Chapter 4
Codes of Professional Ethics

Abstract This chapter outlines the features of the professional practice which lead to the necessity for codes of professional ethics, and which underpin the nature and typical content of such codes. There are a variety of codes and regulations regarding professional practices, which may serve different purposes. Members of a profession possess certain skills, knowledge and capacities that their clients and the general public typically lack. This creates a gradient of power and of relative vulnerability between the professional and others. Codes of ethics aim to mitigate the potentially deleterious effects, or the misuse, of such professional power. Codes of professional ethics may be backed up by hard or soft power. Since each profession deals with a certain area of endeavour, codes of professional ethics typically concern themselves with values, benefit and harms in relation to their own area of expertise. Nonetheless, there are general values underlying such codes, even if these are implicit. These may be hard to articulate and may indeed be controversial. The value of autonomy is examined as an example especially relevant to AI. Codes of ethics can only function effectively with both adequate institutional and societal backing. Understanding the history and context of development of codes of ethics is important to understand their underlying values, and especially where social and technological change is occurring. Codes of ethics may develop in response to catastrophe, in anticipation of problems, and with reference to codes of ethics in key areas, and all of these may give rise to problems. Codes of ethics may have certain failings, and in some cases even make a situation worse.

4.1 Introduction: The Varieties of Ethical Codes

There are different possible formats for ethical codes, regulations or guidance. These include codes of professional conduct produced by various professional bodies for their members, or by other regulatory bodies; safety standards, often produced by industry or governmental bodies and possibly by statutory powers; and research ethics codes and regulation produced by institutions funding, carrying out, or overseeing research. There are also statements produced by prominent members of a profession, such as the Asilomar Recombinant DNA Principles (Berg et al.

P. Boddington, *Towards a Code of Ethics for Artificial Intelligence*,
Artificial Intelligence: Foundations, Theory, and Algorithms,
DOI 10.1007/978-3-319-60648-4_4

1975), or by special interest groups, such as those opposed to autonomous weapons. Discussion papers and interim guidelines may be especially relevant in disputed areas or where technology is developing rapidly. Codes or guidance may be enforceable by law or by other penalty, such as by deregistration or professional disciplinary action, or may offer guidance only.

The project from which this book arises focuses on professional codes of ethics for artificial intelligence researchers. But given the turbulent landscape in which AI is developing, professional codes of ethics will need the backing and support of other formalised or less formal, and institutionalised ways of addressing the ethical questions confronting us. However, for simplicity and ease of explanation, we are going to commence by considering professional codes of ethics, their typical purpose and nature, and then draw out implications for codes of ethics for AI, as well as more widely for how professional and public debate in this area should proceed.

4.1.1 The Purposes of Codes and Statements of Principle

The various codes, declaration of principle, regulations, and laws that exist can have complementary roles, and may differ from each other perhaps because of matters of substance, and perhaps because of their context and intent. There is a role for codes and sets of principles that are aspirational. And there is also a complementary need for codes which can be operationalised into concrete action; this is especially the case where codes of ethics are intended for guidance for engineers at the front line of developing AI.

Codes and regulations in different settings may not translate well to other settings; or they may be very useful for cross-fertilisation of ideas. Codes may be designed for local or national use, or may aspire to international application. A commercial organisation will have its own financial interests which may be nested within legal and ethical concerns, but which will have an impact upon any codes of ethics they produce; government codes and regulations may deal extensively with economic issues but with quite a different agenda than that of a private corporation or a professional body. In a legal context, there are ethical considerations in formulating and applying the law, but the law may lack the nuance that is needed for a rich account of ethics. Contrariwise, the law needs to spell concepts out in sufficient detail that judgements can be made in particular cases. This can mean that the law, including case law, can be a very useful source for considering how to operationalise and add detail to general and abstract concepts in ethics. This could be particularly useful in our area of concern, where developments in technology and changes in social relationship are presenting us with the need to apply central ethical concepts in new contexts.

A code of ethics should not be seen as complete and self-sufficient, for such codes exist in a particular context (Bowden and Surma 2003), and without the backing of a supportive institution, a code of ethics on its own will be of scant use

(Bowie 2009). Accompanying texts, and their institutional context, can be helpful and indeed necessary in their interpretation and implementation (McKerrow 1993). It is often here that key value assumptions are located. Looking at these closely will be especially important in certain contexts: where values are disputed; where values are fundamental and deeply held; where there is rapid technological and societal change occurring. All three apply in the case of AI.

4.2 Professional Codes of Ethics Tend to Have Certain Commonalities

The following is not intended as a full review of the features of professional codes of ethics, but discusses features of particular interest to the question of developing a code of ethics for AI. We need to examine the general rationales for having such codes of ethics or codes of practice in the first place.

4.2.1 Relations Between Professionals, Clients and Others

Gradients of expertise and resources between professionals and others: A code of professional ethics concerns the behaviour and services produced by a professional, who has a certain expertise and who produces something or delivers a service. Thus, the professional has skills and knowledge that the client group typically does not have, producing a gradient of expertise and resources, which then generates a relative vulnerability that gives rise to potential ethical problems that the codes aim to address. In many cases, the professional skill set is accredited, giving prestige to the professional group and presenting barriers to those without the credentials, regardless of their actual level of expertise. The specifics of a particular profession in a particular social context act to shape the resulting codes. Note that their specific professional role gives professionals concomitant additional moral and professional responsibilities; and the opportunities a profession affords also gives opportunities for corruption or unfair use.

Generally, one of the relative vulnerabilities between professionals and others is a general epistemic vulnerability with greater knowledge on the part of the professional, notwithstanding that specific knowledge and practical capabilities on the part of the client might be crucial to the implementation of professional skills. The relative epistemic vulnerability of the client then helps to shape key aspects of professional ethics; for instance: undertakings to assure levels of professional competence, to work only within one's sphere of competence, and to update skills and knowledge appropriately; undertakings of honesty and transparency in dealings with the public and with clients and full disclosure of risks, including taking further advice as needed; undertakings to operate within the law of the appropriate jurisdiction and any relevant local or regional government regulations (which is often simply implied).

The requirements of honesty and transparency will usually involve being able to give an account of actions taken and reasons behind them. An assumption behind this is that individual members of a profession themselves, and the profession as a whole has a significant grasp on its activities and can hence be in adequate control, and at the very least, to insure against unforeseen loss of control.

There will be a working assumption of relative stasis or incremental development in an area, in this sense: that the progress in this area is not outstripping the profession's capacity to understand and control its own area of endeavour. This is of course a matter of degree, since technology and knowledge constantly evolve. But to serve its function, any code of professional ethics has to be capable of addressing significant developments in its area of operation.

Professional codes of ethics are centred on clients but also usually need to refer to the public. The product or service is intended to produce benefit to the clients, and perhaps more widely. There is usually then a concomitant possibility of producing harm, which in the case of some professions can be severe. This harm in particular may affect those other than the clients, hence the need for codes of professional ethics to consider the general public and to make undertakings not to harm (for example, through consideration of the environmental effects of a profession's activities).

4.2.2 Professional Codes of Ethics, Enforcement, and Authority

Codes of ethics ideally outline procedures for reporting problems and violations of codes, which may include protection for whistleblowers and accounts of penalties for proven misconduct. This should draw our attention to the institutional context of professional ethics. Note that there is considerable evidence that, despite professional and legal safeguards, whistleblowers often fear poor treatment and may indeed suffer retaliation (Mesmer-Magnus and Viswesvaran 2005).

The authority and enforcement of codes of ethics may involve professional sanctions, restrictions on membership of professional bodies (which for some professions may make it impossible to practice) and, in worse cases, legal ramifications. Enforcement also occurs through the soft power of the authoritative weight and respect with which the relevant professional body and its codes of ethics are held.

The enforcement of codes may also trade on the relative homogeneity and education of professionals. They have a lot to lose from loss of social standing and income. They have gained a relatively good deal from society, on average. They have been at least to some extent, inculcated into organisations and companies. (This is no guarantor of behaviour, of course. There are many notable examples of spectacular individual failure and institutional corruption. But it forms part of the apparatus of compliance.)

There can be cooperation between different bodies for the enforcement of codes of conduct. For example, concern over the bias in findings of research by pharmaceutical companies by the suppression of negative results has led to moves whereby clinical trials must be openly registered before their start, and academic journals will not publish any trials which are not compliant with this (De Angelis et al. 2004). This may show the effectiveness of outside pressures on professional organisations or companies in helping to change standards of ethics.

4.2.3 Professional Codes of Ethics and Professional Values

There is an assumption of professional value. This relates to a pervading, vital, but sometimes unnoticed background assumption that the product or services of the profession are of general individual and/or society benefit. This assumed value also contributes to the relatively high social standing of members of recognised professions. This assumption is rarely spelled out or argued for in professional codes themselves, but is more implied by the prestige, the training, the professional regard, that surrounds the codes.

Professional practices tend to deal with specific values, arising from a complex of the broad nature of the client group and the nature of the professional services involved. The benefits involved are understood in terms of the particular area of expertise of the profession; avoidance of harms may, of practical and legal necessity, be understood more broadly than the benefits accruing to clients, since they will have to take into account wider consequence. Note, too, that these harms and benefits will tend to be cashed out, not necessarily in terms of a global ethic of human value, but in reference to the particular values of the product or services in question. This will be important in considering codes of ethics in AI.

Linked to this assumption of value, *members of the professions tend to have a relatively high social standing*. Indeed, the very existence of a professional body which produces codes of ethics or conduct also itself helps to contribute to the relatively high status of the professions. Codes may contain undertakings not to bring the profession into disrepute, and undertakings to maintain or improve the social status of the profession. The relatively high social standing also feeds into the soft powers supporting the codes' authority.

4.2.4 Values Underlying Professional Codes of Ethics

There will be explicit values embedded in professional codes of ethics, but also a base of underlying values. The values that lie behind professional codes of ethics will on the whole be values largely shared by the surrounding society, focused towards the particular area of practice of the profession, very often with stricter or additional duties placed on the professionals. As debate and thinking about ethics continues, and as society changes, there may be changes in how these underlying values are articulated and promoted.

However, a fully consistent and agreed set of *underlying* values may be hard to discern. Differences of interpretation and emphasis may mask or reveal deeper differences of opinion, or commonalities, between individuals, groups and communities, and geographical regions, towards these broad underlying values. Even one individual may not have fully consistent understandings of some core value terms: this has been shown for privacy as we saw in Sect. 2.8.3.

4.2.4.1 The Example of Autonomy

Autonomy is not just a core value in contemporary society, not just a core value underlying many codes of professional ethics such as codes of medical ethics, it's of particular concern to us as a key to AI which is developing autonomous systems and machines. It's both a normative value that we aim for in attempts to respect autonomy; and a key notion underpinning our very conception of the moral agent, moral motivation and moral responsibility. It is not just one of our values, it is a presupposition of how we understand our values. It's key, for example, to current understanding of responsibility in warfare, which is challenged by autonomous weaponry (Roff 2013).

Consider: respect for the autonomy of the individual is a core value in codes of medical ethics, expressed in various ways and articulated via concern for issues such as confidentiality and free and informed patient consent. The history of medical ethics over the last century or so can be read in no small way as the history of how patient autonomy has been granted greater and greater emphasis, as opposed to the 'doctor knows best' model (Beauchamp and Childress 2001). At the same time, this then raises questions about the autonomy of medical staff themselves, as seen in debates about the limits of conscientious objection for medics and pharmacists. Such debates are indeed, changing and some would say, undermining, the very idea of the medical profession, and replacing it with a service industry model. There are complex interactions between the expression of value and societal and technological change that it would be very hard to track with complete precision.

There are *philosophical and practical questions* and differences in how exactly the value of autonomy should be understood. Here's one challenge: respect for individual autonomy in clinical medicine may sit in some tension with principles of public health. So we need to understand how to respond when different values that we have clash. The question of the priority of the individual over the group is one of the most central questions of ethics.

There are also large *cultural differences* in how, and to what extent, individual autonomy in medicine is to be valued. A greater emphasis may be placed on community values or social cohesion, for example. Reading journal articles on medical ethics, it's often fairly easy to guess if the authors originated in the USA, or in Northern Europe, by the ways in which autonomy is discussed and ranked alongside other more communitarian or social-oriented values; there are even greater differences visible in discussions of medical ethics from other regions (Padela et al. 2015).

There are *individual differences* in how we value autonomy as well, which may be visible in the work of different moral philosophers, and which also may have strong effects on political affiliations.

Since we are considering the development of technology, note importantly that *scientific findings, technological developments, and brute facts can challenge thinking and action concerning autonomy*. How do we carry on valuing autonomy, for example, in patients with advanced dementia, an increasing problem in advanced societies with aging populations and stresses on social care? (Bridges and Wilkinson 2011). Often, it's advances in science and technology which are presenting us with new, or newly acute issues for autonomy. For instance, the science of genetics challenges simplistic ideas that individuals should have control over 'their' medical information, since genetic information is shared between biologically related individuals (Rhodes 1998). Our views of concepts related to autonomy, such as privacy, individual rights, group rights, and so on, shape our often uncertain and frequently contradictory responses to such developments (Laurie 2001). When we consider the case of AI, the developments of codes of ethics, and assessment of the impact of AI on individuals and societies, we will need to consider such complex interlinking webs.

4.2.4.2 Articulating Values Underlying Professional Codes of Ethics

Providing a definition of underlying values can be surprisingly hard. It's easy to state the goal of medicine is health . . . or is it the elimination of disease? And how do we even draw a distinction between disease and health—this is much harder than may at first appear, and the philosophy of medicine has long grappled with this question (Boorse 1975).

Definitions of such key terms are not simply there to describe 'reality'. They have a function to perform. We should note that the practice of medicine continues: it's in hard cases that these definitional issues are important, and indeed, they are the stuff of difficult policy debates. Yet, at least in medicine, we are considering a long standing practice; developments in AI may be harder to trace and more disruptive of social practices and values.

A Definition of Health, Extreme Social Change, and Some Thoughts for AI

A widely cited definition of health from Aboriginal Australia states:

Aboriginal health is not just the physical well-being of an individual, but is the social, emotional and cultural well being of the whole community in which each individual is able to achieve their full potential thereby bringing about the total well being of their community. It is a whole-of-life view and includes the cyclical concept of life-death-life.

Health to Aboriginal peoples is a matter of determining all aspects of their life, including control over their physical environment, of dignity, of community self-esteem, and of justice. It is not merely a matter of the provision of doctors, hospitals, medicines or the absence of disease and incapacity. (Houston 1989)

This definition not only includes culture, but justice, and in context, therefore includes consideration of historical events. The health status and life expectancy of Australian Aboriginal peoples is far lower than that of the Australian population as a whole. This very broad definition of health therefore needs to be understood with reference to the devastating impact upon the lives and well being of indigenous Australians by European colonisation of their lands (Boddington and Räisänen 2009).

How is this relevant for AI? There are many who consider that the impact of AI on our lives is not just going to be immense, but also unpredictable. We may not be able to understand what's coming (Vinge 1993).

It struck me that this could perhaps be, in very broad terms, analogous to the unimaginable shifts in life that the indigenous peoples of Australia had thrust upon them. *And so note*, it's precisely this rapid and profound change that's one key motivator for the breadth of the Aboriginal Australian health definition and the reference to culture and history. Likewise, in considering ethical issues arising from the advent of AI, it's likely to be important to look very broadly, and to keep an eye on history and on culture, to consider what is lost and what is changing.

These underlying values may also be up for debate, and here it is particularly pertinent that wider scrutiny may occur. This is especially the case when the actions of a profession have wider social significance. It's vital, too, that academic disciplines taking a lead in ethical discussions in a particular area are self-critical, and avoid domination by particular ideologies or factions. For example, in bioethics, some dominant voices currently are those who take certain utilitarian or libertarian views, and it's been argued forcefully that certain core values, including the value of autonomy and of individual contractual obligations, which are shaping discussion, need urgent examination and critique (Dawson 2010). It's such core values that may also shape discussion of the ethics of AI; we need careful scrutiny and broadly based imaginative thinking.

4.2.4.3 Underlying Professional Values May Be Focused Towards Protecting Individuals

The values underlying a code of professional ethics are shaped by views of primary professional responsibilities. Given the client focus of professions, there may be particular attention to protecting individuals and hence a stress on values which pertain to individuals *qua* individuals, such as privacy, autonomy, and individual property. Codes of professional ethics do however (usually) call attention to the need to protect the public; but a code of professional ethics may well assume that the professional's primary duties are to *benefit* their specific client group, whilst *avoiding harm* to the public. At the same time, a code of ethics may be designed to promote and protect the financial and professional interests of a particular group (Hammersley 2009). Indeed, the very bureaucratic machinery of ethical regulation as a whole might serve the purpose of promoting technological, economic and industrial advancement for a particular group or national region (Dingwall 2008).

Note two points. Firstly, that codes of ethics might then focus on values belonging to a relatively individualistic ethic. And, as valuable and as central the protection of individuals may be, additional values are needed in other contexts. For instance, there are somewhat different considerations operating within clinical ethics compared to public health ethics and compared to medical research.

For, secondly, we need to consider how the relevant services or products potentially affect those individuals who are not clients, and indeed how they affect society as a whole. However, if these questions aren't considered the direct responsibility of individual members of a profession, codes of ethics for that profession may be the wrong place to address these.

A particular problem for AI: Moreover, readily identifiable dangers such as structural collapse or the spread of contagious disease might attract scrutiny, but where technologies are new, rapidly developing, and potentially disruptive or transformative of social relations, as in AI, it will be a complex and often difficult task to ascertain exactly what broader ethical and social issues will arise, and even harder to untangle and trace how a particular technology contributes to these. In such cases, greater scrutiny and careful research to uncover impacts will be helpful, indeed, vital.

4.3 Codes of Ethics and Institutional Backing

A code of ethics is only as good as the institution behind it, and the ethos that operates within that institution. Many a company that has collapsed in the midst of corruption scandals, had inspirational codes of ethics languishing untouched in a golden frame on the CEO's penthouse office wall (McLean and Elkind 2013, 2004). A code of ethics plays only a certain part in the ethical conduct of an institution, and only if it is thoroughly embedded into multiple practices within an organisation can it really have a tangible impact (Bowie 2009).

Broader social and political forces can also undermine the integrity of the best codes of ethics. Take a look at one of the first ever codes of professional ethics for medicine. This code distinguished ‘therapeutic’ from ‘non-therapeutic’ research. It included the principles of beneficence and non-maleficence; it was based on an ideal of patient autonomy; it outlined a new legal doctrine of informed consent, which had to be clearly given and based upon appropriate information, with written documentation of consent procedures. There were clear structures of responsibility, and experimentation on the dying was prohibited.

This historically very early and impressive code of medical ethics, ‘Guidelines for New Therapy and Human Experimentation’, was issued by the Reich Minister of the Interior in Germany in 1931. It was not many years before doctors who were fully aware of such a code were involved in some of the worst atrocities of medical ‘experimentation’ that human beings have ever done to other human beings (Vollmann and Winau 1996).

4.4 The Context of Codes of Ethics

To understand many codes of ethics fully, we need to examine the institutional background, history and rationales for their production. This context can help us understand what values were addressed, consider how the landscape has changed, and consider who and what has influenced the codes as they currently stand and why. This can help us to think critically about how to amend or develop the codes, and to recall the root values that motivated their development.

For example, codes of medical ethics cannot be fully understood without at least some awareness of the history behind such codes, including the development of medical ethics in the twentieth century since the Nuremberg trials, the development of the Nuremberg code, the Helsinki declaration and its many revisions, and other such developments around the globe (Shuster 1997). Note however, that accounts of the history of ideas and regulations is always complex, and some people contest any simplistic account of ethics regulation as a straightforward attempt to combat abuse, especially as vested interests sometimes may play a part (Dingwall 2008).

The Nuremberg Trials: A Baseline of Evil

The history of the regulation of medicine cannot be understood without understanding the Nuremberg trials. These addressed the appalling abuse of human beings in medical ‘experiments’ of profound cruelty. One response to this was to draw up codes of medical ethics to protect the individual to try to ensure that such abuse could never occur again.

(continued)

An observation: the event of the Nuremberg trials must then be seen as pivotal in internationally recognised codes of medical ethics. Why were these codes so readily adopted? Because the abuse of subjects in medical experiments in Nazi Germany was so vile, so inhuman, so degrading, that it is impossible not to consider it an evil. [And note that similar atrocities have been committed elsewhere, for example the inhumane medical experimentation carried out during WW2 by Japanese in Unit 731 (Williams and Wallace 1989)]. It is worth remembering this when the issue of relativism and how to develop and apply codes of ethics cross culturally and internationally is considered. In these and other atrocities, the human person—the most complex creature in the known universe—was treated as a mere subject, a thing.

We can also note a key lesson from Nuremberg. The oft-heard plea, 'I was only following orders' was thrown out as an excuse. The atrocities were far too bad for that. Simply going through the motions, simply following a code, a set of instructions, is not a morality. The individual was charged to stand up against the bureaucratic apparatus of evil, as a few in fact had.

But note that the 'following orders' excuse also reduces the human person as moral agent, to something less than human, a cog in an evil wheel. The attempted denial of humanity was doubly tragic in that it applied both to those who acted, as well as to the profoundly suffering humanity upon whom they acted.

What's this got to do with AI? Since advances in AI are precisely raising questions about the nature of the human agent, and the nature of machine agency; since they present us with potentially profound disruptions to our individual and collective lives; since such changes are happening so fast, it will be as well to recall such fundamental moral starting points as we attempt to think through the ethical questions of AI.

Codes of ethics (and laws and other regulations) have developed in response to catastrophes or scandals, and understanding this can help to understand how codes have grown up as they have. For example, responses to the Tuskegee Syphilis trial have had a big impact upon medical ethics, to name just one of many such instances (Reverby 2012). But as vital as responses to catastrophe have been, developing codes in this way can have pluses and minuses. 'Hard cases make bad law', and responding to something tagged as a 'desperate case' may skew our thinking (Moore 1989). For example, responses to the events such as the thalidomide tragedy made it harder to carry out research on pregnant women, and prevented the use of thalidomide even in patients with scant or zero chance of pregnancy (Benatar and Singer 2000). In AI, one hopes to avoid catastrophe of course, especially if we are talking about existential risk, but we need to consider very carefully how to achieve this.

Extrapolating from Examples, Telling Stories and Gaining Insights

In responding to catastrophe and other bad events, or indeed from examples of good conduct, we are extrapolating from one case to the next. Great care is needed. How cases are described and the context in which they are placed will have a large impact upon how they are interpreted and what lessons are drawn from them. Tod Chambers' book *The Fiction of Bioethics* shows how by writing and re-writing cases, different interpretations and conclusions may be drawn (Chambers 1999). It's often common, and understandable, to focus on the most graphic and extreme cases, but this can demonstrably skew resulting policy (Boddington and Hogben 2006).

Note, too, that the way a case is described might block or permit our own moral insights. King David had an adulterous affair with Bathsheba, and, after she became pregnant, deliberately sent her husband, Uriah the Hittite, to his death in battle. The prophet Nathan described a case of a rich man taking a poor man's sheep; King David denounced the actions of the rich man in taking what little the poor man had. By packaging the essentials of King David's heinous acts in parable form, Nathan presented David with the uncomfortable truth: Thou Art the Man (2 Samuel 12) (Butler 1827; MacNaughton 1988).

Note that often it is a third party coming from an outside perspective who is best placed to do this; and someone who is prepared to speak truth to power. In order to keep an outside perspective on one's moral values, Philip Zimbardo argues for the ethical necessity of belonging to more than one social or peer group (Zimbardo 2008).

Codes and regulations may be developed in anticipation of possible problems. For example, the EPSRC Principles of Robotics may be seen as an attempt to avert the kind of public backlash that was seen in the UK over GM crops (Bryson 2012). Hence, such background issues are again important to understand in considering the purpose and final shape of any codes or regulations. The recent government documents such the House of Commons Science and Technology Report (Robotics and Artificial Intelligence 2016), the European Union's Committee on Legal Affairs Report on robotics (European Civil Law Rules in Robotics 2016), and a report by the Obama Whitehouse (Executive Office of the President 2016), are attempts to anticipate particular problems within particular political landscapes.

Codes of ethics may develop in response to other codes of ethics. This may or may not be appropriate. For example, codes of ethics for social science researchers have been historically modelled closely on codes of ethics for medical research. But the risks involved in social science research tend to be of a quite different kind, and of a different degree. Moreover, social science research methodologies may differ greatly from those of medical research (Atkinson 2009). The regulation of social science research has suffered in many respects from being shoehorned into a medical model. We need to think carefully about how AI, in its many different forms, needs to be regulated, rather than simply tinkering with what we already have, and rather than assuming that the same model of codes and regulations will do for all forms of AI.

Codes of ethics may be influenced by certain powerful groups or individuals. This can be useful, but there are drawbacks. Those working on public engagement have long recognised that there are multiple ‘publics’, and that different interest groups among the public may have quite different agendas. Some of these may be avidly pro-science (Novas and Rose 2000). Others may have very different views (Plows and Boddington 2006). Organised groups may be influential, perhaps unduly so; and membership may be skewed, with those who don’t join groups likely to have different opinions. Many patient groups may act as lobbyists, receiving funding from the pharmaceutical industry (Herxheimer 2003). Some codes or sets of principle themselves are of course produced by powerful groups, such as prominent members of a profession, for example the 1975 Asilomar Conference Recombinant DNA Molecules (Berg et al. 1975). One take on the groups producing such statements is that these are the ones who know what they are talking about. Another take is, yes, sure, but others need to have a say as well, and are likely to have very different interests. Yet another consideration is how such groups are selected, or self-selected, for influence and persuasion.

Codes of ethics of professional bodies also often have a wider national or international context. For example, codes of medical ethics for different countries exist in the wider context of the policies of the World Health Organisation. Codes of medical ethics are closely linked to the development of medical law in the relevant jurisdiction; and the development of medical law in separate jurisdictions is itself often influenced by developments and cases in other jurisdictions. Much research takes place in a global context. For example, much research in genomics of necessity needs to study different population groups of humans in order to conduct scientifically robust research. Complex ethical considerations of how to marry global standards with local sensitivities may be needed (HapMap 2004).

Codes of ethics also develop in relation to certain cultural contexts, and these may influence them in ways which are hard to discern, especially if we are also embedded within that context. The development of law, regulation and practice within different geographical areas in itself helps to shape this cultural context. For example, laws regarding the protection of privacy in the use of personal data within the EU are currently more stringent than in the US; this helps to shape debate and opinion, but it’s not clear that this difference in emphasis could have been predicted in advance. This again helps to illustrate how complex, interwoven, and perhaps unpredictable, are such developments in values.

Clues to influential cultural context may be found in literary devices such as the rhetoric used in surrounding text, and allusions made. In the context of technology in general, and AI in particular, reference to science fiction and to various stories regarding robots, computers, and out-of-control creations is frequently made. These may be instructive of underlying beliefs and values.

Science Fiction and Myth in Policy Making for Robotics

The European Parliament's Legal Affairs Commission have published a study on 'European Civil Law Rules in Robotics' (Directorate General for Internal Policies 2016). It illustrates how reference to myth, legend, science fiction and popular culture are routinely referred to in policy and ethics discussions regarding AI in general and robotics in particular. But note how the rhetorical reference to such stories can help shape the thinking that then goes to frame how the surrounding policy is read and understood. Here, 'Western' responses to robots are contrasted with those of the 'Far East':

1 Western fear of the robot

The common cultural heritage which feeds the Western collective conscience could mean that the idea of the "smart robot" prompts a negative reaction, hampering the development of the robotics industry. The influence that ancient Greek or Hebrew tales, particularly the myth of Golem, have had on society must not be underestimated. The romantic works of the nineteenth and twentieth centuries have often reworked these tales in order to illustrate the risks involved should humanity lose control over its own creations. Today, western fear of creations, in its more modern form projected against robots and artificial intelligence, could be exacerbated by a lack of understanding among European citizens and even fuelled by some media outlets.

This fear of robots is not felt in the Far East. After the Second World War, Japan saw the birth of Astro Boy, a manga series featuring a robotic creature, which instilled society with a very positive image of robots. Furthermore, according to the Japanese Shintoist vision of robots, they, like everything else, have a soul. Unlike in the West, robots are not seen as dangerous creations and naturally belong among humans. That is why South Korea, for example, thought very early on about developing legal and ethical considerations on robots, ultimately enshrining the "smart robot" in a law, amended most recently in 2016, entitled "Intelligent robots development and distribution promotion act". This defines the smart robot as a mechanical device which perceives its external environment, evaluates situations and moves by itself (Article 2(1)). The motion for a resolution [calling for 'the immediate creation of a legislative instrument governing robotics and artificial intelligence to anticipate scientific developments over a medium term', p 8] is therefore rooted in a similar scientific context. (p 10)

Here, the West is seen as having a negative attitude of fear towards robotics, and the document itself then expresses a fear of its own, that this may 'hamper the development of the robotics industry'. But note how Western attitudes are presented as emanating from tales and myths, framed only as 'ancient tales' and 'romantic works'. However, the positively presented Far Eastern attitudes are presented with a more solid underlying metaphysics or ideology—in Japan, that of Shintoism. This contrast then automatically

(continued)

frames the Western responses as shallower; subliminally, it's as if the West just scared itself with 'spooky stories'. The motion for a resolution proposed in this policy document is then nested in the positively framed Korean response to legal instruments regarding robots.

However, there are significant currents of Western thought and writings which could be used to explicate a response of fear to robots, which could provide a well-articulated and long established underlying basis of thought. For example, the influential creation story of Genesis presents a creator God as making autonomous beings—us—who quickly behave atrociously and disobey their Maker. Cain murdered his own brother, Abel, in a jealous rage; and on the story goes. So, it seems no mere coincidence that fears about the behaviour of autonomous robots would be strong in Western literature. The control problem of autonomous agents is also precisely a major concern of current attempts to address ethical issues in AI.

Note, too, that the EU report in fact goes on to reiterate that there is reason for Western concerns, 'now that the object of fear is no longer trapped in myths or fiction, but rooted in reality' (p 13) and cites Bill Gates, Elon Musk, Stephen Hawking, and Bill Joy as issuing warnings regarding autonomous AI. Note, however, that this subtly juxtaposes the recent reasoned warnings of scientific experts against the ancient myth-and-story driven fears of the populace, which perhaps by serendipity happen to coincide. This perhaps is not a useful way for a policy document to frame the concerns of 'experts' versus 'the public'. This is especially true given the way that AI does in fact raise questions pertinent to the very foundations of our morality and of our view of ourselves. The document then subtly priorities the expressed concerns of technological 'experts' while diminishing the concerns of 'the Western collective conscience'.

4.5 Can Codes of Ethics Make the Situation Worse? Yes

We've seen that codes of ethics need a strong institutional backing to function effectively. But codes of ethics can actually make matters worse.

Separation of ethics from 'life': The very idea of parcelling ethics into a formal 'code' can be dangerous, if it leads to the attitude that ethics itself is just some separate part of life and of activities. Such a risk exists if the code is presented as a set of instructions to user. 'Perceptions that a code presents the voice of an external authority frequently go along with a defensive and punitive institutional ethos that suggests to code users that it is necessary to lie low and keep out of trouble in order to avoid threats of criticism, negative judgement and punishment' (Bowden and Surma 2003) p 26.

'Can't someone else do it?' Homer Simpson once ran for Sanitation Commissioner of Springfield under this banner (Trash of the Titans 1998). It didn't go so

well. The existence of a code of ethics would be problematic if it encouraged the idea that it was *somebody else's job* to 'do the ethics'; although there can be good reasons to ensure that specific nominated individuals are assigned responsibility for certain issues, as a check against the *diffusion of responsibility* within organisations and looser groups. The appointment of a role such as a 'Chief Values Officer' *might* present such a danger, depending on how the role was implemented. The ways in which responsibility is avoided by individuals and diffused within institutions has been discussed in relation to the very large and often geographically dispersed groups of researchers that may be working on a project (Caulfield et al. 2008). Work in social psychology has turned up some valuable lessons in how easy it is to create the conditions which allow for diffusion of responsibility to occur (Zimbardo 2008). It would be very worthwhile for those considering the effectiveness of codes of ethics for AI, which may be developed by very large groups of people working on different aspects, to contemplate these problems.

And codes and regulations may encourage 'work to rule'—to work up to the regulation, up to the code, and no further; to the letter of the code, not the spirit. This may be especially problematic in some areas, such as those pertaining to safety. Well known examples of operating to the letter, not the spirit, include 'shopping around' for a tame ethics review board, or operating in countries where the standards are not so tight (Gulhati 2005).

So, a code of conduct might produce a 'tickbox' culture of ethical complacency where filling in and complying with paperwork becomes an end in itself, and the goal of ethical compliance is focused on too narrowly, and for the sake of reward or of avoiding penalties. For example, in some situations this may apply to the practice of obtaining informed consent to medical treatment and/or medical research, where the staff have the task of 'consenting the patient' (Jones 1999). Worse, such a mentality can encourage the very behaviour it was intended to discourage (Bazerman and Tenbrunsel 2011; Adams and Balfour 2014).

Indeed, the existence of a code of ethics or other systems of ethical guidance may give rise to ethical display behaviour or 'virtue signalling' (Bartholmew 2015). Easily signed declarations of ethical intent may have no impact and may entice signers to overinflate their self-assessed moral character (Bazerman and Tenbrunsel 2011). It's always worth remembering that in Stanley Milgram's classic Obedience to Authority experiments, where subjects were led to believe they were involved in an experiment on learning and that they were delivering electric shocks to 'learners' (actually stooges), Milgram found that expressing moral doubts enabled subjects to retain a self-concept as a 'good' person, and actually made it easier for them to continue to administer 'shocks' to the stooge (Milgram 1974).

Let me repeat that: *expressing a moral sentiment may in some circumstances decrease the likelihood of behaviour that follows one's conscience.*

There are indeed very hard questions about how to translate institutional ethical policies into practice. For instance, recent work on the 'paradox of meritocracy' shows that institutions which consciously flag meritocracy may in fact show greater bias towards men over women than those which do not (Castilla and Benard 2010), and ethics and HR policies which mandate the currently fashionable implicit bias training must face mounting evidence that such training may even make the situation worse (Duguid and Thomas-Hunt 2015). Any code of ethics which wishes to encourage such good behaviour thus needs to take careful heed of research and developments on the question of how best to bring about such changes.

The flip side of this is that where codes of ethics are *unduly* restrictive, there may be some justification in giving them short shrift. A code of ethics might worsen a situation by tying the hands of professionals whilst those outside the profession can carry on a practice with impunity. For example, restricting research into a particular area because of its dangers might mean reduced capacity to counter any dangers in that area from competent outsiders to a profession.

Ethical Arms Races and Being 'Too Good'

In the children's book, *Super Duper Jezebel*, the main character is a goody-two-shoes little girl who never breaks any rules (Ross 1988). One day a crocodile enters the school playground; refusing to break the rule against running at school, Jezebel alone comes to a sticky end. In a related vein, Hilaire Belloc's *Cautionary Tale* of the truly nauseatingly good child, Charles Augustus Fortesque, paints a comic yet starkly unattractive picture of an entire, unimaginative life lived according to blind conventional attachment to social rules (Belloc 1907).

There is a particular problem with slavish conformity to ethical rules in an unruly world full of 'bad guys'—suppose we render ourselves vulnerable to calamity by ideals of ethical purity that don't equip us to fight dirty if push comes to shove? It's also one thing to sacrifice yourself to a moral ideal, another entirely to sacrifice *others* to your ethical ideals.

Unreflective conformity to existing moral and social convention is particularly problematic where there is a need to address flaws in the existing conventions. This does not necessarily mean that we want everybody to be always questioning existing convention. But we need a dialogue with those who are raising concerns and pointing to shortcomings. There's a particular problem where the wish to conform to otherwise laudable rules makes us powerless in the face of those who flout the rules. As we've seen, OpenAI in fact specifically aims to combat such a problems by producing open source AI, in the hope that this will help to undermine the potential dangers from those creating malicious code (OpenAI).

The very real problem of how to avoid an arms race of autonomous weapons is mentioned here but it would be foolish in a book of this generality and length to attempt to do anything more than point out its difficulty (Roff 2014).

And while we're on this topic: experiments such as Milgram's, and others such as the Stanford Prison Experiment, produced insights into human moral

(continued)

behaviour, yet, therefore perhaps inevitably, had dismaying effects upon the subjects. Philip Zimbardo's Stanford Prison Experiment randomly assigned students, who had all been screened for good mental health, to the role of either 'prisoner' or 'guard'. Many of the 'guards' then meted out harsh treatment to the 'prisoners'. Tellingly, the experiment was called off early only after Christina Maslach, who was not directly involved, observed what was happening and protested. Zimbardo had the sense to marry her (Zimbardo 2008). We've seen how extrapolating from catastrophe can have problematic results as well as good. Such experiments have fed into ethical regulation of social science research which now makes it virtually impossible that such experiments could today get ethics clearance. Yet, these experiments helped to give us valuable insights into unethical behaviour. This is a particularly perplexing conundrum.

Keep calm, dear: There is also a danger that a code of ethics might actually be serving the purpose of calming public anxiety, without actually managing to make an iota of difference to the substance of warranted public concerns. Ethical regulation of new technology might serve to placate concerned groups and individuals and the very existence of regulation around a new technology or practice might make the unpalatable seem more palatable (Bryson 2012; Dingwall 2008).

'Administrative Evil'

Work such as the book *Unmasking Administrative Evil*, which looks closely at the root sources of many technological catastrophes and institutional failings, is extremely pertinent to the consideration of developing codes of ethics for AI (Adams and Balfour 2014).

The failure to apply policy correctly can be a big problem, especially where this failure emanates from pervasive institutional and leadership failings.

But in some cases, it is the very *application of policy* that can inadvertently give rise to deleterious effects, sometimes effects precisely opposing the intent of the policy.

This concerns a 'technical rationality' and how value issues can get lost within large systems. The dangers increase the greater the efficiency of the system and the greater its automation and distance from (uncorrupted) human affective response.

However, it is perhaps not beyond the wit of those studying autonomous systems to consider how to combat such potential for administrative evil. It is a recommendation that such a possibility is studied closely by those drawing up codes of ethics for AI within organisations and systems.

Regulations and boiling frogs: And, since codes of ethics are developed over time, what's seen as problematic is gradually changed or eroded, so that practices can be introduced little by little. This is often known as the 'boiling frogs' problem, although the jury's still out on whether actual frogs do this. (Don't try it out. You'll never get ethics clearance for the experiment.) For example, as recently as 30 years ago, the consensus in the UK at least was that we would never even consider sex selection of human embryos for social reasons. Now, whether for good or for ill, precisely this question has been considered (Holm 2004). Is this progress? Or the reverse? Or is this simply change, and nothing more? Importantly, in the development of regulations in emerging fields, one might expect that changes would be made as our understanding of the technologies changed. But it's common to draw 'lines in the sand' that will 'never be crossed'. Yet, such 'lines in the sand' often do then get put up for further consideration. Is there then no such thing as a 'line in the sand'? Often we are nudged into considering blurring policy lines by the directions of the technology.

Now, this might on the one hand be a genuine acceptance of the new; perhaps with a realisation that it hasn't brought the feared changes, or with a reappraisal of what counts as a plus or a minus. But it could also be step by step introduction of changes which in the end add up to a large change which would never have been accepted, had it been introduced all at once. After all, that is precisely how Milgram got decent individuals to deliver electric 'shocks' to strangers—little by little (Milgram 1974). This will be particularly hard to assess where rapidly developing technology that is embedded in our lives in multiple ways is having a large impact upon how we live.

Oh wait. That's happening with AI.