should always protect a person's privacy and never store or disseminate specifically protected sensitive data.
- *Fairness* (in the sense of being "unbiased") addresses the fact that AI systems can easily become biased in their decisions. For example, systems trained to assess the later success of an applicant for a university may be biased against the person because of bad training procedures or because of bias already present in the training data (e.g., because an institution may have historically accepted fewer women than men and this fact is represented in historic data).

The relations between the different principles are not trivial. It is, for example, not very clear that privacy is a separate principle as it could also be described as following from non-maleficence. Similarly, it could be argued that explicability really follows from autonomy. This is one of the reasons why, at least superficially, many different proposals for ethical frameworks exist as they may group principles differently and have different numbers of principles.

3.2 Philosophical Principlism

From a philosophical perspective, ethical AI frameworks can be considered instances of "principlism." Principlism is a useful approach to support ethical decision-making in practical situations (usually, moral dilemmas). It is often used in medicine and other fields of science as it often facilitates relatively clear decisions based on only a few principles. In the late twentieth century, ethical principles emerged in reaction to medical experiments such as Albert Neisser's experiments in which patients were infected with syphilis without their consent. Later, the horrendous Nazi experiments of no or questionable scientific value on Jews and other prison inmates led to the Declaration of Helsinki (WMA, 2013). Finally, the Tuskegee syphilis study by the US Public Health Service and CDC on 400 African Americans led to the introduction of ethical principles for medical experiments (Beauchamp & Childress, 1979). However, principlism in the medical profession is much older. It is often dated back to the Hippocratic Oath that goes back to AD 245. There are several modern versions of the oath such as the currently relevant Declaration of Geneva[4] used in the medical profession:

> [. . .] I SOLEMNLY PLEDGE to dedicate my life to the service of humanity;
> THE HEALTH AND WELL-BEING OF MY PATIENT will be my first consideration;
> I WILL RESPECT the autonomy and dignity of my patient;
> I WILL MAINTAIN the utmost respect for human life;
> I WILL NOT PERMIT considerations of age, disease or disability, creed, ethnic origin, gender, nationality, political affiliation, race, sexual orientation, social standing or any other factor to intervene between my duty and my patient; [. . .]

This excerpt clearly includes reference to principles like beneficence, non-maleficence, autonomy, and even fairness. While it is positive that principlism seems to work in some professions, it poses the question of whether the principles are sufficiently concrete and tuned for the needs of AI systems and their designers.

3.3 Challenges and Limitations of Ethical Frameworks

Despite its often-intuitive appearance, principlism suffers from a range of challenges and limitations when trying to put it to practice. This of course also impairs the use of ethical frameworks for AI. Firstly, principles are usually formulated without any application context. Indeed, principles require a high degree of abstraction, or otherwise, they lose their character of being a principle. The lack of context means that a principle is often not very helpful or only seemingly clear. Just like the interpretation of "good weather" may depend on whether you are a farmer or a tourist, the question of whether you should tell the truth may depend on the subject matter, a person's situation, age, level of understanding, etc. Although the principle

[4] https://www.wma.net/policies-post/wma-declaration-of-geneva/.

of autonomy says that people should be supported in making their own decisions, this does not apply in medical contexts where individuals might lack decisional capacity, for example, because they are too young, mentally ill, or unconscious. The field of medical ethics has therefore developed a set of practices as well as procedures and structures (e.g., ethics reviews and boards) to deal with these contextual aspects. Medicine has also developed prototypical situations and standardized approaches over time.

Secondly, principles are usually listed without a clear prioritization among them. This makes them susceptible to conflict. As an example, the principle of beneficence may conflict with privacy when a radiological AI system for detecting cancer may require a complete history of sexually transmitted diseases. While this may increase precision, it poses ethical questions including whether it is reasonable to assume completely truthful answers from a patient. Similarly, some techniques for improving the explicability of a neural network model may reduce accuracy (and, hence, conflict with the principle of beneficence) when the tools for explaining a system tend to interfere with its prediction or classification quality. In medical contexts, beneficence can easily clash with autonomy when patients decide against what seems medically beneficial given their value preferences.

Thirdly, some principles are only superficially clear, but it may be very difficult to agree on what they mean precisely and, therefore, when they are fulfilled. For example, the idea of making AI systems explain how they arrived at a prediction sounds reasonable. However, *explicability* is very difficult to specify with precision. We may use the concept of *understanding* in demanding that an AI system's decisions should be understandable for users. But it is not trivial to precisely state what understanding really means. Which type of explanation achieves proper understanding and what would be a test for a person to ensure they have really understood what is going on in a neural network or why a certain decision has been made? For example, how a neural network arrived at its output can be provided in mathematical form. However, this would hardly constitute an explanation for a human who may be better served with an explanation that involves more easily accessible concepts. Sometimes, a counterfactual explanation can be useful, for example, when explaining that the output of an AI system would have been different if only the input had taken a different form or value. Note that this is not just a terminological or conceptual imprecision. It is indeed a philosophical challenge to define the notion of understanding beyond a mere psychological feeling.

In addition, principles alone are usually insufficient to clearly decide which system design steps to take. A medical system proposing measures to an overweight patient's benefit may include anything from exercise to a dietary plan, lifestyle changes, or surgery. All these measures may be judged as beneficial, but it may remain unclear what is the best action to choose. Although strictly speaking, this is perhaps not an ethical problem (as all the actions may be beneficial), it is a practical problem from the point of view of designing the AI system.

Similarly, the principle of *fairness* comes with many challenges including philosophical ones. Again, in many cases, there will be a lack of clarity of what we precisely mean when demanding an AI system's fairness. Consider the case of a

Table 2 Different fairness concepts (metrics), optimization criteria (equalizing for), and examples. Adapted from a longer list in Seng Ah Lee et al. (2021). FP, (number of) false positives or lost opportunity as they are predicted to default when they really would have repaid the loan; FN, false negatives; TP, true positives; TN, true negatives

Fairness metric	Equalizing	Intuition/example
Maximize total accuracy	None	Most accurate model gives people the loan and interest they "deserve" by minimizing errors
Equal opportunity	False-negative rate FN/(FN + TP)	Among *creditworthy* applications, men and women have similar approval rates
Predictive equality	False-positive rate FP/(FP + TN)	Among *defaulting* applicants, men and women have similar rates of *denied* loans
Equal odds	True-positive rate TP/(TP + FN), true-negative rate TN/(FP + TN), positive predictive value TP/(TP + FP)	Both of the above: among creditworthy applicants, probability of predicting repayment is the same regardless of gender
Counterfactual fairness	Prediction in counterfactual scenario	For each individual: if they were a different gender, the prediction would be the same
Individual fairness	Outcome for "similar" individuals	Each individual has the same outcome as another "similar" individual of a different gender

creditworthiness expert system that may have been trained on historic data and therefore is biased in treating men and women differently. What should "fairness" mean in this context, or what precisely does it mean to treat men and women *equally*? It could mean, for example, that the outputs of the system do not change when we change the gender of a person in the input data ("counterfactual fairness"). It could also mean an equal average creditworthy rating for men and women. Or it could mean an equal chance of being denied a loan for both genders, etc. These interpretations (or definitions) of fairness will represent different mathematical functions, as indicated in the table below (cf. Seng Ah Lee et al., 2021). The table lists six variants; there are, however, many more plausible interpretations of fairness including those that discuss continuous functions rather than only binary categorization (Table 2).

These fairness notions also correspond to various philosophical approaches as proposed in the literature; see Seng Ah Lee et al. (2021) for a list. In everyday life, such questions are often political and/or decided through social debate and practical norms (where fairness becomes justice). They are less "mathematical" in their nature than they are social and societal. Also, some approaches to AI ethics aim at societal change and go beyond or correct what may be current practice, which additionally complicates the situation (e.g., affirmative action). This means, we cannot just give a general rule or mathematical function that defines "fairness" free from an application context. It is very well possible that we consider fairness for the case of granting a loan differently from providing subsidies to the poor or taking decisions regarding

permissible insurance premiums. In some situations, it may prove preferable to aim at "equal odds," while in other situations, it could be considered better to optimize for "individual fairness." In AI systems, the problem is exacerbated because every decision-making system will at least implicitly define a "fairness" function (assuming it produces a proper mathematical function). Hence, the question of what we mean by fairness is in the end inescapable, and an AI system designer will always, albeit sometimes only implicitly, define fairness when building a system.

In addition to these concerns, there is a question regarding the human-centeredness of many AI frameworks. It is debatable whether a focus on humans *alone* is sufficient and to what degree principles should also include environmental sustainability or the welfare of animals. Modern ethics has developed in various directions, and there seems to be an addition of topics that ethics should include. As an example, consider the case of "land ethics" (Leopold, 1949). This question concerns an important debate for Digital Humanism as a whole. Despite its name, much of current Digital Humanism goes in fact beyond a purely *human* focus in that the environment, our climate, and the welfare of sentient beings are being discussed by scholars in Digital Humanism.

The challenges listed above have led to a more general criticism of ethical AI frameworks. Some authors have questioned that frameworks can solve the problem, including because the principles were "meaningless," "isolated," *and* "toothless" and because of the gap between "high-minded principles and technological practice" (Munn, 2023).

4 From Principles to Practice

Principles for AI can be employed during the design of the systems and also during the operation so as to ensure that AI systems act in line with ethical principles. However, how to achieve the latter in practice is far from trivial. Technical approaches to realizing ethical AI systems vary widely. In the following, we provide a short overview of selected approaches that have been suggested in practice. For a more complete list, cf. Prem (2023).

Design Phase

- *Checklists* are a straightforward approach to designing ethical AI systems. They are a proven tool in engineering and systems operation (e.g., when operating an airplane). Checklists help make sure that procedures are being followed, that nothing is forgotten, and that certain conditions are met. For example, ethical checklists can be applied to criteria for training data and training processes of AI models or to ensure that all aspects of an ethical framework were considered. Montague et al. (2021) describe a data ethics checklist, for example.

- *Case studies*, *good practice examples*, or *prototypes* can be an efficient way to improve ethical characteristics of a system based on previous experience or existing systems that exhibit the desired ethical characteristics.
- *Process models* aim at guiding a design process in a way that guarantees that ethical concerns are appropriately addressed. Such process models can be recommendations and focus on certain steps or include standardized procedures for certain aspects (e.g., Eitel-Porter, 2021). (See also the chapter by Zuber et al. on "Value-Sensitive Software Design" in this volume.)
- *Data sets* are a type of infrastructure that can be used for training or testing of AI models. Standardized data models can help overcome bias and support the testing or evaluation of the quality of an AI model. For example, the *Equity Evaluation Corpus* consists of more than 8000 sentences "chosen to tease out biases towards certain races and genders" (Kiritchenko & Mohammad, 2018).
- *Algorithms* or *libraries* that help address ethical issues currently are the focus of a large number of research and development activities (cf. Prem, 2023). In fact, so many researchers aim to develop algorithms for privacy-preserving machine learning that these areas could be called a subfield of machine learning. Similarly, developing methods to enable or improve the explicability of AI systems, especially those trained with deep learning, has developed into the subdiscipline of "explainable AI" (or XAI for short). *Software libraries* have the added advantage of being already coded algorithms designed to address ethical issues. Examples include libraries for explainable AI[5] or for measuring systems according to fairness metrics (Wexler et al., 2020).

System Operation

- *Declarations* are statements that assert features of AI systems, typically to their users. This could mean to describe how an algorithm works, which type of training data was used, which fairness or bias concerns were considered, etc. Such declarations may follow formal requirements, e.g., "labels," to facilitate comparisons between systems. Declarations address ethical concerns often indirectly in stating ethical issues explicitly but without necessarily solving them in the system. While declarations are a useful source of information, the choice of appropriate action is often left to the user. Hence, declarations tend to delegate the responsibility for the ethical issue in question to the user.

Ex-Post Approaches

- *Audits* can serve to examine a system after its design and implementation. It can help assess ethical issues after the system was put in operation.

[5] https://github.com/EthicalML/xai.

Other approaches to handling ethical issues of AI systems include *training (education), license models, metrics, design patterns, online communities,* and *codes of practice.*

4.1 Further Research Directions

In summary, ethical frameworks can hardly be considered a solution to the challenge of creating ethical AI systems. They can provide guidance on better understanding many of the issues involved in their design and at times provide guidance and orientation for the design and implementation process. The biggest challenge with ethical AI frameworks today is their lack of providing clear advice on how to build AI systems. An important area of future research therefore concerns technical means to realize ethical AI systems. In addition, the context and evolution of the embedding of an AI application will require much more study in the future. It is therefore likely that future research will address AI systems in specific situations, i.e., within their respective application contexts. Such systems will then have to be analyzed with respect to their behavior and how they are perceived by humans interacting with the system. They need to be reviewed and critically debated so that with time, a practice of ethical AI systems emerges that can help train generations of AI system developers.

5 Conclusions

Principles have played an important role in various science and technology fields. They are current standard practice in various areas, for example, in research where principles are used to decide upon the conditions under which experiments with humans should be performed. They are also used to guide decisions regarding medical treatments, where they inform about priorities such as in the case of medical transplants and guide information provided to patients or subjects of medical experiments.

For the design of ethical AI systems, a large number of ethical frameworks using principles have been designed. The list of principles and the way in which they are formulated show great similarities and often overlap with the ethical principles used in medicine. Several principles included in these frameworks are very general (e.g., non-maleficence), and only a few are specific to AI (e.g., explicability). Frameworks and ethical principles are usually detached from implementation questions. They guide *what* AI and other algorithmic systems should or should not do, but do not explain *how* to achieve this in practical systems. In real-world situations, ethical principles may contradict each other and require prioritization. However, many frameworks for ethical AI do not include a clear order in which the principles should be applied. A prioritization or other decision regarding the principles then may

require careful weighing of the principles against each other, consideration, or debate.[6]

The development of ethical AI systems is at an early stage. To make the principles work in practice, more debate, research, and perhaps clearer rules are required, including regarding concrete application contexts. Tools and techniques to help realize ethical AI systems include checklists, case studies, prototypes, process models, standards, data sets, algorithms, software libraries, declarations, audits, and others.

Discussion Questions for Students and Their Teachers

1. Consider one of the ethical frameworks described above. Show that the order in which ethical principles are applied may change the outcome of an ethical consideration based on ethical frameworks.
2. Discuss the nature of ethical frameworks from an engineering perspective: should they be considered a component of the system specification, an element of the system design process, or part of the resulting AI system? (See also the chapter by Ghezzi in this volume.)
3. Consider an AI model that has a known bias, for example, it may work better for men than for women in a medical diagnostic task. What might cause such a situation? Do you think such situations are completely avoidable, or are there situations where we may have to accept a biased system? Which of the tools described in this chapter could you use to remedy the situation? What are the pros and cons?
4. Which tools or techniques could you use to ensure an AI system's fairness during design, implementation (training), and operation?

Learning Resources for Students

1. Floridi L., Cowls J. (2021) A unified framework of five principles for AI in society.

   This is a good place to get an overview of various frameworks. It also develops a unified version of a framework that extracts common principles from the other frameworks.
2. Prem E. (2023) From Ethical AI Frameworks to Tools: A review of approaches. *AI and Ethics.*

   This is an overview from the perspective of approaches to implementing ethical AI. It collects various tools, standards, declarations, etc. that are proposed in the literature to address various ethical issues. As mentioned above, Munn (2023) presents a critical perspective on frameworks.
3. Stahl B.C., Schroeder D., Rodrigues R. (2023). The Ethics of Artificial Intelligence: An Introduction. In: Ethics of Artificial Intelligence, Springer.

   This is a modern introduction to the ethics of AI. It provides an up-to-date overview and case studies to introduce AI and address some of the key ethical

[6] D. Ross (1930) discussed the dilemmas arising from conflicting principles and a plurality of prima facie duties.

principles such as privacy, unfair discrimination, or the right to life and liberty of persons. It also includes some more general aspects not addressed in this chapter, e.g., surveillance capitalism.

Acknowledgment This research was supported by the "Entrepreneurs as role models" project at the University of Vienna.

References

Asimov, I. (1950). In The Isaac Asimov Collection (Ed.), *Runaround. I, Robot* (p. 40). Doubleday.

Beauchamp, T., & Childress, J. (1979). *Principles of biomedical ethics*. Oxford University Press.

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and risks of machine ethics. *Proceedings of the IEEE, 107*(3), 562–574. https://doi.org/10.1109/JPROC.2018.2865996

EC. (2019). Ethics guidelines for trustworthy AI. Directorate-General for Communications Networks, Content and Technology, EC Publications Office. https://data.europa.eu. https://doi.org/10.2759/177365.

EC AI. (2021). European Commission, Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative act. COM (2021) 206 final.

Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI Ethics, 1*, 73–80. https://doi.org/10.1007/s43681-020-00011-6

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In L. Floridi (Ed.), *Ethics, governance, and policies in artificial intelligence* (Philosophical studies series) (Vol. 144, pp. 5–6). Springer. https://doi.org/10.1007/978-3-030-81907-1_2

GDPR, General Data Protection Regulation. (n.d.) Regulation (EU) 2016/679 of the European Parliament and of the Council. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e40-1-1

Gert, B. (2005). *Morality: Its nature and justification* (Revised ed.). Oxford University Press.

IEEE. (2021). IEEE standard model process for addressing ethical concerns during system design. *IEEE Standards*, 7000–2021. https://doi.org/10.1109/IEEESTD.2021.9536679

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial intelligence: The global landscape of ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kiritchenko S., & Mohammad S. F. (2018). Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of *Sem*, New Orleans, LA, USA.

Leopold, A. (1949). *A Sand County Almanac* (p. 203). Oxford University Press.

Mason, R. O. (1986). Four ethical issues of the information age. *MIS Quarterly, 10*(1), 5–12.

Mitchell M., Wu S., Zaldivar A., Barnes P., Vasserman L., Hutchinson B., & Gebru T. (2019). Model cards for model reporting. Proc. Conf. Fairness, Account. Transpar. FAT*19. https://doi.org/10.1145/3287560.3287596.

Montague, E., Eugene, D. T., Barry, D., et al. (2021). The case for information fiduciaries: The implementation of a data ethics checklist at Seattle children's hospital. *Journal of the American Medical Informatics Association, 28*(3), 650–652. https://doi.org/10.1093/jamia/ocaa307

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00209-w

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI and Ethics.*. https://link.springer.com/content/pdf. https://doi.org/10.1007/s43681-023-00258-9.pdf

Ross, D. (1930). *The right and the good*. Clarendon Press.

Seng Ah Lee M., Floridi L., & Singh J. (2021). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. https://ssrn.com/abstract=3679975.

Stahl, B. C., Schroeder, D., & Rodrigues, R. (2023). The ethics of artificial intelligence: An introduction. In *Ethics of artificial intelligence. SpringerBriefs in research and innovation governance*. Springer. https://doi.org/10.1007/978-3-031-17040-9_1

Weidenfeld, N. (2022). Fictionalizing the robot and artificial intelligence. In H. Werthner, E. Prem, E. A. Lee, & C. Ghezzi (Eds.), *Perspectives on digital humanism*. Springer. https://doi.org/10.1007/978-3-030-86144-5_14

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2020). The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

WMA. (2013). World medical association, declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of the American Medical Association, 310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053