Chapter 3
Concepts of Ethics and Their Application to AI

Abstract Any discussion of the ethics of AI needs to be based on a sound understanding of the concept of ethics. This chapter therefore provides a brief overview of some of the key approaches to ethics with a particular emphasis on virtue ethics and the idea of human flourishing. The chapter reviews the purposes for which AI can be used, as these have a bearing on an ethical evaluation. Three main purposes are distinguished: AI for efficiency, optimisation and profit maximisation, AI for social control and AI for human flourishing. Given the focus on human flourishing in this book, several theoretical positions are introduced that provide insights into different aspects and ways of promoting human flourishing. The chapter concludes with a discussion of the currently widespread principle-based approach to AI ethics.

Keywords Ethical theory · Human flourishing · Purposes of AI · Ethical principles for AI

Ethical issues of AI are hotly debated and sometimes contested. In order to understand what they are and why they might be considered ethical issues, and to start thinking about what can or should be done about them, I start with an introduction to ethics, which is then followed by an empirically based discussion of current ethical issues of AI.

At its most basic level, ethics has to do with good and bad, with right and wrong. However, the term "ethics" is much more complex than that and the same word is used to cover very different aspects of the question of right and wrong. Elsewhere (Stahl 2012), I have proposed the distinction of four different levels, all of which are covered by the term "ethics":

1. Moral intuition, expressed in a statement of the sort: "This is right," or "This is wrong."
2. Explicit morality, expressed in general statements like "One should always /never do this."
3. Ethical theory, i.e. the justification of morality drawing on moral philosophy expressed in statements like "Doing this is right/wrong because …"
4. Metaethics, i.e. higher-level theorising about ethical theories.

B. C. Stahl, *Artificial Intelligence for a Better Future*,
SpringerBriefs in Research and Innovation Governance,
https://doi.org/10.1007/978-3-030-69978-9_3

This view of ethics is compatible with other views, notably the frequently suggested distinction between applied ethics, normative ethics and metaethics. It also accommodates the typical introduction to ethics that one can find in technology ethics textbooks (Van de Poel and Royakkers 2011), notably the dominant ethical theories of deontology and consequentialism.

3.1 Ethical Theories

Ethical theories are attempts to find an answer to the question: what makes an action ethically better or worse than an alternative action? Prominent examples of ethical theories include consequentialism and deontology. (I shall return to virtue ethics later.) Both of these originated during the Enlightenment period (mainly in the 18th century). They aim to provide clear rules that allow us to determine the ethical quality of an action. Consequentialist theories focus on the *outcomes* of the action for this evaluation. The various approaches to utilitarianism going back to Jeremy Bentham (1789) and John Stuart Mill (1861) are the most prominent examples. They are based on the idea that one can, at least in theory, add up the aggregate utility and disutility resulting from a particular course of action. The option with the highest net utility, i.e. utility minus disutility, is the ethically optimal one.

Deontology, on the other hand, is based on the principle that the basis of the ethical evaluation of an action is the duty of the agent executing it. The most prominent representative of this position is Immanuel Kant (1788, 1797), who formulated the so-called categorical imperative. The most often quoted formulation of the categorical imperative says “Act only on that maxim by which you can at the same time will that it should become a universal law” (translation, quoted in Bowie 1999: 14). This categorical imperative stops agents from rationalising exemptions for themselves. The interesting aspect of such a position for our purposes is that this view of ethics pays no immediate attention to the consequences of an action, but exclusively focuses on the motivation for undertaking it.

It is important to underline, however, that deontology and utilitarianism are not the only ethical theories that can be applied to AI, and to technology more broadly. In addition to virtue ethics, to which I will return shortly, there are other general ethical approaches such as the feminist ethics of care (Gilligan 1990) and ethics based on various religions. Applying ethical theories to particular application areas has resulted in rich discourses of concepts such as computer ethics (Bynum and Rogerson 2003, Bynum 2008a, van den Hoven 2010), information ethics (Capurro 2006, Floridi 2006) and technology ethics (Brey 2011) that are relevant to AI.

Entire libraries have been written about philosophical ethics, and I cannot hope to do justice to the many and rich nuances of ethical thinking. It may nevertheless be helpful to outline how ethics links to the human condition. This can explain some of the characteristics of ethics and it can shed light on whether or to what degree non-human artificial agents can be ethical subjects.

A key to understanding ethics, I believe, is that humans recognise that we all, despite many and far-reaching differences, have much in common. We could call this state "the shared features of the human condition". Human beings are fundamentally social. Without social structures and support we would not only die as infants, but also fail to develop the language and thus the conceptual understanding of the world around us that allow us to live our lives. We are possibly the only species that not only recognises that we exist but also knows that we are fundamentally vulnerable and mortal. We not only know this, but we feel it in profound ways, and we recognise that we share these feelings with other humans. The shared fate of certain death allows us to see the other as someone who, no matter how different from us they are, has some basic commonalities with us. We have empathy with others based on our experiences and the assumptions that they are like us. And just as we share the knowledge of death, we also share the experience of hope, of joy, of the ability to (more or less) freely develop projects and shape our world. This world is not just a physical world, but predominantly a social one, which is constructed using the unique capabilities of human language. Ethics is then a way to shape an important part of this social world in ways that take into account the shared aspects of human nature.

This description of human nature and the human condition has direct implications for the concept of ethics and what can count as "being ethical". Ethics does not exclusively reside in an action or an intention. Ethics is part of *being* in the world, to use a Heideggerian term (Heidegger 1993). It is characterised by an agent's ability not only to perceive different possible states of the world and decide between conceivable options, but to do so with a view to the meaning of such a decision for her own world and also for the world at large. This implies that the agent is consciously situated in this world, and understands it, but also has an emotional relationship to it and the fellow agents who co-constitute this world. Such an agent may very well make use of deontological or utilitarian ethical theories, but she does so in a reflective way as an agent who has a commitment to the world where these principles are applied.

This brief introduction to my ethical position points to the idea of human flourishing, which will become vital in later parts of this book: human flourishing linked to *being* in the world, understanding the limits of the human condition and the essential socialness of humans, which requires empathy. Of course, I realise that there are people who have no or little empathy, that abilities to interact socially and use language differ greatly, that many of these aspects apply to some degree also to some animals. Yet, to substantiate my position in AI ethics and the main ideas of this book, it is important that I do not draw inordinately on deontology and utilitarianism, but rather take into account a wider range of sources, and in particular virtue ethics.

3.2 AI for Human Flourishing

Current approaches to philosophical ethics as represented by consequentialism and deontology are largely rational and theoretical endeavours and mostly at home in academic philosophy departments. Ethics, however, has traditionally had a much broader meaning. For the ancient Greeks, philosophy was not just an intellectual endeavour but an attempt to find ways to live the "good life", the answer to the question: how should I live (Annas 1993)? The major philosophical schools of ancient Greece agreed that the cosmos had a purpose and that the individual good life, resulting in happiness (Aristotle 2007), was predicated on people fulfilling their role in society. This is the basis of virtue ethics, which is most prominently associated with Aristotle (2007) but whose main tenets are widely shared across philosophical schools. The focus of this approach to ethics is not so much the evaluation of the anticipated outcomes of an individual act or their intention, but providing guidance for the individual to help them develop a virtuous character.

I do not want to overly romanticise ancient Greece, whose acceptance of slavery and misogyny are not acceptable. However, virtue ethics as an approach to ethics has significant appeal, probably because it offers to provide guidance not only on individual problems but on how we should live our lives. This may explain why it has returned to prominence since the end of the 20th century and seen attempts to translate it into modern contexts (MacIntyre 2007).

Terry Bynum is one of several scholars who have succeeded in translating the ancient principles of virtue ethics into a modern technology-saturated context. He suggests the development of a "flourishing ethics" (Bynum 2006) which draws from Aristotelian roots. Its key tenets are:

1. Human flourishing is central to ethics.
2. Humans as social animals can only flourish in society.
3. Flourishing requires humans to do what we are especially equipped to do.
4. We need to acquire genuine knowledge via theoretical reasoning and then act autonomously and justly via practical reasoning in order to flourish.
5. The key to excellent practical reasoning and hence to being ethical is the ability to deliberate about one's goals and choose a wise course of action.

Bynum (2008b) has shown that these principles of virtue ethics are relevant to and have informed ethical considerations of information technology since its early days and can be found in the work of Norbert Wiener (1954), one of the fathers of digital technology.

Much research has been undertaken to explore how principles of virtue ethics can be applied to technology and how we can live a virtuous life in a technologically driven society. An outstanding discussion of virtue ethics in the context of digital technologies is provided by Vallor (2016), and, given that my approach relies heavily on her discussion, I will return to it later with reference to human flourishing.

As Bynum points out, people are endowed with different skills and strengths. Flourishing includes excellence in pursuit of one's goals, which implies that there are as many ways of flourishing as there are combinations of skills. Flourishing is thus not a one-size-fits-all concept but needs to be filled with life on an individual level. Before I return to a more detailed discussion of the concept of flourishing, I now want to discuss the motivations behind and purposes of developing, deploying and using AI, as these have a direct bearing on the ethical evaluation of AI socio-technical systems.

3.3 Purposes of AI

Understanding the purpose and intention of AI is important when thinking about the ethics of AI. Digital technologies, as pointed out earlier, are highly flexible and open to interpretation. They are logically malleable. They can thus be used for an infinity of purposes, which may or may not be aligned with the intention of the original developers and designers. Despite this openness of AI, it is still possible to distinguish different purposes that determine the design, development and use of systems. I distinguish three main purposes: AI for efficiency, AI for social control and lastly, as an alternative and complement to the two initial ones, AI for human flourishing (see Fig. 3.1).

When looking at current policy documents covering AI, one typically finds a mixture of all three of these motivations: AI can *improve efficiency*, which will lead to cost savings and thereby to economic benefits, which will trickle down, and people's lives will get better. A report to the President of the United States set the tone by highlighting the economic advantages and suggesting that "AI has the potential to double annual economic growth rates in the countries analyzed by 2035" (Executive Office of the President 2016). The European Commission expects that "AI could spread across many jobs and industrial sectors, boosting productivity, and yielding strong positive growth" (Craglia et al. 2018). And a committee of the United Kingdom's House of Lords hopes that "AI could spread across many jobs and industrial sectors, boosting productivity, and yielding strong positive growth" (House of Lords 2018).

A very different view of the use of technology including AI is to see it as a way of exerting *social control*. Rapidly growing abilities to collect data, in conjunction

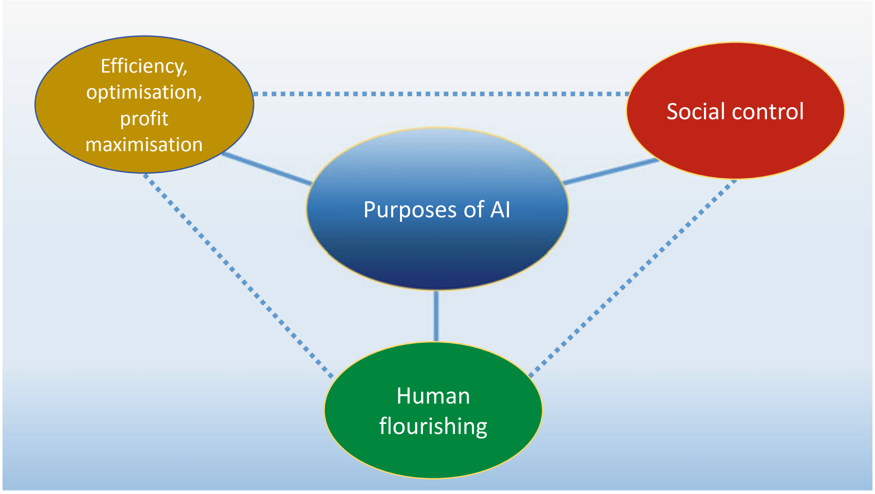

Fig. 3.1 Possible purposes of AI

with AI's ability to detect patterns and correlations between variables, allow for new ways of controlling human behaviour. This can be done in subtle ways, using the idea of "nudging" based on behavioural economics (Mullainathan and Thaler 2000, Camerer et al. 2004) or it can be done more vigorously, as for example in the Chinese social credit scoring system (Creemers 2018, Liu 2019).

> The system intends to monitor, rate and regulate the financial, social, moral and, possibly, political behavior of China's citizens – and also the country's companies – via a system of punishments and rewards. The stated aim is to "provide the trustworthy with benefits and discipline the untrustworthy." (Bartsch and Gottske nd)

AI as social control can also breach the limits of legality, as happened in the Facebook–Cambridge Analytica case, where social media data was used to illegitimately influence the outcome of democratic elections (Isaak and Hanna 2018). Zuboff (2019) offers a forceful argument that social control is a driving force and a necessary condition of success of what she calls "surveillance capitalism". In her analysis she does not focus on the term AI, but her description of the way in which new business models have developed and facilitated enormous profits is fully aligned with the concept of AI as converging socio-technical systems (see Fig. 3.1).

The third purpose of using AI, drawing on the earlier discussion of ethics, is to employ it for *human flourishing*. This means that AI is developed and deployed in ways that promote human flourishing. It can be used as a tool that helps individuals and groups identify worthwhile goals and supports them in their pursuit of excellence in achieving these goals. There are a number of suggestions on how to ensure that AI has positive consequences for individuals and societies, which is part of this third purpose of using AI for human flourishing: for example, attempts to construct a "good AI society" (Cath et al. 2016) or the discourse on AI for good that I discuss in more detail below in the section on the benefits of AI.

The three different views of the purpose of AI are represented in Fig. 3.1.

These three goals may come into conflict, but they are not necessarily contradictory.

The pursuit of efficiency and the resulting economic benefits can lead to a strong economy that provides the material substrate for human wellbeing. By generating wealth an efficient economy opens avenues of human flourishing that would otherwise be impossible. For instance, a move from coal-based energy production to solar energy is expensive. In addition, the pursuit of efficiency and profit creation can be a legitimate area of activity for excellence, and people can flourish in this activity.

Social control is often seen as problematic and in conflict with individual liberties. The use of information and communications technologies (ICTs) has long been associated with violations of privacy and the growth of surveillance (Lyon 2001). This concern traditionally saw the state as the source of surveillance. In these days of corporate giants that control much of the data and technical infrastructure required for AI, the concern includes the exploitation of individuals in new forms of "surveillance capitalism" (Zuboff 2019). But, again, there does not have to be a contradiction between social control and human flourishing. Humans as social beings need to define ways of collaborating, which includes agreement on moral codes, and these

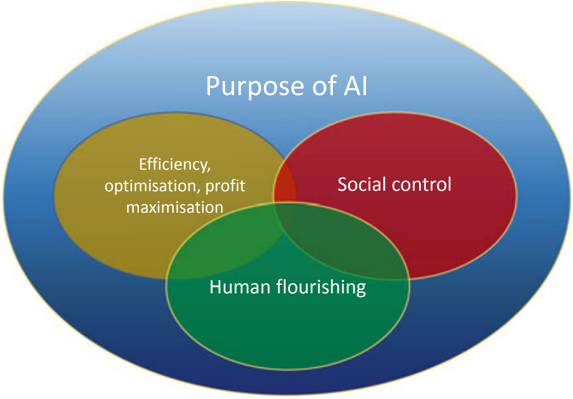

Fig. 3.2 Overlap of purposes of AI

need to be controlled and enforced in some way. While nudging as a policy instrument is contentious, it can be and often is used to promote behaviours that are conducive to flourishing, such as promoting a healthier lifestyle. Used especially in the United Kingdom, Australia, Germany and the US (Benartzi et al. 2017), nudging involves government-led campaigns to achieve given targets, for instance higher vaccination rates. For example, a US campaign involved sending out planning prompts for flu vaccination to citizens, which increased vaccination rates by 4.2% (ibid).

In the technology domain AI can be used to promote privacy awareness (Acquisti 2009), arguably a condition of flourishing. As I write these sentences, much of the world is under lockdown due to the COVID-19 pandemic. In the UK there is a heated debate around apps to be used to support the tracking and tracing of infected individuals (Klar and Lanzerath 2020). What this shows is that even forced social control through digital technologies may in some circumstances be conducive to human flourishing, for example, if it can help save lives and allow society to function. A Venn-type diagram may therefore be a better representation of the relationship of the three purposes (Fig. 3.2).

I must emphasise that the three purposes of AI listed in Figures 3.1 and 3.2 are not intrinsically contradictory, but rather describe the main fields of emphasis or different directions of travel that can guide the development and deployment of AI. My proposal is that the explicit aim to do the ethically right thing with AI can be described with reference to human flourishing.

This is not a novel insight. It draws from the ancient Greek philosophers and has been applied to ICT for decades. It has also been applied to AI. Virginia Dignum (2019: 119), for example, states: “Responsible Artificial Intelligence is about human responsibility for the development of intelligent systems along fundamental human principles and values, to ensure human flourishing and well-being in a sustainable world.” Mark Coeckelbergh (2019: 33) voices a similar view when he states that we “need a positive and constructive ethics of AI, which is not only about regulation in the sense of constraints but which also concerns the question of the good life and human and societal flourishing”. The principle of this argument is unproblematic and can also be found in AI policy proposals (ALLEA and Royal Society 2019). Who, after all, would say that they want to use AI to limit human flourishing? However, it raises the questions: how can we know whether human flourishing is promoted or achieved, and how can this be translated into practice? In order to answer these questions, I will now look at some theoretical positions on technology and its role in the world.

3.4 Theoretical Perspectives on Human Flourishing

Flourishing ethics is part of the tradition of virtue ethics and its historical roots in Aristotelian ethics. In order to answer the question, “How can we understand flourishing in practical terms?” it is helpful to look at other positions that share the aim of promoting human flourishing. Three positions that have been applied to technology, or that were developed specifically with research and technology development in mind, are important in this context: critical theory of technology, capability theory and responsible research and innovation. Each of these three offers an established theoretical approach that is consistent with human flourishing, and each has led to a wealth of insights into how flourishing can be observed and promoted..

Critical theory of technology is my first example of a theoretical approach relevant to AI that encompasses flourishing. Critical theory has a number of different possible roots. In its European spirit it tends to trace its origins to Marx’s criticism of capitalism. There is a recurrence of Marxist thinking in relation to digital technologies (Greenhill and Wilson 2006, Fuchs and Mosco 2017). However, much of critical theory of technology uses later developments of critical theory, notably of the Frankfurt School (Wiggershaus 1995). Andrew Feenberg’s (1993, 1999) work is probably the best-known example of the use of critical theory to understand modern technology. In addition, there has been a long-standing discussion of critical theory in the field of information systems, which draws on further theoretical traditions, such as postcolonialism (Mayasandra et al. 2006) and postmodernism (Calás and Smircich 1999).

Elsewhere I have argued that one central combining feature of the various different views of critical theory is that they aim to promote emancipation (Stahl 2008). The emancipatory intention of critical research, i.e. research undertaken in the critical tradition, means that resulting research cannot be confined to description only, but attempts to intervene and practically promote emancipation (Cecez-Kecmanovic 2011). Myers and Klein (2011), drawing on Alvesson and Willmott (1992), see emancipation as facilitating the realisation of human needs and potential, critical self-reflection and associated self-transformation. The concept of emancipation seems very close to the principle of human flourishing discussed earlier. My reason for bringing critical theory into this discussion is that critical theory has developed a set of tools and a high degree of sensitivity for understanding factors that can impede emancipation. Because of its roots in Marxist ideology critique, critical theory is well positioned to point to the factors limiting emancipation and flourishing that arise from the current socio-economic system, from labour processes and from capitalist modes of production. As will be seen later, these constitute probably the largest set of ethical issues associated with AI.

A second theoretical position worth highlighting in the context of human flourishing is *capability theory*. Capability theory has roots in philosophy and economics and is strongly associated with Amartya Sen (2009) and Martha Nussbaum (2011). The capability approach originated in development economics and the desire to find better ways of describing human development than purely financial and aggregate measures such as the gross domestic product. It is also directly linked to and based on the Aristotelian notion of flourishing (Johnstone 2007), and thus immediately relevant to a discussion of the ethics of AI and human flourishing.

The reason for highlighting the capability approach is that it has a history of application to information technologies (Oosterlaken and van den Hoven 2012), often in the context of studies of ICT for development and its focus on marginalised and vulnerable populations (Kleine 2010). It can thus be used as a way of sharpening the focus on the impact that AI can have on such populations. In addition, the communities working with the capability approach have developed tools for improving human functioning and freedoms and for measuring outcomes that have been recognised at a political level, notably by the United Nations. It is therefore suited to the creation of metrics that can be used to assess whether AI applications and uses benefit human flourishing.

The final theoretical position relevant to AI ethics and human flourishing is that of *responsible research and innovation* (RRI). RRI is a concept that has gained prominence in research and innovation governance since around the early 2010s. It has been defined as the "on-going process of aligning research and innovation to the values, needs and expectations of society" (European Union 2014). There are different interpretations of RRI (Owen and Pansera 2019), including that of the European Commission (2013), which consists of six pillars or keys (engagement, gender equality, science education, ethics, open access and governance), and that of the UK's Engineering and Physical Sciences Research Council (Owen 2014), represented by the AREA acronym (anticipate, reflect, engage and act), which is based on Stilgoe et al. (2013).

A much-cited definition of RRI proposed by Von Schomberg (2013: 63) sees RRI as

> a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society).

The reference to RRI is helpful in the context of AI ethics because it puts research and innovation explicitly into the societal context. The idea that the process and product of research and innovation should be acceptable, sustainable and societally desirable can be read as implying that they should be conducive to human flourishing. RRI can thus be understood as a way of promoting and implementing human flourishing. RRI is important in the context of this book because it is established as a term in research funding and familiar to policymakers. A recent proposal by the European Parliament puts heavy emphasis on RRI as a way to ensure ethical sensitivity in future AI research, development and deployment. The European Parliament (2020: 6) suggests that "the potential of artificial intelligence, robotics and related technologies … should be maximized and explored through responsible research and innovation".

Human flourishing in the broad sense used here is something that I believe most people can sign up to. It does not commit us to a particular way of life or require the adoption of a particular ethical position. It does not prevent us from using other ethical theories, including deontology and utilitarianism, to assess ethical questions (Bynum 2006). It is compatible with various theoretical positions beyond the three (critical theory, capability theory, RRI) introduced here. The choice of human flourishing was guided by the need to find an ethical language that can find traction across disciplinary, national, cultural and other boundaries. AI technologies are global and pervasive, but they have an impact at the local and individual level. An approach to the ethics of AI that aims to provide general guidance therefore needs to be able to build bridges across these many global divides, which I hope the idea of flourishing does.

3.5 Ethical Principles of AI

The main thesis of this book is that flourishing ethics can enlighten AI ethics and provide guidance in the development of practical interventions. The majority of currently existing guidelines were not drafted from one theoretical viewpoint but tend to use a set of ethical principles or values. What are these values?

The most comprehensive review of AI ethics guidelines published so far (Jobin et al. 2019) lists the following ethical principles: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity and solidarity. Each of these is comprised of components. Transparency, for example, refers to related concepts such as explainability, explicability, understandability, interpretability, communication and disclosure. The relationship between these concepts is not normally well defined and they can refer to different ethical positions. Elsewhere we have tried to clarify their normative implications (Ryan and Stahl 2020).

Another example, the ethics guidelines for trustworthy AI proposed by the EU's High-Level Expert Group on Artificial Intelligence (2019), has a tiered level of principles. The expert group proposes a framework for trustworthy AI that consists of lawful AI (which they do not cover), ethical AI and robust AI. This framework is based on four ethical principles: respect for human autonomy, prevention of harm, fairness and explicability. From these principles they deduce seven key requirements for the realisation of trustworthy AI, namely:

1. human agency and oversight
2. technical robustness and safety
3. privacy and data governance
4. transparency
5. diversity, non-discrimination and fairness
6. social and environmental wellbeing
7. accountability.

From these they then develop assessment methods for trustworthy AI and policy recommendations.

It is easy to see the attraction of this principle-based approach. It avoids making strong commitments to typically contested ethical theories. The principles themselves are generally uncontroversial, thereby offering the opportunity of a consensus. Maybe most importantly, the principle-based approach has been the basis of biomedical ethics, the field of ethics with the longest history of high-visibility public debate and need for societal and political intervention. Biomedical ethics in its modern form resulted from the Nazi atrocities committed during research on humans in concentration camps and the Nuremberg Code (Freyhofer 2004) that paved the way for the Declaration of Helsinki (World Medical Association 2008). It was codified and operationalised through the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979), which established the principles of biomedical ethics that remain dominant in the field (Beauchamp and Childress 2009): autonomy, justice, beneficence and non-maleficence.

Biomedical ethics has been hugely influential and underpins discussion of the human rights of patients (Council of Europe 1997). One crucial aspect of biomedical ethics is that it has been implemented via the well-established process of research ethics, based on ethics review, conducted by institutional review boards or research ethics committees, overseen by regional or national boards and strongly sanctioned by research funders, publishers and others.

There can be little doubt that this institutional strength of biomedical (research) ethics is a central factor guiding the AI ethics debate and leading to a principle-based approach that can be observed in most guidelines. This dominant position nevertheless has disadvantages. Biomedical ethics has been criticised from within the biomedical field as being overzealous and detrimental to research (Klitzman 2015). Empirical research on biomedical research ethics has shown inconsistency with regard to the application of principles (Stark 2011). And while largely uncontested in the biomedical domain, though not completely (Clouser and Gert 1990), the applicability of this approach to ethics in other domains, such as the social sciences, has been vehemently disputed (Schrag 2010).

There are two aspects from this discussion worth picking up for AI ethics. Firstly, there is the question of the implicit assumptions of biomedical ethics and their applicability to AI. Biomedical ethics was developed primarily to protect the rights of patients and research participants. This is no doubt transferable to AI, where the individuals on the receiving end of AI systems are worthy of protection. But because biomedical research predominantly aims to understand diseases with a view to finding cures, biomedical ethics is much less concerned with the purpose of the research. It is usually taken for granted that biomedical research pursues an ethically commendable goal: that of contributing to human health and thus to human wellbeing. Ethical concerns therefore do not arise from this goal itself but only from ways of achieving it. In the case of technical research, including AI research, it is not at all obvious that this implicit premise of biomedical research is applicable. The assumption that the research itself and its intended consequences are ethically acceptable and desirable is in need of much more questioning and debate, casting doubt on whether the process-oriented and principle-based biomedical research ethics process is a suitable one to base AI ethics on.

Secondly, biomedical principlism (Beauchamp and Childress 2009) leaves open the question of how to deal with conflicts between principles. This is a well-established problem for any ethical approach that is based on a set of non-hierarchical principles or values. In most cases it is possible to imagine situations where these come into conflict. Looking at the principles used in AI, it is, for example, easy to imagine a case where the principle of transparency would come into conflict with the principle of privacy. In order to successfully guide action or decision, the approach therefore needs to find ways of dealing with such conflicts. In addition, principlism has been criticised for being overly close to its US origins and not generalisable across the world (Schroeder et al. 2019).

Framing AI ethics in terms of human flourishing can address both concerns. By offering an overarching ethical ambition it proposes a point of comparison that can help address value conflicts. It also aligns more closely to 21st-century research ethics, which has been moving away from Western principles to global values (Schroeder et al. 2019). And it furthermore offers a perspective that does not take for granted that all research and technology innovation is desirable per se, but clearly posits flourishing as the overarching goal.

References

Acquisti A (2009) Nudging privacy: the behavioral economics of personal information. IEEE Secur Priv 7:82–85. https://doi.org/10.1109/MSP.2009.163

ALLEA, Royal Society (2019) Flourishing in a data-enabled society. https://royalsociety.org/-/media/policy/Publications/2019/28-06-19-flourishing-in-data-enabled-society.pdf?la=en-GB&hash=D521F71EB21F9369FAC26D7E1313398A. Accessed 23 Sept 2020

Alvesson M, Willmott H (1992) On the idea of emancipation in management and organization studies. Acad Manage Rev 17:432–464

Annas J (1993) The morality of happiness, New edn. Oxford University Press, New York

Aristotle (2007) The Nicomachean ethics. Filiquarian Publishing, Minneapolis

Bartsch B, Gottske M (nd) China’s social credit system. Bertelsmann Stiftung. https://www.bertelsmann-stiftung.de/fileadmin/files/aam/Asia-Book_A_03_China_Social_Credit_System.pdf. Accessed 25 Sept 2020

Beauchamp TL, Childress JF (2009) Principles of biomedical ethics, 6th edn. Oxford University Press, New York

Benartzi S, Beshears J, Milkman KL et al (2017) Should governments invest more in nudging? Psychol Sci 28:1031–1040. https://doi.org/10.1177/2F0956797617702501

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

Bowie NE (1999) Business ethics: a Kantian perspective. Blackwell Publishers, Malden, MA

Brey P (2011) Anticipatory technology ethics for emerging IT. In: Mauger J (ed) CEPE 2011: crossing boundaries. INSEIT, Nice, France, pp 13–26

Bynum T (2008a) Computer and information ethics. In: Zalta EN (ed) Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2008/entries/ethics-computer/

Bynum TW (2006) Flourishing ethics. Ethics Inf Technol 8:157–173

Bynum TW (2008b) Norbert Wiener and the rise of information ethics. In: van den Hoven J, Weckert J (eds) Information technology and moral philosophy, 1st edn. Cambridge University Press, Cambridge, pp 8–25

Bynum TW, Rogerson S (2003) Computer ethics and professional responsibility: introductory text and readings. Blackwell Publishers, Cambridge, UK

Calás MB, Smircich L (1999) Past postmodernism? reflections and tentative directions. Acad Manage Rev 24:649–671. https://doi.org/10.2307/259347

Camerer CF, Loewenstein G, Rabin M (2004) Advances in behavioral economics. Princeton University Press, Princeton, NJ

Capurro R (2006) Towards an ontological foundation of information ethics. Ethics Inf Technol 8:175–186. https://doi.org/10.1007/s10676-006-9108-0

Cath CJN, Wachter S, Mittelstadt B, Taddeo M, Floridi L (2016) Artificial intelligence and the "good society": the US, EU, and UK approach. Social Science Research Network, Rochester, NY

Cecez-Kecmanovic D (2011) Doing critical information systems research: arguments for a critical research methodology. Eur J Inf Syst 20:440–455. https://doi.org/10.1057/ejis.2010.67

Clouser KD, Gert B (1990) A critique of principlism. J Med Philos 15:219–236. https://doi.org/10.1093/jmp/15.2.219

Coeckelbergh M (2019) Artificial intelligence: some ethical issues and regulatory challenges. In: Technology and Regulation, pp 31–34. https://doi.org/10.26116/techreg.2019.003

Council of Europe (1997) The Oviedo Convention: protecting human rights in the biomedical field. https://www.coe.int/en/web/bioethics/oviedo-convention. Accessed 30 Oct 2018

Craglia M, Annoni A, Benczur P et al (2018) Artificial intelligence: a European perspective. Publications Office of the European Union, Luxembourg

Creemers R (2018) China's social credit system: an evolving practice of control. Social Science Research Network, Rochester, NY

Dignum V (2019) Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer Nature Switzerland AG, Cham, Switzerland

European Commission (2013) Options for strengthening responsible research and innovation. Publications Office of the European Union, Luxembourg

European Parliament (2020) Draft report with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies. European Parliament, Committee on Legal Affairs. https://www.europarl.europa.eu/doceo/document/JURI-PR-650508_EN.pdf. Accessed 25 Sept 2020

European Union (2014) Rome declaration on responsible research and innovation in Europe https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf. Accessed 24 Sept 2020

Executive Office of the President (2016) Artificial intelligence, automation, and the economy. Executive Office of the President of the United States. https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF. Accessed 23 Sept 2020

Feenberg A (1993) Critical theory of technology, New edn. Oxford University Press Inc, New York

Feenberg A (1999) Questioning technology, 1st edn. Routledge, London

Floridi L (2006) Information ethics, its nature and scope. ACM SIGCAS Comput Soc 36:21–36

Freyhofer HH (2004) The Nuremberg medical trial: the Holocaust and the origin of the Nuremberg Medical Code, 2nd revised edn. Peter Lang Publishing Inc, New York

Fuchs C, Mosco V (2017) Marx and the political economy of the media : studies in critical social science, reprint edn, vol 79. Haymarket Books, Chicago

Gilligan C (1990) In a different voice: psychological theory and women's development, reissue. Harvard University Press, Cambridge, MA

Greenhill A, Wilson M (2006) Haven or hell? Telework, flexibility and family in the e-society: a Marxist analysis. Eur J Inf Syst 15:379–388

Heidegger M (1993) Sein und Zeit, 14th edn. Max Niemeyer Verlag GmbH & Co KG, Tübingen

High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419. Accessed 25 Sept 2020

House of Lords (2018) AI in the UK: ready, willing and able? HL Paper 100. Select Committee on Artificial Intelligence, House of Lords, Parliament, London. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf. Accessed 23 Sept 2020

Isaak J, Hanna MJ (2018) User data privacy: Facebook, Cambridge Analytica, and privacy protection. Computer 51:56–59. https://doi.ieeecomputersociety.org/10.1109/MC.2018.3191268

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

Johnstone J (2007) Technology as empowerment: a capability approach to computer ethics. Ethics Inf Technol 9:73–87

Kant I (1788) Kritik der praktischen Vernunft. Reclam, Ditzingen, Germany

Kant I (1797) Grundlegung zur Metaphysik der Sitten. Reclam, Ditzingen, Germany

Klar R, Lanzerath D (2020) The ethics of COVID-19 tracking apps: challenges and voluntariness. In: Research ethics. https://doi.org/10.1177/2F1747016120943622

Kleine D (2010) ICT4WHAT? Using the choice framework to operationalise the capability approach to development. J Int Dev 22:674–692. https://doi.org/10.1002/jid.1719

Klitzman R (2015) The ethics police? The struggle to make human research safe, 1st edn. Oxford University Press, New York

Liu C (2019) Multiple social credit systems in China. Social Science Research Network, Rochester, NY

Lyon D (2001) Surveillance society: monitoring everyday life. Open University Press, Buckingham, UK

MacIntyre AC (2007) After virtue: a study in moral theory. University of Notre Dame Press, Notre Dame, IN

Mayasandra R, Pan SL, Myers MD (2006) Viewing information technology outsourcing organizations through a postcolonial lens. In: Trauth E, Howcroft D, Butler T et al (eds) Social inclusion: societal and organizational implications for information systems. Springer Science+Business Media, New York, pp 381–396

Mill JS (1861) Utilitarianism, 2nd revised edn. Hackett Publishing Co, Indianapolis

Mullainathan S, Thaler RH (2000) Behavioral economics. NBER Working Paper No. 7948. National Bureau of Economic Research, Cambridge MA

Myers MD, Klein HK (2011) A set of principles for conducting critical research in information systems. MIS Q 35:17–36

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979) The Belmont Report: ethical principles and guidelines for the protection of human subjects of research. US Government Printing Office, Washington DC

Nussbaum MC (2011) Creating capabilities: the human development approach. Harvard University Press, Cambridge, MA

Oosterlaken I, van den Hoven J (eds) (2012) The capability approach, technology and design. Springer, Dordrecht, Netherlands

Owen R (2014) The UK Engineering and Physical Sciences Research Council's commitment to a framework for responsible innovation. J Responsib Innov 1:113–117. https://doi.org/10.1080/23299460.2014.882065

Owen R, Pansera M (2019) Responsible innovation and responsible research and innovation. In: Simon D, Kuhlmann S, Stamm J, Canzler W (eds) Handbook on science and public policy. Edgar Elgar, Cheltenham UK, pp 26–48

Ryan M, Stahl BC (2020) Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc. https://doi.org/10.1108/JICES-12-2019-0138

Schrag ZM (2010) Ethical imperialism: institutional review boards and the social sciences, 1965–2009, 1st edn. Johns Hopkins University Press, Baltimore, MD

Schroeder D, Chatfield K, Singh M et al (2019) Equitable research partnerships: a global code of conduct to counter ethics dumping. Springer Nature, Cham, Switzerland

Sen A (2009) The idea of justice. Allen Lane, London

Stahl BC (2008) The ethical nature of critical research in information systems. Inf Syst J 18:137–163. https://doi.org/10.1111/j.1365-2575.2007.00283.x

Stahl BC (2012) Morality, ethics, and reflection: a categorization of normative IS research. J Assoc Inf Syst 13:636–656. https://doi.org/10.17705/1jais.00304

Stark L (2011) Behind closed doors: IRBs and the making of ethical research, 1st edn. University of Chicago Press, Chicago

Stilgoe J, Owen R, Macnaghten P (2013) Developing a framework for responsible innovation. Res Policy 42:1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press, New York

Van de Poel I, Royakkers L (2011) Ethics, technology, and engineering: an introduction. Wiley-Blackwell, Chichester, UK

van den Hoven J (2010) The use of normative theories in computer ethics. In: Floridi L (ed) The Cambridge handbook of information and computer ethics. Cambridge University Press, UK, pp 59–76

Von Schomberg R (2013) A vision of responsible research and innovation. In: Owen R, Heintz M, Bessant J (eds) Responsible innovation. Wiley, Chichester, UK, pp 51–74

Wiener N (1954) The human use of human beings. Doubleday, New York

Wiggershaus R (1995) The Frankfurt School: its history, theory and political significance, New edn. Polity Press, London

World Medical Association (2008) Declaration of Helsinki: ethical principles for medical research involving human subjects. https://web.archive.org/web/20091016152009/http://www.wma.net/en/30publications/10policies/b3/ Accessed 24 Sept 2020

Zuboff PS (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power. Profile Books, London