Chapter 7
Some Characteristic Pitfalls in Considering the Ethics of AI, and What to Do About Them

Abstract Those developing codes of ethics for AI must of necessity consider the ethical issues that AI presents. There are some common pitfalls and gaps in argument to watch out for here. A full treatment of this topic would take much longer, but this chapter simply aims to alert readers to some of the main traps to avoid. There is always a balance between abstract and concrete thinking in ethics. Work in AI and ethics may concentrate too much on the idea that what distinguishes humans is their intelligence, and subsequently, idealisation or oversimplification of what is involved in both human and machine agency may occur. There may be different expectations for human and machine agency which are present but not fully articulated. This can have concrete and deleterious impacts upon any ethical conclusions which are drawn. AI is used to enhance or replace human agency. This means we must pay attention to questions about the boundaries of human agency and 'normal' human functioning. There needs to be careful consideration of different cases, given the varying nature of AI. The impacts of AI may not be just on its immediate use, but further afield within complex social systems, and careful attention should be paid to this. Lastly, clarity of language and of definitions is frequently an issue in AI; common language may mask deep disagreement.

Here, I consider some characteristic pitfalls that may be found in attempts to address the ethics of AI. Again, this is not intended as a comprehensive list of problems, nor to suggest that such difficulties are inevitable; neither is it intended as a failsafe manual of how to avoid problems. The broad questions I examine all relate, in some way, to the ways in which AI characteristically enhances or replaces human judgement and agency.

7.1 The Idealisation of Human and of Machine Agency

7.1.1 The Abstract and the Concrete in Ethics

There is a tension in ethical thinking between abstract, general principles and goals, and the concrete particulars of cases. However, in considering general ideas in

P. Boddington, *Towards a Code of Ethics for Artificial Intelligence*,
Artificial Intelligence: Foundations, Theory, and Algorithms,
DOI 10.1007/978-3-319-60648-4_7

ethics, we can allow our thought to become too abstract and hence we can miss important detail necessary for appropriate application. There are many ways in which this broad problem has been tackled by moral philosophers: for instance, John Rawls' famous notion of Reflective Equilibrium presents a methodology for addressing the balance between theory and concrete particulars (Rawls 2009).

It is by no means inevitable that overly abstract or idealised notions of agency will find their way into discussions of AI and ethics. Indeed, this is a great potential strength of work in AI: that decision making and the behaviour of machines has to be thought about in considerable, concrete and applicable detail. This then can help guard against over-abstraction. Concrete, practical work in AI, for example in the area of robotic-human interaction, can itself uncover various ways in which assumptions and idealisations about humans and human agency can cause problems. In a recent interview, Anca Dragan, who runs the InterAct lab focusing on algorithms for human-robot interaction, remarked, 'We have to stop making implicit assumptions about people and end-users of AI, and rigorously tackle head-on, putting people into the equation' (Conn 2017b). Nonetheless, the focus in AI on agency and intelligence can at times nudge us into an overly idealised or abstract approach to ethical questions.

7.1.2 Artificial Intelligence, and Intelligence as the Hallmark of Humanity

Indeed, hype around AI can veer towards idealisation and simplification. For example, focus on artificial *intelligence* might lead us to overemphasise intelligence as humanity's main feature. It's common for those who avidly advocate AI to imply that there is some upward trajectory of advancing intelligence, an arc of moral progress, and that AI—artificial *intelligence*—is the next step to the progress of humanity—or of transhumans or posthumans. 'I regard the freeing of the human mind from its severe physical limitations of scope and duration as the necessary next step in evolution', states Ray Kurzweil (Kurzweil 2001).

But note, such thoughts often rest implicitly on teleological accounts of human evolution. Scientists are usually better known for considering evolution a product of blind chance, whereby species which don't adapt to changing environments simply die out. 'The necessary next step in evolution' implies that there's been some progress in evolution, but not enough; it's as if Nature herself, who fashioned us from inert matter, is now prompting us to wrestle evolution from her own amateurish hands.

Note, too, that such accounts tend to focus exclusively on intelligence as the factor behind humanity's current state of progress. Yet, theories of human evolution point to many other factors; sexual selection, which is a large factor in human evolution; critically, our social nature, including pair bonding, and the operation of dominance hierarchies; and quite possibly, religion (Barrett et al. 2002).

Including these factors in our considerations alongside intelligence may enrich our understanding of what constitutes progress for humanity. Two points for now. Firstly, recall our discussion of the question of work. One major question raised was how we would deal with the issue of the meaning of our lives with large scale AI-driven redundancy. A view of human life, and human progress, based on an account of intelligence alone—even on a wide notion of what 'intelligence' is—is going to be limited. Recall, for example, the question of the value of different sorts of work. Humans have very wide range of abilities and values, and a large repertoire for squeezing meaning out of life, but we are unlikely to be indefinitely malleable.

Secondly, these other factors in human nature are utterly critical to any account of ethics. Aristotle was not just a philosopher but the world's first biologist. He understood well that not just our intellect and our reasons, not just our emotions, but also our sociability is key to understanding human nature, and hence a key to understanding the 'good life for man'. 'For in the case of human beings what seems to count as living together is this sharing of conversation and thought, not sharing the same pasture, as in the case of grazing animals' (Aristotle 1999) Book IX ch 9. Evolutionary biology is catching up with Aristotle; our social natures have been key to our evolution: look, our large brains could never have developed without the empathy and complex society needed for care of the human infant, born helpless only halfway through gestation; in turn, much of our brainpower is concerned with social skills (Morgan 2011). If we see AI as human progress, if we are concerned about the ethics of AI, we must guard against a simplified attention to bare intelligence and to idealised, isolated individual agency.

7.1.3 Idealisation and Overreach Often Applies in Thinking About the Ethics of AI

Since AI has potentially very wide reach and there is concern about having our lives influenced in every which way by intelligent machines, there is a tendency to consider that the ethics of AI has to cover 'everything', so that we have to 'solve' ethics first. Hence, for all but the most cheery optimist about doing this (maybe someone who never picked up an ethics text book, nor ever watched the news), prospects may seem gloomy. Yet, at worst, this would only be an issue for a form of AI that really did affect everyone, and really did affect all areas of life. For many or most AI applications, certainly at present, there will be limited reach, and hence, the ethical questions, including the question of community agreement, is to that extent contained. There may be no need at all to fix the bigger, global ethical questions first; or at least, we may make some useful progress without this.

7.1.4 Idealisation in Thought About Autonomous Vehicles

We may think about agency differently in the case of human beings, and in the case of machines. This may be done inadvertently. This can mean that we have different expectations of machines; and this can infect our thinking about ethical issues.

There is reason to think that autonomous cars of the future will be safer than human driven cars: because if they're not safer, they won't be accepted; we are likely to be less forgiving if a machine kills us, than if a human being does, for a variety of reasons.

There is a general problem with measures which increase public safety. Realistically, these can never be perfect. The people who are kept safe are statistics. The people who are killed or injured are visible. Who reading this knows for sure that they would have been run over and killed, were it not for advances in vehicle safety? As it's been stated: “If self-driving cars cut the roughly 40,000 annual US traffic fatalities in half, the car makers might get not 20,000 thank-you notes, but 20,000 lawsuits” (Russell et al. 2015).

We may also feel that a human driver, otherwise competent and alert, faced with a vehicle collision in which a bad decision was made under great duress, should be forgiven. We are much less likely to ‘forgive’ a machine that does this. This is partly, I suspect, because of how an autonomous vehicle programmes in advance what to do in some crash scenario. This seems ‘cold blooded’; recall the discussion in Sect. 2.3.8 about how anguish and slowness can be used as an indicator of moral sincerity. Whatever answer is preferred, it's going to help add clarity and nuance to the debate by considering directly how we are idealising machine agency and what happens when we substitute a human decision maker for a machine.

Note, too, a paradox: One main point of AI is to make decisions extremely quickly, and to careful formulae. But in ethics, it's often these very features of decision-making which occasion suspicion. This may indicate that trouble may always be on the horizon wherever machines are stepping in for humans in serious or even tragic cases.

We may also idealise in thinking about ethical issues because of the methodology used. The focus on ‘trolley problem’ type approaches to the ethics of autonomous vehicles, for example, may divert attention away from the wider context of the activity in question. For instance, focus on the precise number of people killed while driving in some abstract simulation might not lend itself to asking the bigger question of why you got in the car in the first place, given that you might end up killing someone. We take cars and road deaths for granted. Especially in those parts of the world which are heavily dependent upon private vehicles, it's a common attitude that humans have a right, a need, to drive. It's less likely that anyone thinks in these terms for introducing the new technology of autonomous vehicles. Individually, we also don't tend to get into vehicles thinking we are a danger to other road users. Collectively, although we don't want to be run over by another driver, we want to drive ourselves, and if the standards for driving skills were too high, too many of us would be ruled out. So, we're likely to be softer on humans than on machines in this regard. A downside of this is that safety concerns about autonomous vehicles may delay their use, even after they have reached the stage of being safer than human drivers.

7.2 Building Ethics into AI and the Idealisation of Moral Agency

We've seen how codes of ethics for AI need to build in an extra layer of complexity, one concerning the behaviour of machines. There are various ways of addressing the control problem. Could building ethical behaviour and decision-making into AI be one answer, as a strategy along with developing codes of ethics? Perhaps such codes may even incorporate as a desideratum of work in AI, a recommendation to build ethical behaviour into machines.

But does this make sense? A very brief snapshot of this idea is included here, for many of the pitfalls reveal simplification or idealisation of the notion of a moral agent. There are unarguable reasons to incorporate into the development and use of AI all steps to ensure safety, and to try to ensure that machine behaviour is consistent with ethical values. But the question about building in ethical decision making and action into AI goes further than this, for it concerns judgement in novel, perhaps unpredictable situations, where decisions and actions would be taken without any immediate human oversight; it goes further than simple alignment of outcomes with our ethical values, if it implies that it's the machine itself which is acting morally.

There are, of course, many forerunners, such as failsafe systems built into trains to cope for catastrophes, such as driver collapse. But these work in systems with limited capabilities. For systems of AI where decisions and actions may be made which might have far reaching, and perhaps hard to detect effects, the idea of building ways to make the decisions of the machine 'ethical' might seem a tempting possibility.

Eliminating catastrophe: Discussions around hard moral dilemmas are not just a hallmark of the ethics literature in general, but the ethics literature in AI in particular. So, in the absence of a complete specification of ethics, attempts to build ethics into machines may instead perhaps usefully be focused on at least trying to prevent appalling consequences.

But even here, it's hard to specify what these are. Is running over the baby a catastrophe, or is running over six 59 year-olds a catastrophe? Is it worse if the accident victim is left in a coma, or if they are killed? And, it turns out, *where AI is concerned many of the possible outcomes lie so far at the extreme limits of what we can imagine that they flip from 'wonderful' to 'catastrophe' like a Necker cube flips from one view to the other*. Is AI-induced mass unemployment the ultimate freeing of the human race—or is it a catastrophe? Is uploading my mind into a computer to gain eternal life (so long as you've bought a good policy for sorting out software bugs) a good thing? But this is what happens to victims of the Cybermen in Dr. Who, who are terrified at the prospect of being ‘upgraded’ to into a machine.

Attaining goals: It will be very hard to programme a machine to address ethical questions, unless we have a pretty clear idea of our value goals. But we lack such a clear view especially for such difficult to imagine, complex possibilities. So could we programme a machine to discover our ‘true’ goals? Well, on what basis would the machine work out our true goals? Well, perhaps either we or the machine can work out what our ‘true’ nature is. But . . . do we even have a ‘true nature’? And is our nature fixed then? And can this be something subject to empirical inquiry? This is an immensely complex philosophical question (Stevenson and Haberman 1998).

Moreover, even if we suppose we can create a machine that could determine our moral goals, this bootstraps up the problem in an unverifiable way. We would always need to be able to check that the outcome was ethical, by our own lights. Are we going to accept that, say, wife-beating was ethical after all, particularly if she’s burnt the dinner *and* has sloppily applied make-up, just because we’ve got an app that told us it was okay? I hope not.

Outsourcing ethics: One of the central claims of this book is that ethics must always involve the possibility of development and of dialogue with others who have legitimate interests; perhaps they are affected, or perhaps they might have some insight to contribute. To outsource ethics to a machine that is not embedded in a web of such human dialogue is counter to all of this. And, should machines develop to a point of sophistication where they have as full moral agency as humans, although such a machine might have interesting things to say, handing over moral judgements to that machine is still outsourcing your ethics to another.

There is a serious problem with the whole idea of outsourcing our ethical judgements and actions to a machine, just as there is for outsourcing them to another person. In consequentialism, the only thing that matters ethically is the outcomes of our actions; this is an agent neutral morality where it does not matter how you reached a decision so long as it’s the right one. So, you could, in principle, outsource your final judgement to an efficient machine. But note that this machine would be simply working out a decision procedure to implement a morality, and doing empirical calculations about how best to achieve a moral goal.

And on virtue ethics and on Kantian views of morality, you simply cannot outsource an ethical decision to others. You can’t ask someone what to do and then do it, because to act as a moral agent intimately involves the quality of your motivation, and the nature of your judgement and decision making. You have to do the right thing, for the right reasons, in the right manner. Even many consequentialists are troubled by this, and try to work around it. And remember our discussion of the Nuremberg trials? The quintessentially bad excuse of the twentieth century was, ‘I was only following orders’. That means that what is perhaps the most important moral insight of the twentieth century—upon which subsequent codes of professional ethics and laws have been built—is that we cannot outsource our moral judgements. It is a judgement of inalienable moral responsibility.

7.3 Replacing and Enhancing Human Agency, Boundaries and AI

One of the biggest questions facing AI is to consider the impacts of the enhancement or replacement of human agency by AI, and to start to analyse the multiple issues involved. There will be complex ethical questions; even if such developments are seen as beneficial, the question still remains of how such benefits are distributed. And, as we've seen, assessing the benefits of such complex and far reaching technologies associated with AI will be in any case, extremely hard. Moreover, simply trying to capture benefits and harms does not exhaust our moral discussions.

Hence, codes of ethics for AI need to be formulated in ways which permit and encourage the full complexity of questions about AI and human agency to be addressed. In particular, codes of ethics for AI research need to encourage research which actively investigates these issues where appropriate. Much research in AI already is looking at large complex systems, and hence could be a promising line of inquiry for including consideration of the ethical questions involved in displacing or supplementing human agency or human agents within such systems.

I noted earlier how thinking about ethical questions may be broader or narrower, and how those with different personality types may be more or less concerned with issues of boundaries in ethics. Given the central questions of how human agency, and even human bodily boundaries for some forms of proposed AI, affect boundaries, when we think about the ethics of AI, we should watch out for the tendency to reject or ridicule attention to boundary issues.

7.3.1 Case by Case Consideration Is Needed

Some AI may extend our capacities in incremental or relatively ethically insignificant ways. But the question of drawing the boundary between ethically significant and ethically insignificant will be contentious. We can see value questions about the enhancement of humans already operating in sport, and in medicine, with questions arising about the boundaries between curing disease or illness, and enhancing human capacities. As with AI, answers to such questions will depend upon ideas about what constitutes 'normal' human functioning, and the appropriateness of going 'beyond' this. There are problems about how to distinguish between incremental changes which have big effects—this is the question of the Sorites paradox of 'when does a few grains of sand become a heap'. One way of determining if a pile of sandgrains has reached the level of a heap, or if emergent properties are exhibited, is by looking further afield at the knock-on effects of the AI. A small change in AI capacity might have a substantial impact elsewhere in a system. For example, it might render a whole class of jobs redundant, and then lead to large institutional restructuring. But this will involve considering AI within its concrete, real world setting. Codes of ethics must therefore take note.

There is only time in a book of this length for brief indication of some of the issues. Let's consider a few examples. In 2016, a Robotic Retinal Dissection Device (R2D2) trial at Oxford was used for the first time to remove a membrane 100th of a mm thick from the retina of a patient. The membrane was distorting the shape of the retina. The robot was placed inside the eye through a hole less than 1 mm in diameter. The remotely operated robot eliminates tremors in the surgeon's hand. Such precision would be impossible for an unaided human hand. The device can perform movements as precise as 1000th of a mm (Parkin 2017).

It is hard not to see such a use of robotics as anything other than a great advance. The robot is controlled by the surgeon at all times, and extends human agency merely in terms of adding precision to human movements. Surgeons already have the ability to perform very delicate operations. And the purpose of the robot, to restore sight to as close as normal functioning as possible, can also be taken as uncontroversial—indeed, of great value.

However, consider a different possibility from a health care setting. There is much work on the potential for using robotics in nursing care, for example, for routine care work. Such work may be used to supplement human labour or to replace it. Working out its impact will be highly complex.

Take the use of robotics to assist with the feeding and toileting of patients on a hospital ward. Will patients benefit or not? For obvious reasons of privacy, a patient may well prefer robotic assistance with using the toilet to human assistance. But it remains to be seen if the same is true for other assistance. Routine care work provides opportunities for human interaction which may make a big difference to quality of life for patients, which in turn affects health outcomes, and may provide opportunities for exchange of useful information about the health status of patients. However, robots could also possibly record various details about patients, producing complex issues about data storage and communication within the hospital system.

7.3.2 What Kind of Questions Do We Need to Ask in Such Cases?

These are far more complex than simply assessing the benefit for patients. Hospital wards are intricate social environments where staff at different grades and functions operate in often varying local cultures, and where social hierarchies operate (Bridges et al. 2013). Within this social setting, there are complex lines of communication of morally relevant knowledge. In recent years, there has been an increasing professionalization of nursing, with nurses often using specialised equipment, and with routine bodily care more and more undertaken by lower status health care assistants (Twigg 2000). Technology seems to track social status. It's hard to predict what impacts there might be on relative status within a ward of the introduction of robotics for various aspects of nursing and routine care. And note that status within the ward is a critical element in how knowledge flows. Health care assistants may have particular, and useful, knowledge about patients, but they may or may not be shut off from ward meetings, and research finds that their low status, combined with the stigmatising nature of the bodily care work they perform, further isolates them as a relatively insular group within the ward (Lloyd et al. 2011). This has implications both for their own wellbeing, and for that of patients.

7.3.3 AI, Ethics, and Effects on Complex Systems

This is merely an indicator of a very complex issue. But in assessing the impact of AI entering a complex social system, it's going to be important to ask questions about how changes might occur elsewhere within that system, and to do so, one needs to understand how the system operates. I noted earlier the error of thinking of humans too much along the dimension of intelligence, a pitfall that might occur if we focus on bringing in artificial *intelligence*. We need to look at our social nature too. We need to look closely at bringing AI into human societies. One reason why I used a hospital ward for my thumb-nail sketch of the use of AI is because it highlights questions about social dominance hierarchies, something to which ethics needs to pay closer attention. Placing robotics into such systems may have unexpected effects, which may be trivial, or may be profound. We need to pay particular attention to how this might affect the transmission of information within a social system; crucially, this affects what issues are even seen as ethical issues, and how. Because those lower in social hierarchies, such as the health care assistants mentioned above, are less likely to be listened to, it may be especially useful to take such dominance hierarchies into account when considering an appraisal of the ethical impact of implementing AI within a social setting. Work in social epistemology could be useful here, and again, the kind of systematic thinking in which many experts in AI are adept may be useful. Codes of ethics might usefully consider explicitly addressing such matters (Goldman and Blanchard 2015).

Take note: Consider social systems. Consider social hierarchies. Consider the impact of technology on these and on the nature of communication. Consider how this might impact upon how ethical issues are uncovered. Consider whose views are least likely to be heard.

7.3.4 Pay Attention: Technology Can Hide, and Technology Can Blind Us

In ethics we need to consider not just what the right thing to do is. We need to consider how ethical questions are seen, how they do and do not come to our notice. The perennial issue in AI of how it supplements, enhances or replaces human agency means that we need to pay attention to what's going on with the human beings affected by the use of AI, and the complexity of human social systems may make it hard to see what impact the AI is having, without close attention.

There are additional questions about ethical visibility that arise with AI. AI takes many different forms. It may be so tightly and so invisibly embedded in complex technological systems that we don't even notice it's there (until it causes us some problem, perhaps); recall that once it's used, we may no longer think of it as AI. Contrariwise, AI may be whizzy, hi-tech, dazzling and exciting. Both these features—invisibility and prominence—are typical of AI, and both present ethical challenges.

We saw above in the brief discussion of robot camel jockeys how focus upon the robots as a solution to a moral problem might distract from considering other important aspects of the situation. Technology can over-complicate matters: Anti-Slavery International commented wryly of proposals to introduce robot jockeys into the UAE: '*This seems a complicated alternative to implementing fair labour conditions for adult jockeys.*' (Anti-Slavery International 2006). Technology which dazzles us can also prevent us from looking closer at other issues, as we saw above in discussions of how it seemed to be the lure of the technology which enticed owners to replace child jockeys, rather than any moral realisation of the wrongs of using children. Codes of ethics for AI need to consider carefully how these sometimes opposing aspects of technology—hidden, or revealed in full, chrome-gleaming lustre—may impact upon what ethical problems are visible, and what ethical solutions are sought.

7.4 Addressing the Increased Gradient of Vulnerability

We've seen how a distinctive issue for codes of ethics for AI is how the problem of control decreases the gradient of vulnerability between AI professional and others, which in turn threatens the authoritative base of professionals and of any codes of ethics. Attempts to address this are key to developing autonomous AI, and include discussions about how to retain meaningful control over AI, and indeed, what such meaningful control would even look like. It is obviously impossible in a book of this length to address what precisely to do about this. If I could answer the control question, this little book would be at the top of the Amazon best seller lists for sure.

However, the question of how to develop codes of ethics for AI, given the control problem, is somewhat different. It again reinforces the need for wide public communication and involvement.

It might be a crumb of comfort to AI professionals to see that there are similar issues elsewhere. In medicine the professional status of the doctor is being gradually transformed, eroding traditional notions of professional authority and causing numerous troubling questions for professional ethics.

Developments in technology, such as remote devices, some of which include AI, allow individual patients to collect and be in charge of a great deal of data concerning their own illnesses and health status, and, together with the rise of patient groups and more widely disseminated medical knowledge, this has led to the rise of the ‘expert patient’ (Department of Health 2001). This weakens the traditional expertise gradient between medical professional and patient, challenging the superiority, integrity and validity of medical knowledge, whilst notably validating one of the central values of medical ethics, patient autonomy. The boundaries and power base of expertise of the medical profession is also challenged now that much health data is in the hands of mobile phone companies rather than in the control of the medical profession, so new loci of struggle for control are arising as medical practices extend beyond the traditional clinical encounter; indeed, it is developments in AI, inter alia, which are producing such a challenge (Boddington 2016). However, there are not inconsiderable challenges to the authority of any resulting codes, since changes in the ‘vulnerability gradient’ both diminishes and modifies the power of the professionals and professional bodies who are producing these codes. It could be very useful to keep track of how medical ethics is dealing with such changes. This use of technology also raises questions about the potential loss of power of the medical profession. Changing patterns in knowledge and expertise between individuals and groups is a common feature of emerging technologies.

One response is to recognise the legitimacy of public concerns and to express these in forms such as widespread public consultations. These should be very welcome for AI.

7.5 Common Language, Miscommunication and the Search for Clarity

There is a pressing need for clarity of communication in any enterprise of investigating and developing ethics, and this is a particular issue with the ethics of AI both because of its technical complexity, and because of the need to add as much transparency as possible, given the difficulties with transparency in some forms of AI. All interested parties need to be able to understand the ethical issues, and so there’s a need for technical language and concepts to be communicated clearly; but note of course, this need for communication goes both ways—those working in AI need to understand the concerns of those outside the field.

A particular problem in AI is that there are terms which are used in technical sense which are also in common parlance. Perhaps the prime example of this is the word ‘autonomy’. This is used in particular ways by those working in AI; it’s used in common speech; and it’s used by philosophers. There may not be complete agreement between different uses of the word. And any misunderstandings thereby generated are likely to be important ethically; it’s a concept to which great value is attached.

Despite the necessary calls for clarity on language and definitions regarding AI and ethics, a common vocabulary can mask disagreements. Only a depth of dialogue and an understanding of underlying background issues will reveal this. But don't be mistaken in thinking that all we need to do to improve understanding is simply need to come up with a robust and agreed definition of autonomy.

7.5.1 Common Language May Mask Disagreement: A Tale of Two Autonomies

Within ethics itself, autonomy can also be understood in radically opposing ways. You could literally fill an entire book case with material on this. Here I just illustrate briefly two contrasting approaches. This also serves to illustrate how deep debates about the ethics of AI are likely to go, and to warn how language may mask serious disagreements.

Autonomy may be used to signify human agency, responsibility and freedom, and it's frequently used in this context to flag the importance of allowing individuals to make decisions for themselves and to hold their own personal values, without any outside influence. It represents the rejection of external demands. It can be used to mean, 'I set the rules for me.' Recall that earlier, we discussed the view that morality only concerns how I treat others; that I can do what I want if it only affects me.

Yet we can trace back emphasis on autonomy in ethics to Kant's philosophy, in which it is an essential feature of human beings that they are rational agents, capable of autonomy. BUT note this: For Kant, to be a rational agent is to recognise the pull of rationality; we *participate* in rationality. Rationality gives rise to the demands of morality (Kant 1972). This means that in acting with full autonomy, we also act morally, motivated by our reverence for the Moral Law. And the Moral Law, based as it is on reason, gives universally applicable answers (at least in theory). This is freedom, this is autonomy, not because we are doing 'what the hell we like', but because we are acting in accordance with our natures as rational, autonomous beings.

To argue that autonomy means the rejection of 'external' demands and hence, that each person can do what he or she likes, is correct then, only if you ignore that for Kant and for his followers, the demands of rationality, and hence the demands of morality, are not 'external' to us.

Some tricky concepts that are likely to crop up: Note that there are many concepts where values are deeply implied, but which may not at first sight seem purely value terms themselves (Williams 1985). For example, consider the word 'parent' which we discussed earlier. We also looked at the notion of 'bias' which seems at first sight always wrong, but which on inspection, things are not so clear. The notion of trust is often used in relation to AI, especially in robotics, but again, trust is a two-edged sword. Adults grooming children for sex are very good at eliciting trust, for example; it's actually quite easy to do.

In conclusion, it's important to note that language is not a set of labels fixed to the world, but serves multiple purposes; even as a description of the world, we rarely need a 'full' description, but pick a description to suit various purposes. And there will be notions of value included in many words which are not straightforwardly 'value' terms. Moreover, there will be different implications, and different connotations, for different people. Hence, in looking for definitions of key terms we may not need to get 'the' definition. Rather, it may be better to flag up possible misunderstandings, and make sure that common language does not elide complexity and mask disagreement. The masking of disagreement may occur where codes of ethics are trying to formalise language. Glossaries can be helpful, but not if they shoehorn complex concepts into a box; and it would often be useful to note the difficulties of producing a simple, standard definition.