UK £6.99 USA $12.99 CANADA $13.99

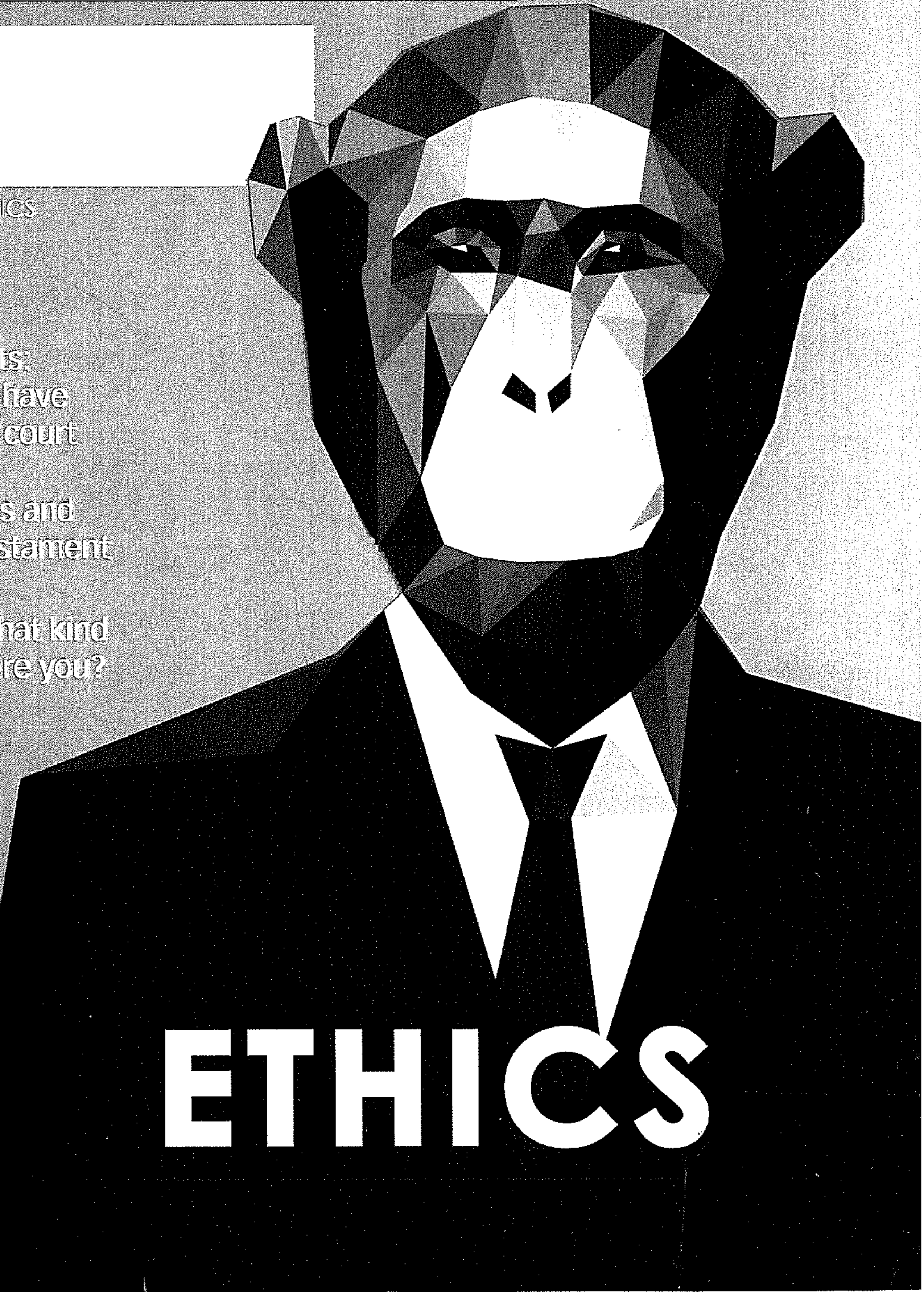
The Ultimate Guide

ISSUE ONE: ETHICS

Animal rights:
When apes have
their day in court

Virtue ethics and
the New Testament

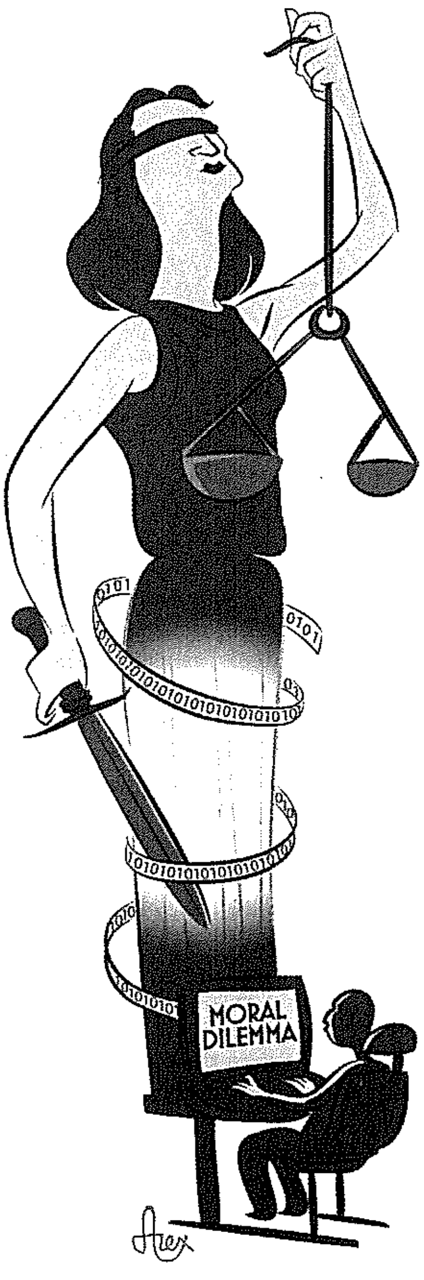
Find out: What kind
of ethicist are you?



ETHICS

The Morality Machine

Phil Badger considers what it would take to make truly justifiable moral decisions.



Imagine for a moment that it will soon be possible to program a computer in such a way that it can generate the answer to any particular moral dilemma we might be faced with. Now further imagine the process that might lead to the writing of such a program, and what the content of that program might be. We might imagine a huge project in which people from different backgrounds and cultures come together to debate and discuss not just moral problems, but the principles upon which they should be decided. Many suggestions will be made as to which principles should be adopted. Some might suggest a single principle, such as, 'Do for others what you would want them to do for you', while others will want a hierarchy of principles, such that the less significant ones are moderated but not annulled by the principles to which we grant higher priority. The conversation about which principles to build into the program might go on a very long time and involve the active philosophical engagement of vast numbers of people.

Perhaps you think that such a thing could never happen, or that, if it did, it would turn out not to be a long conversation, but an eternal one, in which no agreement was ever reached. Perhaps the effort would lead to bloodshed as people, incapable of altering their perspectives, settled their differences through violence.

Let us assume, however, that progress *can* be made. This requires that morality is a puzzle that, regardless of its difficulty, does have principles that are amenable to discovery. So the Morality Machine entails a particular kind of what is called 'moral realism': it means that there are such things as moral truths. Finding valid moral principles will involve a shared search for truth; but of a kind that might be unfamiliar to some readers.

In the ordinary course of things, we tend to think that we can establish the veracity of propositions by empirical means – we look for evidence for our propositions in our experience. This approach seems fairly straightforward if we are talking about the claim that 'the cat is on the mat', but becomes more awkward for claims like 'murder is wrong'. Indeed, some moral realist theories involve the idea that knowing the moral truth involves some kind of special 'moral perception'. This was the thought behind the 'Intuitionism' of the twentieth century Cambridge philosopher G. E. Moore. However, most of us don't hold out much hope for the detection of what Ronald Dworkin, in his 2010 book *Justice for Hedgehogs*, called 'moral particles', or, in satirical mode, 'morons'.

To be fair to G.E. Moore, he didn't actually think that moral truths were the kinds of things that you found 'in the world'. Rather, moral truth was 'non-natural', and subject to a special and fairly mysterious form of perception called 'moral intuition'. For Moore, moral perception was only similar to ordinary perception in the sense that it was something you either just had or didn't have. And just as some people have defective colour vision, they might also have defective moral vision (although most people won't), and no amount of discussion will change that. From the intuitionist viewpoint, giving reasons is not part of the business of determining moral truth. In this (if in nothing else), intuitionism has something in common with sceptical positions such as emotivism. Emotivism asserts that our moral attitudes are neither the product of reason nor of perception, they are merely matters of taste. By contrast, giving reasons is at the heart of what Dworkin calls the 'interpretative' version of realism. In this view, moral truths are not things we *see* or otherwise perceive, but *conclusions* we arrive at. And the program for our Morality Machine is constructed via a process of moral interpretation.

To see the difference between the emotivist and the interpretativist views, consider the difference between a discussion with a friend about the merits of liquorice and another about the film you saw last night. In the former case, the discussion is likely to be a short one. You either agree or disagree that liquorice is horrid, and nothing anyone can say can change your immediate and visceral feelings about the matter. In the case of the film, however,

"The Morality Machine entails a particular kind of what is called 'moral realism': it means that there are such things as moral truths."

things are very different, and you have much to discuss. Perhaps you agree that it was terrible, and begin exploring the reasons you have for thinking it so. As you do so, almost certainly you'll find areas of overlap between your thinking: you agree that the direction was a bit ham-fisted and that the lead actor was even more wooden than in the last film you saw him in. On the other hand there will be moments of disagreement. One of you might consider the cameo by the veteran actor with the limp to have been charming and a saving grace of the whole movie, while the other thinks that whole section of the film is manipulative and bolted on. Later, it strikes you that you had misunderstood the nature of the film, such that you now conclude that it really was a satire on its topic, and that, in this new light, it was a triumph. Your friend initially disagrees, but is ultimately won over by your arguments.

The point is that the truth about the film is not arrived at by simple perception, nor is it reached by logical deduction (there is no mathematics of film criticism), but remains something other than merely a matter of taste. Liking a film is not like liking liquorice, and the reasons we might give for doing so are clearly more than simple rationalisations of our tastes.

A pertinent question is how similar are our moral views to our opinions about films? Clearly in both the film and morality cases, reasoning is important, and as in the film case, I might be able to persuade you to change your mind on some important moral issues. I might, for example, point out an inconsistency in your views about suffering that lead you to reevaluate your views on euthanasia or vegetarianism. As far as our Morality Machine program is concerned, however, that degree of persuasiveness and openness to discussion may not be nearly enough. What is required for the Morality Machine to function at all is that there be an ultimate point of resolution where the arguments stop. And for this to happen we need to be able to recognise that this or that set of principles really will do the job of correctly answering any and all moral questions.

The Moment of Truth

For many years I've aspired to know what the contents of the Morality Machine's program would be. I've toiled, at times in the pages of this very magazine, to solve problems concerning the relative moral status of our rationally-derived duties and the consequences of our actions for others; in other words, I've been torn between the ethics of Kant and those of the Utilitarians. I've indulged in 'trolleyology' (I'll explain that soon), and I've speculated about justification for torture. Now I sense I've reached the moment of decision. By the end of this article, I will either abandon the search for a coherent moral viewpoint and a program for my Morality Machine, or I'll stick my neck out, define the program, and wait for readers to agree with me, or otherwise.

Let me first explain why I might be tempted to give up. The issue comes down to the possibility of what we might call 'the irreducible difficulty of moral decision-making'. This is the thought that, no matter how hard we work at it, there is simply no program that could respond to the demands of all the moral dilemmas we might face.

> **"One of the attractions of the idea of a Morality Machine is that thinking along these lines might keep us honest and prevent moral discourse amounting to the *post hoc* rationalisation of our decisions."**

Years ago I created an 'ethics game' in order to facilitate rational moral thinking in my students. Participants had to imagine themselves as having two responsibilities. Firstly, they had to draw up a constitution (in effect, the program of the Morality Machine) by ordering a series of moral priorities. Secondly, they had to apply the constitution they had created to a series of difficult cases. Effectively, this meant that they became the Morality Machine and ran its program. You may be familiar with the kinds of cases I threw at them: my students had to deal with all the popular moral issues philosophers like to think about. One of the beauties of the game was that I could add more and more cases as time went on, and the game has grown accordingly.

The result was mayhem. The most common demand from frustrated players was that they be allowed to rewrite their constitutions (programs) in accordance with how they felt about each case in turn. In other words, they wanted a licence to be inconsistent (they wanted to be 'situation ethicists'). This was something I absolutely disallowed. I conceded that they might re-order principles in the light of the insights they gained from applying them to the cases; but the ultimate goal remained that of coming to a settled list of principles in a settled order of priority which could be applied to any and all cases. I demanded, in short, that interpretation should come to an end.

As well as wanting to promote debate, I wanted to create a critical distance between my students and their gut instincts. One of the attractions of the idea of a Morality Machine is that thinking along these lines might keep us honest and prevent moral discourse amounting to the *post hoc* rationalisation of our decisions – as many psychologists and philosophers think it to be. However I had another, less strictly pedagogical, reason for setting up my game. I was mindful of a critique that had been made of Rawls which I was anxious to address.

In his thought experiment concerning what he called the 'Original Position' for moral deliberation, Rawls asked his students to imagine that they were disembodied beings residing behind a 'veil of ignorance', waiting to be born into a body with unknown characteristics and an unknown life. By these means, Rawls wanted to distance his participants from any biases which might lead them to unjustifiably favour particular kinds of people when drawing up their 'principles of justice'.

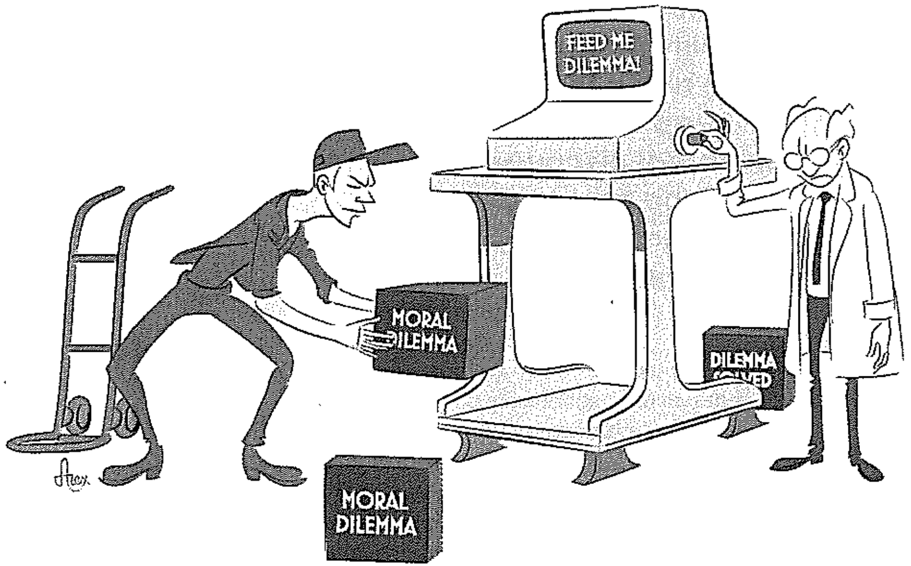
Rawls took much philosophical flack from many directions for his way of thinking about morality. Communitarians claimed the whole thing to be psychologically implausible, and, through abstraction from everything that defines a person, to hollow out the personal, leaving us not with an unbiased arbiter as Rawls imagined, but with no individual at all. Others, such as Jürgen Habermas, have argued that thought experiments cannot replace the 'communicative rationality' of actual people, although for Rawls – as for Kant – the universality of rationality meant that any and all individuals would arrive at the same principle, independent of actual debate. It was partly to avoid all such objections that I asked real people to decide on an ethical constitution.

So far so good; and things arguably got better when I managed to convince the vast majority of players that the following constituted the ideal constitution:

1. Respect the autonomy of autonomous

beings in respect of their 'large-scale concepts of the good'. (Autonomy is someone's capacity to direct their own life.) This is a 'negative' principle, or principle of self-restraint, which demands that we don't interfere with or coerce people in respect of their deepest commitments, values and beliefs;

2. Minimise suffering where possible. This is a 'positive' principle, or principle of action, which demands that we don't just refrain from hurting others, but actively intervene to reduce the suffering of other beings;
3. Aim to promote the autonomy of potentially autonomous beings. This is another positive principle. It suggests, amongst other things, that we should educate children, or make sure that disabled people have the opportunity to participate fully in society, etc.



"I'm suggesting that the conversation about morality has an end point, and so, in this way, ethics comes closer to the deductive logic of mathematics than it does to the interpretation of cinema."

There are many good reasons for choosing precisely these principles and ordering them in precisely this way (if we reverse the order of the first two principles, for example, we end up with a justification of involuntary euthanasia – overdosing a terminally ill patient without consulting him because we know that this will lessen his suffering); but the point right now is to concentrate on those moments where this constitution/program comes under critical stress.

Here's one such moment, which involves the alleged distinction between 'acts' and 'omissions' and an excursion into the world of 'Trolleyology':

Philippa Foot's 'trolley problem' is a famous thought experiment in ethics. We are faced with the moral dilemma of diverting or not diverting a runaway railway carriage. We can pull a lever that will cause the carriage to switch tracks, saving the lives of five innocent track workers, but killing a different individual; or we can leave things as they are and allow the five to die, and the other to live. I've argued in previous articles that, because the people involved cannot be consulted about their 'large-scale concepts of the good', we should rule such concepts as irrelevant to the issue (they may or may not be inclined to be heroic, but we have no way of knowing which). In this context I'm prepared to be a consequentialist (i.e., to minimise suffering) and pull the lever. However, some of my students took the 'Kantian' line that it is wrong to try to 'add up' suffering in this way rather than valuing each individual person as an 'end in themselves'. They further argued that there was a distinct difference between 'allowing' the deaths of five people and actively intervening to 'cause' the death of another.

It might occur to some readers – as it did to some of my students – that a relatively minor tweak to my constitution might produce a result rather closer to their moral intuitions. These students conceded that the large-scale concepts of the good of the innocent workers in the trolley scenario were irrelevant (because they were unknowable), but they proposed the following idea. Why not, they asked, simply insert a requirement *not to cause suffering* between the first and second principles? The impact of this clause, they suggested, would be to make it okay to act positively to minimise suffering, but only on occasions when this did not involve creating suffering elsewhere (they rejected my notion that the principle of not causing suffering was in a sense contained within the principle of minimising it). In this way, diverting trolleys to squash an innocent man again becomes an immoral thing to do.

The problem with this response, especially in respect of the Morality Machine, is that the injunction not to cause suffering suggests that causing *any* suffering in the cause of the greater good is immoral. Yet imagine for a moment that the lone person on the other train track is fully aware that you could divert the runaway train in his direction, but he has his foot caught in the rail. He is making frantic efforts to escape, but you have serious doubts that he will free himself before the trolley hits him. However, you also realise that by him arranging his body in a particular way, he can ensure that the trolley only cuts off his foot! Now what would you do?

I suspect that many who were unwilling to kill this individual to save the others might be less perturbed about maiming him to save them. This suggests that the 'not causing suffering' principle is in further need of modification; perhaps by a careful addition of the word 'significant'. The problem with this move is that we then have to decide what constitutes *significant* suffering. Perhaps causing the loss of a foot and the associated agony is acceptable, but a whole leg would not be? Certainly, so far as writing a morality program is concerned, we're in deep trouble. Machines are going to find it very difficult to make judgements about an issue as intrinsically subjective as what constitutes significant suffering.

What I'm edging towards here is the idea that I should leave my original constitution/program for the Morality Machine as it is. Respecting people's large-scale concepts of the good takes precedence so long as they are not impacting on others; but after that, it's legitimate to prevent suffering; and finally, to try to promote autonomy. Using these rules, it might indeed be possible to have computer-aided moral decision-making. We have what philosophers call a decision procedure for moral dilemmas, and so a way to define our moral duties. Given almost any situation, we feed our moral dilemma into the machine and wait for the systematic application of the program to tell us what to do. We turn the moral crank on the side of our machine, and we get our answers.

Horrified Responses

You might be feeling a little bit horrified right now. You might feel that I should be willing to accept what I called the irreducible difficulty of moral decision-making, and so give up my search for a morality program. On the contrary, I'm suggesting that the conversation about morality has an end point, and so, in this way, ethics comes closer to the deductive logic of mathematics than it does to the interpretation of cinema. If you're attracted to the kind of postmodern perspective which celebrates rather than despairs in the loss of moral certainty, you might be disappointed by this conclusion. You might for instance see a life without moral foundations as a place where real freedom might be found. I tend to disagree with that view because it seems to me to be inherently self-defeating: to begin an argument for the value of freedom on the grounds that there is no moral truth seems more than a bit of a non-starter. (Incidentally, the 'irreducible difficulty' idea suggests a particularly nasty version of the problem of evil, which might give some theists out there pause for thought. If you combine a belief in a creator with the thought that it is impossible to do the right thing in certain situations, you might begin to question if that creator can be intrinsically good. Could an all-good and all-powerful being have created a world in which it was impossible to make correct moral choices?)

Another, equally horrified, response to my conclusions will likely come from the modern followers of Aristotle known as 'Virtue Theorists'.

For Aristotle, there could be no moral decision-making procedure in the way I've outlined here. For him, it was wisdom or judgement that defined the moral individual, and that wisdom or judgement did not result from the application of abstract principles but from the development of a balance in our characters between different virtues (such as courage balanced against compassion etc). But for me this thinking has an air of circularity about it: good people are those who do the right things, and they know what is right because they are good... There are no clues here as to how to escape this circle, and modern virtue theorists have a tendency to resort to identifying virtue in terms relative to the values or traditions of particular societies.

I also have to admit here to a kind of personal tendency that may have not served me well in life or in thought. Put simply, I've wanted to cut to the chase – to generate moral answers based on rational deliberation. I've had little desire to wait for the dawning of wisdom. I might have, weirdly, seen philosophy as a kind of short cut to the answers. The metaphor of the Morality Machine was a particularly apt one for me in my writing this article. My approach to philosophy has always been cerebral, emotionally detached, universalistic, perhaps even stereotypically male. All of this might have been detrimental to the development of my judgement and even my personality. I mention this because it seems to me that what may be true of me is likely to be true of others, and so illuminating about the unconscious inclinations which shape our thinking.

But perhaps I'm being too hard both on myself and on the Enlightenment tradition of thought from which I derive my perspective. I don't really think that the moral conversation we have can have an end in the sense of an ultimate solution. I do, however, think that it is pretty clear that some answers are manifestly better than others, and that clarity of thought in matters of ethics is important. Moral judgement is not like mathematics; but nor is it like disliking liquorice. It is, perhaps, a little more objective than film criticism. Equally, the fact that my position sees us as *constructing* morality is no argument that what we make is unreal. People constructed the pyramids, and more abstract things like 'Government', but both these things are real in most ordinary senses of the word. Indeed, abstract constructions such as governments are perhaps more real for most of us than the distant, unvisited pyramids.

In deciding to not give up on the Morality Machine, I realise that I'll incur the displeasure or possibly the pity of many readers. Nevertheless, I'd ask you all to attend to my first principle: what makes life meaningful to us – our deepest commitments and values – are *ours*, and are not justifiably subject to any coercion. Persuasion is another matter, and I anticipate your responses with pleasure as well as trepidation.