# Chapter 1
# The Ethics of Artificial Intelligence: An Introduction


**Abstract** This chapter introduces the themes covered by the book. It provides an overview of the concept of artificial intelligence (AI) and some of the technologies that have contributed to the current high level of visibility of AI. It explains why using case studies is a suitable approach to engage a broader audience with an interest in AI ethics. The chapter provides a brief overview of the structure and logic of the book by indicating the content of the cases covered in each section. It concludes by identifying the concept of ethics used in this book and how it is located in the broader discussion of ethics, human rights and regulation of AI.

**Keywords** Artificial intelligence · Machine learning · Deep learning ethics

The ethical challenges presented by artificial intelligence (AI) are one of the biggest topics of the twenty-first century. The potential benefits of AI are said to be numerous, ranging from operational improvements, such as the reduction of human error (e.g. in medical diagnosis), to the use of robots in hazardous situations (e.g. to secure a nuclear plant after an accident). At the same time, AI raises many ethical concerns, ranging from algorithmic bias and the digital divide to serious health and safety concerns.

The field of AI ethics has boomed into a global enterprise with a wide variety of players. Yet the ethics of artificial intelligence (AI) is nothing new. The concept of AI is almost 70 years old (McCarthy et al. 2006) and ethical concerns about AI have been raised since the middle of the twentieth century (Wiener 1954; Dreyfus 1972; Weizenbaum 1977). The debate has now gained tremendous speed thanks to wider concerns about the use and impact of better algorithms, the growing availability of computing resources and the increasing amounts of data that can be used for analysis (Hall and Pesenti 2017).

These technical developments have favoured specific types of AI, in particular machine learning (Alpaydin 2020; Faggella 2020), of which deep learning is one popular form (see box) (LeCun et al. 2015). The success of these AI approaches led to a rapidly expanding set of uses and applications which frequently resulted



B. C. Stahl et al., *Ethics of Artificial Intelligence*,
SpringerBriefs in Research and Innovation Governance,
https://doi.org/10.1007/978-3-031-17040-9_1

in consequences that were deemed ethically problematic, such as unfair or illegal discrimination, exclusion and political interference.

**Deep Learning**

Deep learning is one of the approaches to machine learning that have led to the remarkable successes of AI in recent years (Bengio et al. 2021). The development of deep learning is a result of the use of artificial neural networks, which are attempts to replicate or simulate brain functions. Natural intelligence arises from parallel networks of neurons that learn by adjusting the strengths of their connections. Deep learning attempts to perform brain-like activities using statistical measures to determine how well a network is performing. Deep learning derives its name from deep neural networks, i.e. networks with many layers. It has been successfully applied to problems ranging from image recognition to natural speech processing. Despite its successes, deep learning has to contend with a range of limitations (Cremer 2021). It is open to debate how much further machine learning based on approaches like deep learning can progress and whether fundamentally different principles might be required, such as the introduction of causality models (Schölkopf et al. 2021).

With new uses of AI, AI ethics has flourished well beyond academia. For instance, the Rome Call for AI Ethics,[1] launched in February 2020, links the Vatican with the UN Food and Agriculture Organization (FAO), Microsoft, IBM and the Italian Ministry of Innovation. Another example is that UNESCO appointed 24 experts from around the world in July 2021 and launched a worldwide online consultation on AI ethics and facilitated dialogue with all UNESCO member states. Media interest is also considerable, although some academics consider the treatment of AI ethics by the media as "shallow" (Ouchchy et al. 2020).

One of the big problems that AI ethics and ethicists might face is the opaqueness of what is actually happening in AI, given that a good grasp of an activity itself is very helpful in determining its ethical issues.

> [I]t is not the role nor to be expected of an AI Ethicist to be able to program the systems themselves. Instead, a strong understanding of aspects such as the difference between supervised and unsupervised learning, what it means to label a dataset, how consent of the user is obtained – essentially, how a system is designed, developed, and deployed – is necessary. In other words, an AI Ethicist must comprehend enough to be able to apprehend the instances in which key ethical questions must be answered (Gambelin 2021).

There is thus an expectation that AI ethicists are familiar with the technology, yet "[n]o one really knows how the most advanced algorithms do what they do" (Knight 2017), including AI developers themselves.

Despite this opacity of AI in its current forms, it is important to reflect on and discuss which ethical issues can arise due to its development and use. The approach to AI ethics we have chosen here is to use case studies, as "[r]eal experiences in AI ethics present … nuanced examples" (Brusseau 2021) for discussion, learning and

[1] https://www.romecall.org/.

analysis. This approach will enable us to illustrate the main ethical challenges of AI, often with reference to human rights (Franks 2017).

Case studies are a proven method for increasing insights into theoretical concepts by illustrating them through real-world situations (Escartín et al. 2015). They also increase student participation and enhance the learning experience (ibid) and are therefore well-suited to teaching (Yin 2003).

We have therefore chosen the case study method for this book. We selected the most significant or pertinent ethical issues that are currently discussed in the context of AI (based on and updated from Andreou et al. 2019 and other sources) and dedicated one chapter to each of them.

The structure of each chapter is as follows. First, we introduce short real-life case vignettes to give an overview of a particular ethical issue. Second, we present a narrative assessment of the vignettes and the broader context. Third, we suggest ways in which these ethical issues could be addressed. This often takes the form of an overview of the tools available to reduce the ethical risks of the particular case; for instance, a case study of algorithmic bias leading to discrimination will be accompanied by an explanation of the purpose and scope of AI impact assessments. Where tools are not appropriate, as human decisions need to be made based on ethical reasoning (e.g. in the case of sex robots), we provide a synthesis of different argument strategies. Our focus is on *real-life* scenarios, most of which have already been published by the media or research outlets. Below we present a short overview of the cases.

*Unfair and Illegal Discrimination* (Chap. 2)

The first vignette deals with the automated shortlisting of job candidates by an AI tool trained with CVs (résumés) from the previous ten years. Notwithstanding efforts to address early difficulties with gender bias, the company eventually abandoned the approach as it was not compatible with their commitment to workplace diversity and equality.

The second vignette describes how parole was denied to a prisoner with a model rehabilitation record based on the risk-to-society predictions of an AI system. It became clear that subjective personal views given by prison guards, who may have been influenced by racial prejudices, led to an unreasonably high risk score.

The third vignette tells the story of an engineering student of Asian descent whose passport photo was rejected by New Zealand government systems because his eyes were allegedly closed. This was an ethnicity-based error in passport photo recognition, which was also made by similar systems elsewhere, affecting, for example, dark-skinned women in the UK.

*Privacy* (Chap. 3)

The first vignette is about the Chinese social credit scoring system, which uses a large number of data points to calculate a score of citizens' trustworthiness. High scores lead to the allocation of benefits, whereas low scores can result in the withdrawal of services.

The second vignette covers the Saudi Human Genome Program, with predicted benefits in the form of medical breakthroughs versus genetic privacy concerns.

*Surveillance Capitalism* (Chap. 4)

The first vignette deals with photo harvesting from services such as Instagram, LinkedIn and YouTube in contravention of what users of these services were likely to expect or have agreed to. The relevant AI software company, which specialises in facial recognition software, reportedly holds ten billion facial images from around the world.

The second vignette is about a data leak from a provider of health tracking services, which made the health data of 61 million people publicly available.

The third vignette summarises Italian legal proceedings against Facebook for misleading its users by not explaining to them in a timely and adequate manner, during the activation of their account, that data would be collected with commercial intent.

*Manipulation* (Chap. 5)

The first vignette covers the Facebook and Cambridge Analytica scandal, which allowed Cambridge Analytica to harvest 50 million Facebook profiles, enabling the delivery of personalised messages to the profile holders and a wider analysis of voter behaviour in the run-up to the 2016 US presidential election and the Brexit referendum in the same year.

The second vignette shows how research is used to push commercial products to potential buyers at specifically determined vulnerable moments, e.g. beauty products being promoted at times when recipients of online commercials are likely to feel least attractive.

*Right to Life, Liberty and Security of Person* (Chap. 6)

The first vignette is about the well-known crash of a Tesla self-driving car, killing the person inside.

The second vignette summarises the security vulnerabilities of smart home hubs, which can lead to man-in-the-middle attacks, a type of cyberattack in which the security of a system is compromised, allowing an attacker to eavesdrop on confidential information.

The third vignette deals with adversarial attacks in medical diagnosis, in which an AI-trained system could be fooled to the extent of almost 70% with fake images.

*Dignity* (Chap. 7)

The first vignette describes the case of an employee who was wrongly dismissed and escorted off his company's premises by security guards, with implications for his dignity. The dismissal decision was based on opaque decision-making by an AI tool, communicated by an automatic system.

The second vignette covers sex robots, in particular whether they are an affront to the dignity of women and female children.

Similarly, the third vignette asks whether care robots are an affront to the dignity of elderly people.

*AI for Good and the UN's Sustainable Development Goals* (Chap. 8)

The first vignette shows how seasonal climate forecasting in resource-limited settings has led to the denial of credits for poor farmers in Zimbabwe and Brazil and the accelerated the layoff of workers in the fishing industry in Peru.

The second vignette deals with a research team from a high-income country requesting vast amounts of mobile phone data from users in Sierra Leone, Guinea and Liberia to track population movements during the Ebola crisis. Commentators argued that the time spent negotiating the request with seriously under-resourced governance structures should have been used to handle the escalating Ebola crisis.

This is a book of AI ethics case studies and not a philosophical book on ethics. We nevertheless need to be clear about our use of the term "ethics". We use the concept of ethics cognisant of the venerable tradition of ethical discussion and of key positions such as those based on an evaluation of the duty of an ethical agent (Kant 1788, 1797), the consequences of an action (Bentham 1789; Mill 1861), the character of the agent (Aristotle 2000) and the keen observation of potential biases in one's own position, for instance through using an ethics of care (Held 2005). We slightly favour a Kantian position in several chapters, but use and acknowledge others. We recognize that there are many other ethical traditions beyond the dominant European ones mentioned here, and we welcome debate about how these may help us understand further aspects of ethics and technology. We thus use the term "ethics" in a pluralistic sense.

This approach is pluralistic because it is open to interpretations from the perspective of the main ethical theories as well as other theoretical positions, including more recent attempts to develop ethical theories that are geared more specifically to novel technologies, such as disclosive ethics (Brey 2000), computer ethics (Bynum 2001), information ethics (Floridi 1999) and human flourishing (Stahl 2021).

Our pluralistic reading of the ethics of AI is consistent with much of the relevant literature. A predominant approach to AI ethics is the development of guidelines (Jobin et al. 2019), most of which are based on mid-level ethical principles typically developed from the principles of biomedical ethics (Childress and Beauchamp 1979). This is also the approach adopted by the European Union's High-Level Expert Group on AI (AI HLEG 2019). The HLEG's intervention has been influential, as it has had a great impact on the discussion in Europe, which is where we are physically located and which is the origin of the funding for our work (see Acknowledgements). However, there has been significant criticism of the approach to AI ethics based on ethical principles and guidelines (Mittelstadt 2019; Rességuier and Rodrigues 2020). One key concern is that it remains far from the application and does not explain how AI ethics can be put into practice. With the case-study-based approach presented in this book, we aim to overcome this point of criticism, enhance ethical reflection and demonstrate possible practical interventions.

We invite the reader to critically accompany us on our journey through cases of AI ethics. We also ask the reader to think beyond the cases presented here and ask

fundamental questions, such as whether and to what degree the issues discussed here are typical or exclusively relevant to AI and whether one can expect them to be resolved.

Overall, AI is an example of a current and dynamically developing technology. An important question is therefore whether we can keep reflecting and learn anything from the discussion of AI ethics that can be applied to future generations of technologies to ensure that humanity benefits from technological progress and development and has ways to deal with the downsides of technology.

## References

AI HLEG (2019) Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence. European Commission, Brussels. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419. Accessed 25 Sept 2020

Alpaydin E (2020) Introduction to machine learning. The MIT Press, Cambridge

Andreou A, Laulhe Shaelou S, Schroeder D (2019) D1.5 Current human rights frameworks. De Montfort University. Online resource. https://doi.org/10.21253/DMU.8181827.v3

Aristotle (2000) Nicomachean ethics (trans: Crisp R). Cambridge University Press, Cambridge

Bengio Y, Lecun Y, Hinton G (2021) Deep learning for AI. Commun ACM 64:58–65. https://doi.org/10.1145/3448250

Bentham J (1789) An introduction to the principles of morals and legislation. Dover Publications, Mineola

Brey P (2000) Disclosive computer ethics. SIGCAS Comput Soc 30(4):10–16. https://doi.org/10.1145/572260.572264

Brusseau J (2021) Using edge cases to disentangle fairness and solidarity in AI ethics. AI Ethics. https://doi.org/10.1007/s43681-021-00090-z

Bynum TW (2001) Computer ethics: its birth and its future. Ethics Inf Technol 3:109–112. https://doi.org/10.1023/A:1011893925319

Childress JF, Beauchamp TL (1979) Principles of biomedical ethics. Oxford University Press, New York

Cremer CZ (2021) Deep limitations? Examining expert disagreement over deep learning. Prog Artif Intell 10:449–464. https://doi.org/10.1007/s13748-021-00239-1

Dreyfus HL (1972) What computers can't do: a critique of artificial reason. Harper & Row, New York

Escartín J, Saldaña O, Martín-Peña J et al (2015) The impact of writing case studies: benefits for students' success and well-being. Procedia Soc Behav Sci 196:47–51. https://doi.org/10.1016/j.sbspro.2015.07.009

Faggella D (2020) Everyday examples of artificial intelligence and machine learning. Emerj, Boston. https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/. Accessed 23 Sept 2020

Floridi L (1999) Information ethics: on the philosophical foundation of computer ethics. Ethics Inf Technol 1:33–52. https://doi.org/10.1023/A:1010018611096

Franks B (2017) The dilemma of unexplainable artificial intelligence. Datafloq, 25 July. https://datafloq.com/read/dilemma-unexplainable-artificial-intelligence/. Accessed 18 May 2022

Gambelin O (2021) Brave: what it means to be an AI ethicist. AI Ethics 1:87–91. https://doi.org/10.1007/s43681-020-00020-5

Hall W, Pesenti J (2017) Growing the artificial intelligence industry in the UK. Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy, London

Held V (2005) The ethics of care: personal, political, and global. Oxford University Press, New York
Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2
Kant I (1788) Kritik der praktischen Vernunft. Reclam, Ditzingen
Kant I (1797) Grundlegung zur Metaphysik der Sitten. Reclam, Ditzingen
Knight W (2017) The dark secret at the heart of AI. MIT Technology Review, 11 Apr. https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/. Accessed 18 May 2022
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
McCarthy J, Minsky ML, Rochester N, Shannon CE (2006) A proposal for the Dartmouth summer research project on artificial intelligence. AI Mag 27:12–14. https://doi.org/10.1609/aimag.v27i4.1904
Mill JS (1861) Utilitarianism, 2nd revised edn. Hackett Publishing Co, Indianapolis
Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507. https://doi.org/10.1038/s42256-019-0114-4
Ouchchy L, Coin A, Dubljević V (2020) AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. AI & Soc. https://doi.org/10.1007/s00146-020-00965-5
Rességuier A, Rodrigues R (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc 7:2053951720942541. https://doi.org/10.1177/2053951720942541
Schölkopf B, Locatello F, Bauer S et al (2021) Toward causal representation learning. Proc IEEE 109(5):612–634. https://doi.org/10.1109/JPROC.2021.3058954
Stahl BC (2021) Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies. Springer Nature Switzerland AG, Cham. https://doi.org/10.1007/978-3-030-69978-9
UNESCO (2021) AI ethics: another step closer to the adoption of UNESCO's recommendation. UNESCO, Paris. Press release, 2 July. https://en.unesco.org/news/ai-ethics-another-step-closer-adoption-unescos-recommendation-0. Accessed 18 May 2022
Weizenbaum J (1977) Computer power and human reason: from judgement to calculation, new edn. W.H. Freeman & Co Ltd., New York
Wiener N (1954) The human use of human beings. Doubleday, New York
Yin RK (2003) Applications of case study research, 2nd edn. Sage Publications, Thousand Oaks