Chapter 9
The Ethics of Artificial Intelligence: A Conclusion

Abstract The concluding chapter highlights broader lessons that can be learned from the artificial intelligence (AI) cases discussed in the book. It underlines the fact that, in many cases, it is not so much the technology itself that is the root cause of ethical concerns but the way it is applied in practice *and* its reliability. In addition, many of the cases do not differ radically from ethics cases related to other novel technologies, even though the use of AI can exacerbate existing concerns. Ethical issues can rarely be resolved to everybody's full satisfaction, not least because they often involve the balancing of competing goods. What is essential is space for human reflection and decision-making within the use of AI. Questions about what we can and should do, why we should act in particular ways and how we evaluate the ethical quality of our actions and their outcomes are part of what it means to be human. Even though Immanuel Kant believed that a good will is the only thing in the world that is ethical per se, a good will alone does not suffice where complex consequences may not be obvious. The complex nature of AI systems and their interaction with their human, social and natural environment require constant vigilance and human input.

Keywords AI ethics · Socio-technical systems · AI ecosystem · Solutions · Mitigation

This book of case studies on ethical issues in artificial intelligence (AI), and strategies and tools to overcome them, has provided an opportunity for learning about AI ethics. Importantly, it has also shown that AI ethics does not normally deal with clear-cut cases. While some cases provide examples of events that are obviously wrong from an ethical perspective, such cases are often about the *reliability* of the technology. For instance, it is obvious that AI-enabled robots should not present health, safety and security risks for users such as the death of a passenger in a self-driving car, or the smart-home system which allowed a man-in-the-middle attack. More difficult are cases where deliberation on the ethical pros and cons does not provide an immediate answer for the best approach—for instance, where robot use in elderly care reduces pressure on seriously overstretched staff but outsources important human contact

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_9

to machines, or where sex robots can be seen as violating the dignity of humans (especially girls and women) but at the same time helping realise sexual rights.

Looking at the cases across the different example domains in this book, one can make some general observations. The first refers to the application context of AI. Our case studies have aimed to be grounded in existing or realistic AI technologies, notably currently relevant machine learning. The ethical relevance of the cases, however, is almost always linked to the way in which the machine learning tool is applied and integrated into larger systems. The ethical concerns, then, are not focused on AI but on the way in which AI is used and the consequences this use has. For instance, the unfair dismissal case (Chap. 7) and the gender bias case (Chap. 2) are about the application of AI. Both the dismissal of staff without human input into the sequence and the training of AI devices on gender-biased CVs are about the use of AI. This is not to suggest that AI is an ethically neutral tool, but rather to highlight that the broader context of AI use, which includes existing moral preferences, social practices and formal regulation, cannot be ignored when undertaking ethical reflection and analysis.

This raises the question: how do AI ethics cases differ from other cases of technology ethics? As a first approximation it is probably fair to say that they usually do not differ radically. Many of the ethics case studies we present here are not fundamentally novel and we do not introduce issues that have never been considered before. For instance, the digital divide discussed in Chap. 8 has been debated for decades. However, the use of AI can *exacerbate* existing concerns and heighten established problems.

AI in its currently predominant form of machine learning has some characteristics that set it apart from other technologies, notably its apparent ability to classify phenomena, which allows it to make or suggest decisions, for example when an autonomous vehicle decides to brake because it classifies an object as an obstacle in the road, or when a law enforcement system classifies an offender as likely to commit a further crime despite a model rehabilitation record. This is often seen as an example of AI autonomy. It is important, however, to see that this autonomy is not an intrinsic part of the machine learning model but an element of the way it is integrated into the broader socio-technical system, which may or may not allow these classifications in the model to affect social reality. Autonomy is thus a function not of AI, but of the way in which AI is implemented and integrated into other systems. Ibrahim Diallo might not have been dismissed by a machine and escorted from the company building like a thief (see Chap. 7) if the AI system had been more transparent and required more human input into the dismissal process.

Indeed, another characteristic of current AI based on neural networks is their opacity. It is precisely the strength of AI that it can produce classifications without humans having to implement a model; that is, the system develops its own model. This machine learning model is frequently difficult or impossible for humans to scrutinise and understand. Opacity of this kind is often described as a problem and various approaches around explainable AI are meant to address it and give meaningful insight into what happens within an AI system. This raises questions about what constitutes explainability and explanations more broadly, including questions

about the explainability of ethical decisions: questions that may open up new avenues in moral philosophy. And while explainability is generally agreed to be an important aspect of AI ethics, one should concede that most individuals have as little understanding of how their internal combustion engine or microwave oven works as they have of the internal workings of an AI system they are exposed to. For internal combustion engines and microwave ovens, we have found ways to deal with them that address ethical concerns, which raises the question: how can similar approaches be found and implemented for AI?

A final characteristic of current AI systems is the need for large data sets in the training and validating of models. This raises questions about ownership of and access to data that relate to the existing distribution of economic resources, as shown in Chap. 4. As data sets often consist of personal data, they may create the potential for new threats and aggravate privacy and data protection harms. This may also entrench power imbalances, giving more power to those who control such information. Access to data may also be misused to poison models, which can then be used for nefarious purposes. But while AI offers new mechanisms to misuse technology, misuse itself is certainly not a new phenomenon.

What overarching conclusions can one draw from this collection of cases of ethically problematic uses of AI and the various interpretations of these issues and proposed responses and mitigation strategies?

A first point worth highlighting is that human interaction typically results in ethical questions. Adding AI to human interaction can change the specific ethical issues, but will not resolve all ethical issues or introduce completely unexpected ones. Ethical reflection on questions of what we can and should do, why we should act in particular ways and how we evaluate the ethical quality of our actions and their outcomes are part of what it means to be human. Even though Immanuel Kant believed that a good will is the only thing in the world that is ethical per se, a good will alone does not suffice where complex consequences may not be immediately obvious. For instance, as shown in Chap. 8 about AI for Good, the most vulnerable populations might be hit harder by climate change, rather than helped, as a result of the use of AI-based systems. This was the case with small-scale farmers in Brazil and Zimbabwe who were not granted credit to cope with climate change by bank managers who had access to forecasts from seasonal climate prediction. Likewise, seasonal workers in Peru were laid off earlier based on seasonal climate forecasting. In these cases, helicopter research to aid vulnerable populations in resource-limited settings ought to be avoided, as local collaborators are likely to be in a better position to predict impacts on vulnerable populations.

Ethical issues can rarely be resolved to everybody's full satisfaction, not least because they often involve the balancing of competing goods. AI raises questions such as how to balance possible crime reduction through better prediction against possible discrimination towards disadvantaged people. How do we compare access to novel AI-driven tools with the ability and motivations of the tool holders to benefit from the use of our personal data? Or what about the possibility of improving medical diagnoses amid crippling human resource shortages versus the downsides of automated misdiagnosis? Can an uncertain chance of fighting a pandemic through AI

analysis justify large-scale data collection? How could one justify the deployment of AI in resource-limited areas in the light of the intrinsic uncertainty and unpredictability of the consequences this may have on different parts of the population? What about the elderly lady whose only companion is a pet robot? All our cases can be described in terms of such competing goods, and it is rarely that a simple response can be given. The conclusion to be drawn from this is that awareness of ethical issues and the ability and willingness to reflect on them continuously need to be recognised as a necessary skill of living in technology-driven societies.

Another conclusion to be drawn from our examples is that the nature of AI as a system needs to be appreciated and included explicitly in ethical reasoning. AI is never a stand-alone artefact but is always integrated into larger technical systems that form components of broader socio-technical systems ranging from small-scale local systems in individual organisations all the way up to global systems such as air traffic control or supply chains. This systemic nature of AI means that it is typically impossible to predict the consequences of AI use accurately. That is a problem for ethical theory, which tends to work on the assumption that consequences of actions are either determined or at least statistically distributed in a way that can be accurately described. One consequence of this lack of clear causal chains in large-scale socio-technical systems is that philosophy could aim to find new ways of ethical reflection of systems.

In practice, however, as our description of the responses to the cases has shown, there is already a significant number of responses that promise to be able to lead to a better understanding of AI ethics and to address ethical issues. These range from individual awareness, AI impact assessments, ethics-by-design approaches, the involvement of local collaborators in resource-limited settings and technical solutions such as those linked to AI explainability, all the way to legal remedies, liability rules and the setting up of new regulators. None of these is a panacea which can address the entire scope of AI ethics by itself, but collectively and taken together they offer a good chance to pre-empt the significant ethical problems or prevent them from having disastrous consequences. AI ethics as systems ethics provides a set of ethical responses. A key challenge that we face now is to orchestrate existing ethical approaches in a useful manner for societal benefit.

AI ethics as an ethics that takes systems theory seriously will need to find ways to bring together the approaches and responses to ethical challenges that we have presented. The responses and mitigation strategies put forward here do not claim to be comprehensive. There are many others, including professional bodies, standardisation, certification and the use of AI incident databases, to name but a few. Many of these already exist, and some are being developed and tailored for their application to AI. *The significant challenge will be to orchestrate them in a way that is open, transparent and subject to debate and questioning, while at the same time oriented towards action and practical outcomes.* Regulation and legislation will likely play a key role here, for example the European Union's Artificial Intelligence Act proposal, but other regulatory interventions, such as the creation of AI regulators, may prove important (Stahl et al. 2022). However, it is not just the national and international policymakers that have to play a role here. Organisations, industry

associations, professional bodies, trade unions, universities, ethics committees, the media and civil society need to contribute. All these activities are based on the effort and contributions of individuals who are willing to participate in these efforts and prepared to reflect critically on their actions.

Tackling AI ethics challenges is no simple matter, and we should not expect to be able to solve all ethical issues. Instead, we should recognise that dealing with ethics is part of what humans do and that the use of technology can add complexity to traditional or well-known ethical questions. We should furthermore recognise that AI ethics often cannot be distinguished from the ethics of technology in general, or from ethical issues related to other digital and non-digital technologies. But at the same time, it has its peculiarities that need to be duly considered.

Our aim in this book has been to encourage reflection on some interesting cases involving AI ethics. We hope that the reader has gained insights into dealing with these issues, and understands that ethical issues of technology must be reflected upon and pursued with vigilance, as long as humans use technology.

Reference

Stahl BC, Rodrigues R, Santiago N, Macnish K (2022) A European agency for artificial intelligence: protecting fundamental rights and ethical values. Comput Law Secur Rev 45:105661. https://doi.org/10.1016/j.clsr.2022.105661