Chapter 2
What Do We Need to Understand About Ethics?

Abstract Consideration of ethical questions in AI requires an understanding of some central questions and ideas in ethics. This chapter provides an introduction to ethics which will be used as a basis for further explanation of the particular questions about ethics in AI. Ethics is sometimes seen entirely negatively as restricting developments, but can also be used more positively as assisting in the promotion of beneficial activities. Standard normative ethical theories are outlined, but the focus here is on spelling out underlying questions in ethics. We need to understand that there are diverse accounts of the root need for ethics, questions about the nature of ethical concerns, and questions about who, or what, is the proper object of our moral concern, all of which need to be addressed in thinking about AI. There are also contentious questions about the nature of argument and justification in ethics, including questions about moral relativism, which are especially pertinent to the issue of developing codes of ethics, and which we will need to consider carefully. The issue of transparency in ethics parallels concerns with transparency in AI. Questions about the nature of moral agency and moral motivation are also of prime relevance to discussions of AI.

From reading much of the literature on AI and ethics, and from taking part in many hours of discussions with a range of people from a variety of disciplinary backgrounds, I've realised more and more that there are some questions and issues in ethics which are omnipresent in many of these discussions, but which are not always articulated.

It is a central contention of this book that developments in AI require that we consider and perhaps reconsider some fundamental questions in ethics.

It's obviously impossible to present a full characterisation of ethics here. There is disagreement among philosophers on every one of the issues we will discuss. The points raised are pertinent to codes of ethics for AI; to considering some of the central ethical questions of AI more generally; as well to as the thorny question of whether or not, and how, we can build ethics into machine behaviour.

P. Boddington, *Towards a Code of Ethics for Artificial Intelligence*,
Artificial Intelligence: Foundations, Theory, and Algorithms,
DOI 10.1007/978-3-319-60648-4_2

2.1 A Preliminary Plea: Ethics Is Not About 'Banning' Things

Very often, talk of 'ethics' and in particular 'ethical regulation', conjures up the idea that 'ethics' is simply out to stop activity, prohibit or mandate various actions. In some circles, the word 'ethics' has attained negative connotations (Bowie 2009). Indeed, some 'ethical' regulation can with some justification be found guilty of excessively hampering valuable research—and to this extent then, '*un*ethical' (Atkinson 2009). We'll directly consider later the possible negative impacts of codes of ethics. But this 'spoilsport' notion of ethics is limited. Ethics can and should be seen more positively as helping to promote or enhance an activity.

Note that we may recognise that an activity merits our attention and requires ethical discussion, without deciding in advance that this means it's going to turn out to be problematic. We need to be aware of how changes are impacting on our values. Self-awareness, both as individuals and as societies, is itself of value. In considering ethics in the context of artificial intelligence, amidst talk of the possibility or otherwise of self-aware machines, we must here of all places recognise its value.

2.2 Normative Ethical Theories

Many accounts of practical ethical questions will start off with a broad characterisation of different normative ethical theories. These are accounts of how to act; in other words, theories about the basis for making decisions in ethics. The three most commonly outlined theories are:

Consequentialist theories, which broadly claim that the right action is the one that brings about the best consequences. This is most commonly held as some form of utilitarianism, which aims to bring about the greatest balance of happiness over unhappiness, or pleasure over pain, for the largest number of people.

Deontological theories, which claim that what matters is whether an action is of the right kind, that is, whether it is in accordance with some general overarching principle, or with a set of principles, such as 'do not take innocent life', 'do not lie', and so on.

Virtue ethics, which focuses of the character of the ideal moral agent, and describes the range of different virtues such an agent has, and, broadly, claims that the right thing to do in any given situation is to do what the fully virtuous person would do.

There is much that can be said about these theories, their differing interpretations, and the vexed question of how to 'apply' theory to practice in ethics. However, normative ethical theories will not be our focus in this book. Important elements of morality which lie behind and outside these theories need to be examined to gain a fuller appreciation of the ethical challenges of AI.

2.3 Ethics and Empirical Evidence

Ethics deals with normative issues; it is not purely descriptive of empirical reality. Normative issues are ones we feel have a certain weight and import, although it’s surprisingly hard to characterise precisely what the weight and import of ethical issues are, and there is philosophical disagreement about whether ethical issues should always override other considerations.

The normative nature of ethics means that simply describing the way people act will not give an account of ethical action. Ethics requires discrimination between ways of acting and of being. Nonetheless, empirical questions about how we do think and act, and the possibilities of human psychology and society may be relevant to any consideration of ethics, for a variety of reasons. As the philosopher Kant observed, ‘Ought implies can’—we can’t require an individual person, humans in general, or indeed, machines, to do something that they *cannot* do (Kant 1998). We need to know what’s possible for human action, what might be effective strategies for assisting with obstacles to moral judgement and action, what effects there might be on human health and wellbeing of various possible policies, what pitfalls of action and judgement await us as we strive to think and act for the good, and so on.

What this means for AI: We need to think carefully about what relevant empirical evidence we have to collect to assess the impact of AI. This is harder than it might seem, and for interesting reasons. The evidence we need to consider is about the impact upon complex, feeling, living beings, immersed in sophisticated, dynamic cultures; it’s about human beings who only partly understand themselves, and who only partly understand their own cultures and societies. It’s about untangling what appears to be the case, and what is the case. AI, now and in the future, is deeply embedded within other technologies and with social practices; so measuring impact and attributing it to AI will be extremely challenging.

2.4 So Why Do We Even Need Ethics?

It’s worth pondering this, for there are different answers. Often these answers are strongly shaped by the disciplinary background of the questioner, be it sociology, anthropology, evolutionary biology, philosophy. Again, the aim here is not to produce ‘an answer’, but to indicate that whatever answer is given, it will reveal issues of central relevance to questions of ethics and AI.

One broad brush answer is that ethics exists because the world is not perfect, and we think we could improve it if we tried hard enough. But if this were the only ethical problem, then we’d simply need to sort out how to improve the world, and then, improve it. Simple! There are at least two further problems.

One, the world is imperfect *in a really complicated way*. It’s often hard to work out what *precisely* is wrong, let alone have a clear idea of what to do about it.

And two, *we* are not perfect, whether as individuals, or as groups. Even when we know what do to, we don't always do it—there is a problem with moral motivation. We lament with Rodney King, 'Why can't we all just get along?' And note, we ourselves often think we personally could have done better. We have some idea of what St Augustine meant when he prayed, "Grant me chastity and continence, but not yet." (Augustine 2014)

Many philosophers have considered that we need morality because things are 'inherently such that things are liable to go very badly' (Warnock 1971), and that we can't sort this out if left entirely to our own individual whim. Morality is a 'device for counteracting limited sympathies' (Mackie 1977). But even among those thinkers broadly in this tradition of 'double deficit' where both we and the world are broken, there are significant disagreements. For instance, not all agree on what particular shortcomings we have as humans. Some focus on problems with reasoning, some on our emotional responses. Some consider that if only we fix bias, we'll do the right thing. Some consider the end result of fixing bias must be some kind of equity. Some are idealistic utopians about human perfectibility. Some consider that the price of civilisation will always be a certain amount of discontent (Wiseman 2016; Freud 2002). And so on.

Morality as a Solution to Competition for Scarce Resources: What's AI Got to Do With It?

It would make a rather interesting project of its own to consider how different models of the function of morality, and the concomitant picture of human nature and the world, interacted with the development of AI.

But to illustrate, and to see how deep questions about AI and ethics go:

Suppose you consider that we need morality to combat our bias towards ourselves and our kin, given that there is scare competition for resources in the world and these need to be shared with some measure of fairness. Then, we usher in a glorious future of advanced AI.

We'd still be biased, of course. So do we outsource our ethical judgements to AI? Note the precise details of how we do this will depend not just on how we understand our own biases, but also on how we understand the ultimate goals of morality.

And even if we do this, why would we obey the AI, given our shortcomings in moral motivation? So, should we tie ourselves in to being forced to obey the AI? Should we go for individual enhancement via AI to combat this bias, so each of us is morally 'corrected'? In which case, we no longer have the same picture of the need for morality.

And what is the point of AI if it can't solve the problem of scare resources? So now we live in abundance. But abundance of what? Material goods, perhaps; but what do we do all day? Many scenarios foretell mass unemployment; goods aplenty, jobs scarce. If morality combats a problem of resource

(continued)

distribution, and the resources which are scarce change, we might need quite different moral tricks up our sleeves to address the rather different challenges of plenty.

Another case: suppose we say the task of morality is to make sure that each person lives a decent life, despite scarcity in the world and human shortcomings.

Or suppose we say the task of morality is to make sure that there is as little suffering in the world as possible, despite scarcity in the world and human shortcomings.

The former implies that AI should be geared towards making sure that those rendered unemployed by machinery all have good lives, suited to their individual situations. It might even double back on the use of AI to prevent individual misery.

The latter leaves it wide open that AI might be geared towards trying to ease out of existence the class of people who don't cope well with AI induced redundancy, whether by a programme of eugenics or of enhancement.

Note that an account of ethics will explicitly or implicitly rest upon underlying views of moral agents—us—and of our place in the world. It will implicitly rest upon underlying views of the value and nature of that world. It will implicitly rest on views of the relationship between us, as moral agents, and other moral agents, and the rest of the world. It will rest on an account of what inclines humans to behave badly, and what enables them to behave well. It will rest upon assumptions about how good a job we can do of perfecting 'human nature' and the world. Such underlying issues will surface, at some point, in discussions of codes of ethics for AI. They may be in disguise. But they will be there.

What this means for AI: The take home message is that understanding ethics means understanding moral agency. And how we understand human agency in particular, and agency in general, is a critical question in AI.

2.5 So, With What Sort of Issues Is Ethics Concerned?

Let's start with a popular answer to this. Ethics concerns important questions of welfare and harm, or if you prefer, pain and happiness, along with important questions of justice and fairness. The questions of justice and fairness bring with them questions about balancing the interests of individuals and groups. These questions tend to predominate many formal academic discussions of ethics, but there are other values which are important to recognise, such as the value of loyalty, of (justified) respect for authority, and ideas that relate to some notion of sanctity or purity—drawing what are seen as proper boundaries between different elements of our world (Haidt 2013).

It's easy to get an intuitive handle on many of the core ethical values, but very hard to specify them in detail without running against problems. Let's take an example: human health. This seems like a sound moral goal to pursue. But it turns out to be impossible to characterise health without addressing many other value issues. Should we have as a goal of human health, the maximal extension of human life, the postponement of death for as long as possible? Yet some would consider that there is a 'natural' termination to human life, others not. Should we extend the life of someone with such advanced dementia that their personality is no longer apparent? Addressing such a question involves asking and answering questions about the nature of the human person over time, and questions about how one person relates to their past and future selves, and to other people. These questions then rest upon accounts of human nature, human agency, what it is we value about life, and about personhood. And such questions come to the fore in many questions involving the development and application of AI.

Likewise, consider the fundamental question of whether we should aim at maximising happiness, in the sense of maximising pleasure, in humans. On many views, this gives an impoverished account of what human life should be about. Surely we want to do more than sit around with the pleasure centres of our brains firing away? Or is this really what we do actually want? Do we then need to address questions of the meaning of human life? Of its point?

There's also the question of where the boundaries lie between questions of ethical value, and other sorts of value, such as aesthetic value, and political questions. In drilling down to fine detail, there will be substantial questions raised. For example, how does the value of equality play out in relation to the complex and heated debates about what behaviour does and does not count as 'sexist'? Translation of values into the behaviour of AI has already raised many detailed questions of interpretation, such as the question of how Siri responds to sexist 'banter' (Fessler 2017).

What this means for AI: These questions turn out to be utterly crucial in considering the replacement of human activity, whether in whole or part, by machines. To that extent, these are questions already raised by mechanisation, but the developments of AI heighten our concerns. We will discuss these issues later.

2.6 Who (or What) Is The Proper Object of Moral Concerns, and How Widely Should Our Concerns Extend?

It's easy to assume that ethics must have universal reach (however this is defined), and that a sound ethic has to reach beyond individual, tribal or group concerns. It's commonly held that everyone shares a universal ethic, but this is demonstrably false. The views of Aristotle are particularly influential among many moral philosophers currently, but he did not take a universal view of ethics, distinguishing not just between men and women, free man and slave, Athenian and barbarian, but also held that one had significant duties to one's parents and one's children, yet no particular duties towards grandchildren (Aristotle 1999). Many actual systems of ethics have different rules of behaviour for different classes of people.

Moreover, even for those who hold that moral demands apply universally, there's the question of who counts morally: humans as a species; or persons, a class which may include some who are not human and exclude some who are; or any creature that is capable of suffering; or wider still, as some environmental philosophers argue? The philosopher Immanuel Kant held that moral concern should extend equally to all rational beings, and that would apply to rational creatures from other planets. He might or might not then have added that it could apply perhaps to some forms of AI (Kant 1972).

What this means for AI: Could we ever have moral obligations to sophisticated artificial intelligence? This depends on the basis for our moral obligation, and for why others—other creatures, other machines—have value. On some views of human and the human brain, we are pretty much like calculating machines, like computers, with various goals built in. On such a view, it's then more feasible that we might build AI which, like us, has moral standing, and can act as a moral agent. But others hotly dispute the initial premise that this is a good view of what humans are like. These are not questions extraneous to ethics. These are questions which underpin any account of ethics we might have. Hence, the question of who merits our concern, has large ramifications for considering ethics in AI.

2.7 Four Domains of Ethics: Self, Friend, Stranger, World

For some, ethics is essentially about how we treat other people. The view that it's all about combating disparate interests under a condition of scarcity suggests this. Such views may posit an egoistic motivation, and often assume self-interest: we can act in any way we like, so long as we don't harm others, and it's assumed either that we always act in our own interests, or that it's none of anyone else's business, and of no moral import, if we don't.

But on other views, we may have ethical responsibilities towards ourselves. If you're someone to whom this seems counter-intuitive, simply ask if it's okay voluntarily to get rigged up to a machine that stimulates your pleasure centres, rather than actually acting in the world. This seems abhorrent to many people and a travesty of a good life; it may even seem a morally wrong waste of a life. Others of course, beg to disagree. This debate can be very polarised; I've noticed that those few undergraduates who tough it out and insist that they'd be rigged up to the pleasure machine often find others respond with horror.

What this means for AI: Note that the potential of AI to powerfully transform our sources of choice, value and pleasure raises issues which come very close to these concerns. Whatever your own views, and even if you reject this idea, failure to appreciate that others do not will limit understanding and debate.

The question in ethics of how we treat others can be usefully split into questions about how we treat others known to us and within our circle of everyday concern, and how we treat more distant strangers; the second person, and the third person. Briefly, although these may be collapsed into the dichotomy between self and other, they tend to arise in different ways and tend to need different approaches.

What this means for AI: In AI, the former set of questions may concern how we are changing how we relate to friends, family and colleagues through AI-mediated technology, and how we interact with robots; the latter, by questions such as the societal impact of technology, employment, taxation, discrimination in the use of algorithms. Indeed, the prospects of human extinction at the hands of AI raises questions of a different complexion again.

And the question of what if anything we owe the non-human world is raised in AI. For we are changing the world, AI will hasten these changes, and hence, we'd better have an idea of what changes count as good and what count as bad. Again, failure to appreciate the range of views on this question will limit debates.

2.8 What Counts as Adequate Justification and Argument in Ethics?

Consider the contestable nature of moral justification: We start from premises of uncertainty. We know that there is disagreement on questions of ethical value. Should we pursue happiness as the sole value? Should we care for all others equally? And so on. And we also know—by a quick perusal of philosophers' debates—that there is disagreement on the nature of justification in ethics, and what would count as a good ethical argument.

But the weight of ethical concerns means we can't simply put these questions in the 'too hard' basket. And it seems to be part of the nature of moral issues to require justification. The question 'why' always seems appropriate, especially if it comes from those affected by decisions. So, where ethical questions are concerned, we just have to solider on, somehow.

Moral Foundations Theory

Here are some related problems:

How to construct a code of ethics for AI, given that at least some of this AI will have global reach;

How to construct a code of ethics for AI that will be largely acceptable internationally;

How to embed ethical decision making and agency into AI: what ethics do we chose to embed in the machine, given variation in ethics cross culturally, and indeed, within societies?

(continued)

One way into examining these questions is to start from research into how people across the world actually do think about values. This in itself won't be enough, since ethics is normative, not simply descriptive, as explained earlier: it's no good pleading, 'but in some areas of London, gang culture holds that raping a rival gang leader's sister is a viable form of revenge, and they're increasingly finding that acid attacks are a cheap and handy response to insult', and leaving things at that.

Moral Foundations Theory is research that aims to understand what lies behind the variations in morality around the world (http://moralfoundations.org/). Researchers have probed moral views and claim that behind variation lies concerns that can be grouped in five or six main headings: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, and perhaps additionally, liberty/oppression.

There will be societal and also individual personality variations in the emphasis given to these values. So note, that this has not found that 'really' humans have the 'same' morality at base. But it does indicate that there are common values, even if the emphasis is placed differently on these by some communities, belief systems, and individuals.

One important take-home lesson from this: understanding different ways of approaching ethical questions is the first step to seeing opposing points of view, and is a promising way to open dialogue with others.

2.8.1 How Do We Gain Moral Knowledge?

For some philosophers, this consists in gaining an appreciation of an independent moral reality. For others, it involves setting out one's moral goals (for example, the goal of maximising happiness and minimising pain) and then gaining the empirical knowledge to work out how best to do this in any given situation. For others still, morality is based not on objective reasons, but on subjective emotions. Such an approach will still be interested in conducting empirical inquiries, but these will be asking quite different questions. Others consider that we can best work out to do by considering the response of an 'ideal observer'. But is this someone stripped of all bias, of all emotion? Or someone who can see all biases, understand all emotions, and take them into account?

In drawing up codes of ethics for AI we need to assume certain broadly accepted notions concerning ethics. We can't just start from scratch. But it may be that in the very throes of discussing and implementing AI that some of the deepest disagreements about fundamental ethical issues bubble to the surface. We also need to consider how we can come up with the best, the most robust, the most workable set of guides and principles, given various disagreements about ethics, that pragmatically will attain assent and actually have a positive impact on action and outcome. And we need to think about how the process of arguing and debating all this needs to proceed.

2.8.2 The Elimination of ‘Bias’

One of the first things people think about in improving moral arguments is the question of bias. It’s now quite rightly routine, for instance, that conflicts of interest must be declared by participants in debate.

Eliminating bias in arguments seems an obvious goal, and some indeed hold out the hope that AI might help us to eliminate bias in ethical decision making. But what is ‘irrelevant’ bias? It can’t simply be the presence of emotion, since (even if the views of those moral philosophers who place the basis of ethics in our felt responses to situations are rejected), in ethical judgement, it’s often emotionally charged responses like empathy that help us to see what the moral issues are, and notice who’s affected. Neither can it be any *simple* account of partisanship to one group, since a certain bias towards those who are suffering the most may be morally justified.

What this means for AI: This question is vitally important for the issue of who is involved in developing and implementing codes of ethics, as well as for projects to embed ethical decisions into machines. ‘Getting rid of bias’ may be a great goal, but to understand what it means, and how to do it, is another matter.

Algorithmic Bias in AI

There is increasing awareness that algorithms used to facilitate various operations can reproduce or create bias. This may be because the training data sets for the algorithms are themselves biased in some way, or because the operation of the algorithm itself creates bias. This will be an especially difficult issue where the AI involved lacks transparency.

But what is bias, anyway? Recruiters are going to favour the competent, other things being equal. Is this bias, if certain groups in society are less represented in the group picked out as most competent? There are a whole host of legal, political, sociological and moral arguments to be had here.

Okay, so we can at least start with the bias that law requires us to eliminate. But that’s hard too. Here’s just one of many potential problems.

A ruling by the European Court of Justice in 2011 has required that in order to eliminate the bias of gender discrimination in setting insurance policy rates, insurers must not give lower premiums to female drivers, (nor give men better pensions in view of their shorter life spans) (Kuschke 2012). But statistics show that women are in fact on average less likely to have motor vehicle accidents. Insurance works on precisely assessing risk. So any machine learning algorithm trying to work out premiums is going to end up finding proxies for gender. But, discriminating against a group indirectly through the use of proxies for a protected characteristic is also against the law. Ways to try to circumvent this problem include more and more personalised insurance calculations, such as reductions in premiums for drivers who install devices in their vehicles to track how well they are driving.

(continued)

But insurance is pooled risk. The end trajectory of highly personalised insurance premiums could be the end of insurance as we know it. Some people will have extremely small premiums, and some could well be priced off the road. Statistically, these will be disproportionately males, members of the very class who'd originally benefited from legal protection from discrimination with respect to motor insurance. However, on the bright side, pricing accident prone drivers off the road might be a relief to other road users.

Interestingly, one could point out that it's the very efficiency of the algorithms which has alerted us to the inherent difficulties created by changes in the law.

2.8.3 When Is Ethical Justification 'Finished'?

Ethical questions are often so complex that it's hard to make our answers exactly precise. But is there always a 'right answer'? Or are there some genuine moral dilemmas, where, whatever we do, there is some moral cost? It may be that we are sometimes faced with situations where different moral values clash, where they're *incommensurable*.

Why Is this relevant for AI? Where rapid technological and societal change is occurring which affects our relationships with each other and with the world, many of our values will be in flux. This makes it all the more likely that we won't have a fully worked out, coherent and consistent set of values. It's better to recognise this than to chase a false consistency. Witness current debates about privacy, an issue of particular concern in AI, where attitudes have developed significantly in relation to the use of technology, vary greatly depending on the context, and are also arguably internally inconsistent for many individuals (Nissenbaum 2004, 2010). An individual may value privacy in one area, while posting indiscrete personal information all over social media, and may see some data collection as routine, other data collection as a violation, but may lack consistent reasons for these distinctions.

Midnight Anguish and Slow Torment in Moral Reasoning

Especially where it's particularly hard to know what to do, and all the options have some pluses and some minuses, it's often noticeable that the subjective, felt quality of the decision making process is sometimes flagged as sort of place-holder for moral justification. 'Finding a decision particularly difficult to make' is sometimes accepted as a proxy for making a good decision. Watch out for this. It may or may not be something to worry about.

An example can be found in a report of an interview with Elon Musk and Sam Altman regarding the launch of OpenAI. This comes from a magazine write-up, so it's doubtless an incomplete account of Musk and Altman's own

(continued)

views of the matter; the example is meant simply to demonstrate the seductive idea that effort and difficulty indicates moral sincerity.

Interviewed on announcing the launch of OpenAI in December 2015:

Stephen Levy: I want to return to the idea that by sharing AI, we might not suffer the worst of its negative consequences. Isn't there a risk that by making it more available, you'll be increasing the potential dangers?

Altman: *I wish I could count the hours that I have spent with Elon debating this topic and with others as well and I am still not a hundred percent certain.* You can never be a hundred percent certain, right? But play out the different scenarios. Security through secrecy on technology has just not worked very often. If only one person gets to have it, how do you decide if that should be Google or the U.S. government or the Chinese government or ISIS or who? There are lots of bad humans in the world and yet humanity has continued to thrive. However, what would happen if one of those humans were a billion times more powerful than another human?

Musk: I think the best defense against the misuse of AI is to empower as many people as possible to have AI. If everyone has AI powers, then there's not any one person or a small set of individuals who can have AI superpower.

[(Levy 2015) (Emphases added.)]

The interview is interesting in many ways. There is an admission of uncertainty about whether OpenAI might increase the dangers of AI. But note how the opening proviso by Altman about the difficulty of the decision seems intended to provide assurance. Note, too, that this prolonged debate was said to take place between just two main people and an unspecified number of unknown others.

And note, too: There is however, a serious question to consider about what we are looking for in our moral decision making. In the context of AI, which focuses on speed, and which may operate using black boxes which no one fully understands, the reference by none other than Sam Altman to the slowness and difficulty of an ethical decision as markers of its probity, is telling. How machines operate, and how humans demonstrate the sincerity and integrity of their moral decision-making are poles apart on this account. Work on the psychology of time and decision making shows how different perspectives on the present and the future can affect conclusions and sometime distort judgements (Zimbardo and Boyd 2009).

2.8.4 Can We Necessarily Even Fully Articulate All Our Key Values?

Given the complexity and the importance of ethical questions, and given the social and technological changes being brought in by AI, it's highly likely that there are some profound values at play that we may find hard to articulate. We need to balance the demand to make our moral reasoning as robust as possible, with safeguarding against making it too rigid and throwing the moral baby out with the bathwater by rejecting anything we can't immediately explain. This point is highly relevant both to drawing up codes of ethics, and to the attempts to implement ethical reasoning in machines.

There is a good reason why we might not be able to articulate fully our most deeply held ethical responses. These may be more like the procedural memory we have for the deeply learned, automatic things we do each day that are driven into the fabric of our lives. There is a tendency among some philosophers to insist that the considered, articulated, coherent responses are the best, or the only ones allowable. But we would not dismiss as a fraud a concert pianist who could not explain precisely how their feats of virtuosity were achieved, finger movement by minute finger movement. Something similar might be occurring in our everyday and rapid moral reasoning.

And note it's our most fundamental values that are often hardest to articulate, for precisely the reason that these are the values from which we *start* articulation. The US Declaration of Independence (July 4th, 1776) states 'We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.' Note the necessity of stating the self-evidence of these claims; this is a declaration of *faith*. No deeper ground of justification can be given. *"If I have exhausted the justifications, I have reached bedrock and my spade is turned. Then I am inclined to say: This is simply what I do."* (Wittgenstein 1973).

What's this got to do with AI? When we are trying to ensure that machines keep to our values, when we are trying to articulate those values in times of profound technological and societal change, we both need to be able to spell them out as rigorously as possible; but at the same time be aware, that the inability to do so may not mean that there is nothing of value there to be grasped.

2.8.5 Can There Be Such a Thing as Moral Progress?

There are various answers to this, from the optimistic answers of the utilitarian and social reformer John Stuart Mill (1863), to more pessimistic answers from those who see history as moving in cycles or just randomly. One lurking danger is a view that change is *ipso facto* change for the better.

What's this got to do with AI? In the context of AI and of technological change, one view is to see technological change as inevitable, and something we must adjust to but cannot realistically halt. It's useful to consider one's own assumptions about moral progress and social change. Excitement about AI often includes calls for its use in human enhancement. But in order to understand that something counts as enhancement in this context, we need to have a clear idea about what the desired end result is—and as should be apparent by now, that's still on *homo sapiens*' 'to do' list. Assessing and advancing moral progress, whether in individuals or in humans as a group, is highly complex (Wiseman 2016).

Transparency in Ethics and in AI—'What Plato Did'

Transparency in ethics has at least three aspects.

One is *visibility to others*. If others can see what you are doing, it makes it more likely you'll behave well. Philosophers have long known this. In Plato's *Republic*, Protagoras considered the Ring of Gyges, which magically renders its wearer invisible. Possessed of this, Protagoras argued, one would of course commit all manner of wrong-doing (Plato 1974). Conversely, much recent research lends support to the view that even *imagined* scrutiny by others helps us do the right thing (Zimbardo 2008).

The second is *comprehensibility to others*. Ethics demands a shared system of justification. In the *Republic*, Plato infamously argued that those in the top rung of society, the Philosopher Kings, dubbed the 'gold', had a grasp of moral truths but that the lower orders, or those dubbed the 'silver' and 'bronze' in society, were incapable of full access such knowledge.

And a related aspect is *accountability to others*. A corollary of Plato's views on knowledge and government is that, in governing those under them, the 'noble lie' could be justified to keep the *hoi polloi* in order. I take it that a view is abhorrent in any democratic society. It goes without saying that you can't claim to be adequately addressing ethical questions, if you refuse to explain yourself to rightly interested parties. Of course there will often then be a further question about who such parties are and what claims they have on you.

What this means in AI:

Firstly, The very complexity of much of AI means that there is often a particular question of transparency. If even its creators don't know precisely how an algorithm produced by machine learning is operating, how do we know if it's operating ethically or not? The frequently posed fears that without our knowledge we might be manipulated by powerful machines or very powerful corporations armed to the teeth with the opaque machinations of AI, gives a modern take on the Ring of Gyges myth. Only, now it's not actually a myth.

Having specialist knowledge, as professionals in AI have, does not entitle you to 'lie' to the people, nor to be in sole charge of questions that concern them; quite the reverse. Such specialist knowledge should mandate a duty to explain.

However, the question of how much transparency is legitimate in respect to certain activities is an open question. Only a fool wants the security services of their country to be fully transparent given the existence of real enemies; nonetheless drawing the line may be hard. Commercial companies also have reasons for secrecy. Which brings us on to the next point:

Secondly, there are many powerful actors involved in AI whose activities may affect billions of others; perhaps then, in some ways, a technological

(continued)

elite with access to arcane knowledge—AI professionals—are the new ‘Philosopher Kings’. How they handle ethics, how they explain themselves, and whether they manage any system of accountability and dialogue, will be critical to any claim they might make to be truly concerned with ethics.

Some Notes on Disgust

We want our ethical arguments to be rigorous and we want them to be complete. But these aims may be in tension.

Some argue that some responses to moral issues are simply emotional reactions based upon what has been called the ‘yuk’ factor: an automatic response of disgust to an issue (Edmonds and Warburton 2010). This may seem like the reaction of someone uneducated in restraining their thoughts and submitting them to the test of reason. Those who warn against basing views on ‘disgust’ may find support in experimental evidence indicating that manipulating disgust responses can alter moral judgements (Haidt 2013).

But, psychologists have argued that disgust reactions track a kind of immune response to protecting the self, the community and its boundaries. Disgust reactions are linked to notions of sanctity or purity. Work on the range of values in moral psychology shows that those with certain political views (broadly, liberals) tend to focus on a narrower range of moral values than those with opposing views (broadly, conservatives). The latter include values of sanctity and purity which may result in responses of disgust (Schnall et al. 2008).

But, it is among those philosophers who themselves tend to argue for a narrower range of values (autonomy, welfare, justice, for example) that the arguments for eliminating considerations of disgust (and dignity) can generally be found. So, are these philosophers simply more rigorous in their quest for moral justification? Or are they more limited in their appreciation of a range of values?

What’s this got to do with AI?

One: Some of the ethical questions in AI concern how we should delineate the boundaries between humans and machines. So, we should expect that some responses to some possibilities will involve disgust (for example, calls for the development of post-human cyborgs).

Two: Since we know that different groups of people see such reactions as relevant to ethics, or as irrelevant to ethics, this has implications for how we constitute our discussions of AI ethics.

Three: Interestingly, as mentioned, disgust responses are linked to notions of sanctity or purity. Those calling for the removal of consideration of disgust, or ‘woolly’ notions like human dignity, from discussions of ethics, are themselves, of course, exhibiting a variant of a call to purity.

2.9 Moral Relativism, Moral Justification and AI

How can justification of ethical arguments proceed, given that there is a large variety of moral systems and ethical beliefs, not just within a society and culture, but between different cultures?

Why is this an issue for AI? Because many forms of AI, by their very nature, affect people across societal boundaries. Because AI is predominantly being developed in certain parts of the world. Because AI, along with other technologies, is helping to connect individuals and groups from different social and cultural groups.

What should we do about it? Again, a book of this length cannot hope to answer the question. But we need to be aware of the questions. Communication, open dialogue and debate, and diversity in participation, go some way towards recognising the issues.

Note too that there are many responses one might take to moral diversity around the world. Recognising differences between cultures in moral codes, and valuing the contributions from a variety of cultures, has not led all to conclude that moral beliefs are simply relative to different societies.

Having Your Relativist Cake and Eating It: ***Not Such a Good Idea***

Here is a commonly expressed argument behind a particular view of moral relativism:

PREMISES: Morality is simply the expression of socially constructed value judgements. Other societies have their standards of judgement, we have ours.

CONCLUSION: Therefore, we should not judge other cultures.

Such a view is often motivated by the finest principles—concern not to condemn what we don't understand, and concern for power imbalances between the wealthier and the less wealthy. There are many examples where havoc was wrought by 'interfering' in other cultures. And we have much to learn from dialogue with others.

However, this view involves taking what the philosopher Bernard Williams once described as the 'mid-air' position (Williams 1976). The premises state that all value judgements only make sense *relative to a social system*. But the conclusion—a value judgement—is announced as if it is some *universal truth*.

But, if morality is *always and only relative to societies*, from what society do we judge that it's wrong to judge other societies? From some 'mid-air' position, outside of any culture, from which it is possible to pronounce universal truths? But . . . I thought you said all value judgements only make sense from within some society or other?

Moreover, such a simplistically sketched view may rest on an assumption of a series of isolated and homogenous societies which each contain their own autonomously created set of values. This is a greatly simplified view of the

(continued)

complex world we face today, and raises specifically difficult issues in regard to international issues. The possibilities that AI itself brings are indeed helping to further create and disrupt links between cultures and to disseminate information.

Furthermore, such crude cultural relativism tends to present individual societies as harmonious clubs where everyone agrees on the presiding values. But this is not true of many societies, and perhaps true of none. There are almost always some groups in society whose views are not adequately heard, and whose interests get short shrift. Moreover, taking notice of such people is of the very essence of ethics**.** So, this commonly held form of relativism may end up doing the reverse of what the often well-meaning people behind it wished to do—it may end up supporting the dominant views of the most powerful people in other cultures.

And note this complexity. The currently dominant views of morality in Western thought are universalist in nature. This is behind moves like the Universal Declaration of Human Rights. But, if 'our' morality is universalist in nature, then, from a relativist view of morality, who can argue that we should not be universalist? So, paradoxically, if we maintain a crude moral relativism, then there is reduced ground to argue against imperialist expansion or a global takeover of systems of AI.

What does this mean for AI? AI crosses national and cultural boundaries. We need to think about how we develop a robust ethic which addresses this without simply degenerating into a 'pick and mix' approach, where if someone else wants to use AI to instigate, say, the total surveillance of their population in an attempt to fine-tune brainwashing, we *simply* say, 'oh well, each to their own'. This is a crude example; the point is how hard it is to draw a line between praiseworthy respect for other cultures, and turning a blind eye to moral wrongs.

2.10 A Distributed Morality?

Note that calls to end bias, and many notions of justification in ethics, often rest upon an assumption that there is one thing that it's right to do, and that this is the same for all agents. But many argue, often on the basis of research in various branches of the social sciences as well as in philosophy, that morality is (at least sometimes) socially distributed, so that differently placed actors within a situation have different moral roles to play; and that *this is better than a 'homogenous' morality*. As ever, there are variations of detail in how a distributed morality might be understood (Floridi 2013; Floridi and Sanders 2004).

What does this mean for AI? It has of course implications for the responsibilities of individuals and teams in AI, for questions about autonomous systems, including systems involving both humans and machines, and for questions around building in ethics into intelligent machines.

2.11 Moral Agents

What is it to be a moral agent, what motivates us to act morally—and what prevents us from acting morally? On some ethical theories, all that matters is that the best result obtains. Such accounts are neutral with respect to agency; it doesn't matter who acts, so long as the job gets done. On others, agency matters, and matters crucially in a multidimensional way. Deontological and virtue ethics theories take such a line. This is a fascinating and complex area that has received intense debate and scrutiny. Here are a couple of pointers for why this matters for AI, and for developing codes of ethics in AI.

If agency does not matter, then we can outsource our moral decisions and actions to another competent person, or even to a machine. But on the most plausible views of ethics, the intention with which something is accomplished makes a difference to its moral assessment, and it matters who it is who is acting, and why they act as they do. Even consequentialists usually see the point of the questions they are asked about the place of agents in their account of ethics (Scheffler 1988).

So, what does this mean for AI? If our actions are mediated by a machine which lacks transparency in some respect, how do we ensure that they are ethical? Suppose I used an algorithm designed by machine learning to make a policy decision. How can I be held accountable for decisions made in such a way? On some views of ethics, well, never mind, so long as the outcome is okay. On others—not so fast.

But note that addressing such highly complex questions can mean examining the basis of claims of agency and autonomy, in ourselves as well as in machines. And in part, much work on the development of intelligence and agency in machines is examining the nature of intelligence and agency in humans. This means that we might perhaps upset the philosophical applecart on which certain views of ethics rest. For instance, are we using ideas of moral agency which assume humans have free will?

This debate is far too interesting to pursue in great detail in this little book. But, note how again, how deeply questions about AI go when we think of them alongside questions of ethics. In drawing up codes of ethics for AI, it will be important to examine what assumptions are being made about moral agency.

2.12 Moral Motivation

There is also the question of *moral motivation*. Authority, and motivation to adhere to codes, may stem from 'soft' powers such as the respect for the originating body, or for the colleagues or the process by which codes of ethics were drawn up and discussed. For those assuming that so long as the very clever people who work in AI produce codes of ethics, this will be enough to inspire confidence in those codes, some humility may be found in personality research which indicates that there is no correlation between how intelligent you are, and how likely you are to follow codes of conduct. Psychological studies have found a null or negative correlation between IQ and the trait of conscientiousness, which roughly translates as 'character' (Luciano et al. 2006; Moutafi et al. 2004).

What's this got to do with AI? Let's face it, AI is run by people who are generally pretty bright, at least in certain ways. But it's a mistake confidently to presume that clever people will draw up, and implement, good codes of ethics, simply in virtue of their intelligence.

2.13 AI, Codes of Ethics and the Law

There is a strong and complex relationship between ethics and law. Codes of ethics are nested within the appropriate legal jurisdictions of local, national and international laws, and seek to adhere to these. However, especially when technology is rapidly advancing, the law might not be able to keep up, and professional bodies and others considering ethical aspects of that technology might well lobby for appropriate changes to the law. It may be possible to amend codes of ethics issued by professional bodies more flexibly and more rapidly than national, and especially international, laws.

There may be great differences in some aspects of the law between different jurisdictions, some of these being differences of great relevance to AI. For example, there are significant differences between the laws on data protection and privacy in the US and in Europe, which can potentially be highly relevant to codes of ethics for regulating AI, and indeed, to how AI is developed.

Meanwhile, how can technology cope when a legal regime might be a stumbling block to its development? For example, legal regimes may be rightly concerned about the development of autonomous vehicles, yet this might slow the development of technology which in the longer term could have a beneficial impact on road safety.

One possibility is to test technology in more permissive jurisdictions. One problem might be certain countries paying a price for the development of technologies from which other countries are more likely to benefit. Suspicion has been raised that testing for paediatric medicines may take place in less developed or developing countries where children are not so vigorously protected (Gulhati 2005). Another more attractive possibility is to have prescribed certain areas where experimentation with technology was permitted, subject to improved regulations (Pagallo 2011).

Law has to be applied, and applied rigorously and consistently across a wide range of circumstances. Attention to how the law might be updated to accommodate various developments in technology, including AI, may proceed with an attention to detail from which ethics could sometimes benefit. Contrariwise, close attention to legal judgement in relation to AI as it unfolds in case law can be both useful for considering ethical issues, and important to note for critical commentary as thinking in AI unfolds.

For example, the 2016 decision in State v Eric Loomis (State of Wisonsin v Eric Loomis 2016) concerned whether the use of the COMPAS algorithm in determining sentencing was fair or whether it violated the Constitutional right to due process. The finding was that it was used appropriately. A legal decision such as this will make reference to precedent and law in the appropriate jurisdictions of course. There can naturally also be broader debates about whether such legal decisions really do capture 'fairness' in such cases. Indeed, in this case, Loomis filed a petition for the writ of *certiorati* concerning the judgement; in an unusual move, the Supreme Court of the US ordered the State of Wisconsin to respond, and on March 6th 2017 in an even more unusual move, the Supreme Court issued a CVSG, a call for the views of the Acting Solicitor General (Admin 2017). This reflects the gravity of the concerns about the lack of transparency in the use of such algorithms and the possible threat to procedural justice and fairness. This level of scrutiny by the courts is to be welcomed and is indeed necessary with the introduction of AI which is potentially altering fundamental tenets of our legal system.

Additionally, the very fact that there are sometimes important relevant differences between jurisdictions on the law, which then shapes debates about ethics and codes of ethics, means that examining the possibilities of different legal regimes can be a good way of thinking more laterally about what is possible and what kinds of legal reform might be desirable.