Chapter 8
Some Suggestions for How to Proceed

Abstract This final chapter makes some provisional suggestions for the development of codes of ethics based upon the discussion so far. This will be of necessity incomplete, but there is a need to contribute to ongoing debate. Any code of ethics needs to be embedded well into an organisation and its culture, and specific ways in which codes of ethics for AI might face problems are indicated. Procedures for drawing up and implementing codes need to take note of diversity of thinking style and of experience in participants. The problems of transparency inherent in the operation of some AI, together with the important public concerns about the impact of AI, means that maximising transparency and openness in codes of ethics, appropriate to a particular organisation, is highly desirable. Codes of ethics need to balance attention to abstract principles with specificity, especially in AI where application of ethical ideals must be translatable into concrete practice. Procedures for revision and critique of codes are essential. Ethical discussion leading up to codes of ethics, as well as the codes of ethics themselves, must include consideration of issues concerning boundaries of human functioning, which is a key issue in AI and which may be left out of some ethical debates. Particular attention to the implications of replacing or extending human agency, and impacts upon complex social systems, would be useful. Lastly, the Asilomar AI Principles are briefly discussed, as an example of a recent attempt to produce principles intended to stimulate debate and discussion about beneficial and ethical AI.

8.1 Organisations and Codes

A code of ethics are only as good as its organisational backing. The way in which development of codes of ethics for AI is managed, and how such codes are implemented, will be one element of such organisational integrity, for good or for ill. These points apply to codes of ethics in general; but some problems are likely to be especially acute in AI.

Codes of ethics may function more as window dressing than real applied policy. In AI, where fears abound, the temptation to produce a wonderful sounding code of ethics simply to ward off criticism may be especially acute. Conspicuous virtue can also be a trap for the content of the codes: overstating certain values or virtues might

© Springer International Publishing AG 2017

P. Boddington, *Towards a Code of Ethics for Artificial Intelligence*,
Artificial Intelligence: Foundations, Theory, and Algorithms,
DOI 10.1007/978-3-319-60648-4_8

make it impossible for the good to combat the bad. The control problem in AI makes this an especially important issue.

Given how AI challenges the basis of standard professional codes of ethics, there is particular reason for hard thinking about how to develop and implement such a code. There should be explicit attention to how values are imbued in practices and how they may be present subliminally in the language and framing around codes and regulations.

Appointing someone with specific responsibility for the institutional memory, to keep track of the organisation's own history and thinking regarding value issues might be valuable and useful in an area of such rapid change as AI, especially given how technological development can lead to incremental changes in value which over time may cross boundaries which were once 'lines in the sand', although this might be unfeasible for small organisations.

We've seen how the control problem in AI affects the authoritative basis of professional power in AI. It would be wise for organisations clearly to state these difficulties, ideally specified in relation to the specific forms of AI that concern them. We've also seen how widely some forms of AI may affect and disrupt society. Again, it would be wise for institutions to show awareness of when issues are touching on wider political, social, and cultural issues that are beyond their capacity to address sufficiently, even though these institutions may have a vital role to play in societal dialogues.

Not all organisations will have the same range of value concerns; private industry has different concerns from governmental organisations, and some organisations have more local, others more global, concerns. Precision and self-awareness in such matters is valuable, and likely to go further to gaining public trust than bland statements of very general value. The task of specifying values in relation to concrete particulars should also be easier.

8.2 Procedures for Drawing Up and Implementing Codes

Diversity in participation is needed in drawing up, revising, critiquing and implementing codes of ethics. The potentially transformative nature of AI heightens the need for diverse, constructive and creative input. We need diversity of opinion, thinking style, status, interests, experience, and of position in hierarchies; however, beware of falling into the trap of having 'tick box' quotas for 'diversity'. Consideration might be given to ensuring that diverse personality types are represented, to gain a full range of thinking styles. In addition, subject matter experts from outside the realm of AI, such as lawyers, economists, social scientists, public engagement, and others, will be useful. The inclusion of members who have serious interests outside the world of AI may be useful for maintaining an outside perspective. Attention should be given to the leadership of discussions regarding ethics.

Input from those with expertise in areas such as the social impact of technology, and those who stand in diverse relationships to the technology of AI, would be especially welcome. People with expertise in the history of ideas, and understanding of the historical sweep of changes in both technology and in ethics could make valuable contributions, given the disruptive nature of at least some AI.

Transparency: The problems of transparency inherent in some forms of AI mean that gaining maximum transparency elsewhere whereever possible is particularly desirable. Although private corporations may be chiefly answerable only to themselves, their boards of directors, and shareholders, as much transparency as possible about membership and recruitment is desirable, as well as steps to ensure a measure of independence for those with especial responsibility for ethics within an organisation. This must include openness about the operation of any ethics committee or board. This is especially true for those forms of AI which have wide or ubiquitous impact on the lives of millions or even billions.

Good communication with other bodies, and willingness to participate in public discussions and consultations, would be a virtue. This should include discussions about legal change and development since AI concerns questions of agency and the distribution of responsibility, also key concepts to legal systems.

Revision and critique: There must be provision for the revision of codes, and provision for whistleblowing procedures, and as well, good lines of communication to reduce any need for whistleblowing. Thought should be given to procedures for ascertaining the impact of codes.

Timing: Attention needs to be paid to the timing of discussions drawing up codes. There may be some need for swift responses to issues, but in general, where these issues are concerned, careful thinking which takes time is needed.

8.3 The Content of Codes

This section is not intended to be comprehensive, but merely indicates some suggestions based on discussions earlier in the book.

The specificity of codes: There is always a balance between the generality and precision of codes of ethics. In AI, where codes of ethics relate to the development of AI itself, they need to be in a form such that the engineers will be able to translate them into realisable steps. There may be a tension between producing codes of ethics that retain general principles, and that can be embedded in workable practice. This relates to questions of the distribution of responsibility and tasks throughout an organisation. General ethical statements about ‘producing benefit for all’ and so on, will simply have no impact unless they can be translated into concrete ways of making a positive difference on the ground. Codes may therefore need to be presented at different levels of specification.

Ethical Uncertainty and Rigid Rules: Can Virtue Ethics Come to the Rescue?

One common response to the difficulty of producing future-proofed codes of ethics in areas of rapid development or contextual uncertainty is to refer to virtue ethics (Atkinson 2009). This recognises the importance of equipping researchers and professionals with the ethical skills to make nuanced decisions in context, to provide careful contextualised interpretation of rules, and to judge when rules are no longer appropriate. For example, the Association of Internet Researchers have suggested a strategy of equipping people with *phronesis*, (practical wisdom) drawing on the Aristotelian conception of this (Aristotle 1999; AoIR 2012).

Note that the AoIR suggests an Aristotelian approach to deal with situations where the right ethical path is unclear. Aristotle is frequently quoted as claiming that in any matter of inquiry, one can only hope to produce the degree of precision which that subject area permits (Aristotle 1999). This is sometimes erroneously used to justify vagueness or a range of acceptable, (yet perhaps mutually incompatible) answers. Yet, for Aristotle, making the appropriate ethical decision was understood as getting the appropriate answer, as hitting a target as closely as possible, and he certainly did not intend to allow for ethical pluralism. The call for *phronesis* as a desiderata in codes of ethics for rapidly developing technologies may not in fact provide an answer, so much as indicate the depth of the problem.

Moreover, for Aristotle, crucially, few actually possess *phronesis*. It indicates wisdom achieved over years; on the point of the rarity of true moral wisdom, he was probably correct. The virtues are habits of thought and action—to do the right thing, in the right situation, with the right motivation and thoughts—and note, these habits are acquired within a stable cultural context, by learning from those older and more virtuous, and with the starting assumption that those embarking on the path to virtue already have a good understanding of ethical action, and a strong motivation to live a good life. The application of virtue ethics in a diverse setting of rapid technological development is questionable to say the least.

Note that many Aristotelian virtues would not fit with current values (e.g., he had slaves and women were kept out of public life). In other words, to talk of having a virtue ethic as a framework is to leave wide open what the virtues are. To know who exhibits *phronesis* we have to be able to identify who the good guys are. (It's interesting that a frequent theme of sci fi is the precise difficulty of knowing who's the good guy and who's the bad guy—this is no coincidence.)

There are foundational issues with an Aristotelian account of the virtues, since it is linked intimately to a teleological account of human nature basing the 'good for man' on the 'function' of mankind, which is our unique nature.

(continued)

But not only is such an account far more controversial in the twenty-first century. One feature of AI is the way in which raises questions about humanity's 'uniqueness' or otherwise, and raises questions about what our 'true nature' really is. By presenting us with such destabilising thoughts, by potentially bringing wide ranging changes to society and to how we interact with the world, AI produces precisely the polar opposite of the relatively stable and small world of ancient Athens in which Aristotle could write with confidence about the virtues.

The level of specificity and detail of codes will also be relative to the specific forms of AI in question: self-driving cars for international export, robots for local use in care homes, algorithms for use in search engines, all present different challenges. There may or may not be need to address global or cross cultural issues. Indeed, fine tuning the values of AI may well involve looking very closely at localised values and priorities.

Responsibility: Questions of responsibility and accountability, their distribution within an organisation, and attention to how the implementation of AI itself affects responsibility and accountability, should be included.

AI in context: Attention to issues concerning AI in use is important, although may be difficult where the particular context of application is not specified in advance. It may be important to consider procedures for liaison with others concerning the downstream application of AI and how it might impact upon complex settings.

AI and the law: Attention to legal regimes local and internationally will of course be needed; a lesson that can be learned from elsewhere is to raise the question of whether or not legal loopholes are being used exploitatively.

Support for further research, and active collaborations, would be welcome, including research into the ethical issues, and how best to further constructive developments in the ethics of AI.

8.4 Thinking About Ethical Issues in Developing and Implementing Codes of Ethics

Benefits of AI: It must be explicitly recognised how hard it is to assess the 'benefits' and 'harms' of AI, and how differently these may be understood; given the potentially transformative nature of AI, this especially important.

AI, agency, and idealisation: As described earlier, it would be a good idea to take note of the particular dangers of idealisation of human and machine agency in discussions of the ethics of AI, and of the question of how hype can distort thinking in AI.

Checking for incompleteness of ethical discussion: One way of attending to distortions of thinking is by implementing procedures to consider the different viewpoints of all those affected by particular developments in AI, and different ways that the ethical issues may be understood.

Including consideration of boundary issues in ethics: We've seen how different thinking styles in ethics include or exclude consideration about issues of boundaries, especially relevant in AI concerning boundaries of human agency and action and even physical boundaries. This often relates to debates about what is 'natural' or the issues which may inspire 'disgust'. Although many philosophers may argue against the relevance of these issues, they may be particularly important in AI, and particularly important for some of groups of people whose voices may currently be less heard in academia, as discussion earlier indicated.

Replacing or surpassing humans: Specific attention to the impact of replacing or supplementing human agency with machine agency on humans, and on how this then affects wider social systems, would be useful.

The limits of expertise: This will include specific recognition of how there might be wider impacts beyond the knowledge and immediate control of AI professionals. This will include recognition of the problems within the AI community of combatting unwise or even malicious AI.

Language and communication: As we've seen, there is a need for precision and understanding regarding AI and in particular some key terms such as autonomy and transparency. It's important to bear in mind the different ways such terms may be understood and implemented, and to check and recheck for good communication.

The public: There's a particular need to pay attention to how issues are communicated to members of the public, or rather, the many different publics. It is preferable to think of developing a dialogue with members of the public, rather than simply 'educating' them about AI.

8.5 Asilomar AI Principles

The recently developed Asilomar AI Principles, drawn up in January 2017 (Future of Life Institute), can serve as an example of an initiative to begin to draw up principles for AI, including ethical principles. They were specifically intended to promote discussion, as is appropriate, given the early stage of consideration of value issues in the development of AI, and given the desirability of wide involvement in the debates around the ethics of AI.

8.5.1 The Process of Producing the Principles

The Principles were discussed by participants at a conference in Asilomar organised by the Future of Life Institute. The process of developing the Principles is described on their website. The basis for inclusion in this conference is not specified, but it appears to involve various prominent people working in AI as well as those from other disciplines, including law, philosophy, economics, industry, and social science. Many participants were holders of Beneficial AI grants awarded by the FLI; as invitations were extended to Principle Investigators, I was not myself present. Although there was a range of expertise involved, the participants cannot be said to be 'representative' of their particular areas of specialisation in any formal way, in the absence of a specific process for ensuring representativeness. Principles drawn up by the prominent have their place, but may miss elements that might be uncovered by the inclusion of those with less visible power. Bearing in mind our discussions earlier about diversity and the facilitation of group intelligence, the list of names of attendees suggests that that approximately 20% of participants were women.

Prior to the conference, members of the FLI compiled various recent reports into AI and from these, distilled a list of opinions about how society should best manage AI, from this list they distilled out a set of principles that expressed some level of consensus. These were then sent out to conference participants in an iterative process that saw a revised list of principles put up for discussion at Asilomar, and refined again over several days of debate. Attendees finally voted on each Principle and only those with 90% approval were included in the final set of 23 Principles. The Principles are available online and those who wish to can add their names.

The process thus was designed to achieve consensus; this is of course one method of generating material for discussion, but debate is also especially worthwhile in contested areas, and it would have been interesting to know if there were any issues on which firstly, the reports initially used to draw up the Principles, and secondly, the Asilomar participants, were in serious disagreement. Reports with minority opinions clearly expressed can provide valuable material for debate. Note, too, that consensus may sometimes be achieved at the expense of abstraction and of choosing words which may mask disagreement. The FLI website recognises that the Principles are open to varying interpretations and are likely incomplete, and considers them aspirational.

The 23 Principles are divided into three sections: Research Issues, Ethics and Values; and Longer Term Issues. Below I give brief commentary on aspects of these Principles, drawing on the discussions throughout this book.

8.5.2 Research Issues in the Asilomar Principles

1. Research Goal: The goal of AI research should be to create, not undirected intelligence, but beneficial intelligence.
2. Research Funding: Investment in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies (with example questions added).

3. Science-Policy Link: there should be constructive and healthy exchange between AI researchers and policy makers.
4. Research Culture: A culture of cooperation, trust and transparency should be fostered among researchers and developers of AI.
5. Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

It is hard to disagree with any of these Principles. But are there ways they could be improved? One major omission in the groups mentioned are members of the public. This is unfortunate, especially given the difficulty of defining the key notion of what would constitute 'benefit' in anything, especially AI, which may drive deep into the heart of our entire account of value and meaning.

Notwithstanding the consensus-driven and aspirational nature of the Principles, some recognition of the institutional, financial and policy burden of these Research Principles would be useful in any development of them. Who will provide the funding for research into the beneficial use of AI? Consider the case of private corporations doing such research. It's common for such corporations to aspire to ethical principles—but they also have duties towards their shareholders and a need to make a profit, or at least keep solvent. Moreover, if research into beneficial uses of AI does come from private sources, this will leave many questions open, given the contested nature of what counts as a benefit. Would a private company be more likely to think that a 'benefit' involves steps which lead the populace to be dependent upon their products and services, or those of their corporate friends? Some indication of what *specific* issues there might be in AI would be welcome too.

And while recognising that there is not space for detail in such Principles, much of what is indicated here will depend upon the institutional and governmental context within which AI is being developed. Principle 4 regards Research Culture, but this requires robust and healthy institutions; this could be mentioned; and a note about why AI in particular has a difficulty with cooperation and transparency would be useful and would help give more precise direction to any thoughts about the implementation or further elaboration of the Principles.

8.5.3 Ethics and Values in the Asilomar Principles

6. Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

Comment: it's hard to argue with this one. There are of course challenges concerning assessing safety with regard to complex human-AI interactions

7. Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.

Comment: As an ideal, this is laudable. But there is uncertainty if it can be achieved technically. There are various moves available to deal with cases where the cause of harm is unverifiable, for example in law with regimes of strict liability, where attributions of the cause of harm are not necessary to assign responsibility for redress. I'd suggest: 'AI systems should be developed so that, as far as possible, it will be possible to ascertain the causes of any resulting harm, and steps taken to assign responsibility and redress where this is not possible. Full consideration to what constitutes harm should be given'.

8. Judicial Transparency: Any involvement with an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

Comment: It's pretty much up to judicial systems to decide on this one, and such questions are currently receiving much scrutiny, as we've seen in Wisconsin vs. Loomis. Cooperation between legal scholars and law makers, and the AI community, is of course essential. AI needs to be fully integrated into human systems, and legal systems already have their own set of ideals of operation and notions of procedural justice, which AI must only enhance, not weaken.

9. Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

Comment: again, a laudable sentiment, and aspirational. As I've argued, figuring out how to distribute and maintain responsibilities across a large network of often loosely connected people and institutions is a very vexed question. The notion of responsibility is also rather elastic and has various uses in context; one common reason why people resist calls to responsibility is because of how swiftly it leads, or may be perceived to lead, to blame. There are some good reasons for this: among them, that attribution of responsibility without adequate control is a major dimension of work place stress, with concomitant serious health effects (Marmot et al. 1997). Although it is desirable for designers and builders of AI to consider the misuse of their systems, calls to responsibility might be counterproductive if done in ways which suggest responsibility for problems over which they have scant or no realistic control.

10. Value Alignment: Highly autonomous systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation.

Comment: this is of course again aspirational. I would add explicit reference to the embedding of autonomous systems within human social and work settings and the necessity of understanding the possible complexities here. As Francesca Rossi stated in an interview on the Principles, '... when you have human and machine tightly working together, you want this to be a real team. So you want the human to be really sure that the AI system works with values aligned to that person. It takes a lot of discussion to understand those values' (Conn 2017c).

There is a tendency in the Principles to talk of AI as a whole. I'd also add that value alignment will be highly specific to each instance and context of use. In any event, it will only be in examining specific circumstances that value alignment can occur. This process could at its best even improve the value alignment for certain activities, if it involves clarity and explicitly operationalising underlying values.

11. Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

Comment: Again, naturally human dignity, rights and freedoms should be aspired to. However, the knotty question as always is, how do you achieve dignity, and which rights and freedoms? This vexed problem can often be cut to size by again noting that many AI systems will operate in certain contexts only. The extremely complex question of cultural diversity has been addressed earlier. Respecting people from other cultures is a given; it's part and parcel of a universalist ethic. Yet allowing unfettered cultural diversity of values is, as a matter of verifiable empirical fact, inconsistent with implementing certain understandings of human rights; cultures concern values. A set of Principles for AI can't be expected to sort this one, but given that AI professionals deal all day long with ironing out bugs and inconsistencies in computer programmes, they might have noticed the tensions here.

12. Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyse and utilize that data.

Comment: this is an example of a Principle which does at least mention the relevance of AI in particular to the issue. A major conceptual question is what counts as 'data they generate': for example, since individual data may be pooled, data needs to be analysed with considerable sophistication so it's not necessarily clear what the basis and extent of individual rights are. However, these questions are questions for data analysis in general and attention to the particular role of AI might add clarity. The role of AI can indeed involve helping to address the issues of individual control over data with AI driven solutions.

13. Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

Comment: Again, mention of the role of AI in escalating concerns about personal data use, and attention to any specific responsibilities that this produces, would tighten this Principle from a general issue about privacy to one focused on AI. Moreover, liberty is frequently in tension with AI; posing this Principle in the form of raising the question about what counts as 'reasonable' curtailment, and whether AI has anything to do with shifting conceptions of 'reasonable curtailment of liberty' in one direction or another, would be welcome. Everything hangs on what is construed as 'unreasonable'.

14. Shared Benefit: AI technologies should benefit and empower as many people as possible.

Comment: there's no reason given to explain why AI has any particular reason to be concerned with benefit and empowerment in general. If AI is produced by private companies, it will be in their economic interests to ensure good corporate reputation and a consistent customer base who can afford their products, but that they have any further duties to general benefit is unclear. But if AI were responsible for the loss of benefits or power, this does give reason for its producers to guard against this, and mitigate or offer redress. Again, a statement which more explicitly cited ways in which AI might reduce the power or benefits of people, and looked to specific ways of combatting this, might provide a more precise and hence fimer basis for moving forward.

15. Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

Comment: it's left entirely unclear how this could be achieved. I would suggest that some clarity about whose responsibility this is would be welcomed. We might be left in a situation where governments are forced to mop up the economic and social mess created by AI-induced redundancies and escalating wealth disparities.

A set of aspirational Principles without any indication of whose responsibility it might be to bring them about, or how this is to be implemented, is to that extent weaker. It's very early days for AI, but yet, Principles for AI would have more weight, the more they can be linked to concrete specifications.

16. Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

Comment: the Principle of keeping human control and choice over delegation of decisions is good; but note that it's ambiguous about whether this means 'some human should choose, not a machine' or 'all humans should be able to choose'—the former case might still mean that many other humans are subject to human-machine systems. The reach of AI in certain areas indeed makes this likely. This is an issue for the differential spread of power and influence in society under AI, and mention of this in any Principles would be welcome. This links of course to Principles 14 and 15 which concern the differential benefits of AI. Loss of control over aspects of one's life is one such possible harm of AI for many.

17. Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

Comment: this is an interesting Principle which raises an abundance of issues. It points to how far AI can reach into our lives. The problem raised suggests that there is a need for a variety of groups overseeing and commenting on how AI is interacting with our social and civic processes—this is important to recognise the importance of debate, and the difference of viewpoints possible here, as well as the impact upon views and levels of influence of issues like funding sources, representation in such groups, and so on.

There are many examples of how developments in AI are likely to impact upon social and civic processes, too many to illustrate here. The recent EU and Whitehouse reports raise concerns about its possible impact upon taxation, and the need for government intervention and support in developing essential areas to support the long term overall social interests of AI, where private financial interests may have insufficient individual reason to invest (European Civil Law Rules in Robotics 2016; Preparing for the Future of Artificial Intelligence 2016). The concerns of the EU with harmonisation (European Civil Law Rules in Robotics 2016) indicate a wish to step in before advances in other jurisdictions force those lagging behind to fit in with others. Hence, international relationships are also implicated. Long term, and global, thinking is needed. Yet, our current civic processes have been noted to work against the need for longer term thinking about AI (Conn 2017d).

18. AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Comment: achieving this will be challenging. One way to avoid an arms race is to let the enemy win; presumably this is not what those who signed these Principles had in mind. A topic for another book, or indeed, for many volumes.

8.5.4 Longer-Term Issues in AI

There are various longer term issues included in Principles 19–23. Just one will be discussed here.

23. Common Good: superintelligence should only be developed in the services of widely shared ethical ideals, and for the benefit of all humanity, rather than one state or organisation.

Comment: this seems to suggest that ensuring that superintelligence can be produced to align with widely shared ethical ideals is possible. And much hangs on how any such ‘widely shared’ ideals are identified. Ideals held by large minorities are nonetheless ‘widely shared’; ideals held by majorities can do untold damage to minorities.

8.5.5 General Comments on the Asilomar Principles

These Principles are of course an early step in the process of thought about beneficial AI. There are advantages to attempting to achieve consensus, but nonetheless, expressing some of these Principles in terms of the questions to be raised around these points, rather than as statements expressed with a degree of certainty, might help to open up and continue discussion, without forgoing consensus.

Likewise, although aspirational, it would be beneficial to try to focus them as closely as possible on the distinctive or typical role of AI, and to avoid statements of very general principle which raise issues which are not unique to AI, but might apply to any technology, or to any commercial or industrial enterprise. Contrariwise, the Principles tend to refer to AI in general, which then implies we need to consider value issues for AI in general, whereas very often, the value issues we need to consider are much more local and contextualised—and therefore, to that extent, easier to address.

More explicit reference to the way in which AI will be closely embedded in complex human systems, and therefore, to that extent more complex to assess, would be helpful in indicating the necessary direction of much future work. For this and for other reasons indicated above, although the work of professionals in AI is absolutely necessary, including technological work on issues such as safety, verification, and transparency, emphasis also needs to be given to the role of others, including members of the public, and the representativeness of those involved in discussions about the ethics of AI. This is certainly the case given the difficulty, discussed throughout this book, of ascertaining what constitutes ‘benefit’ in the development of AI; the Principles could usefully indicate awareness of this issue.