AI and Ethics (2021) 1:109–111
https://doi.org/10.1007/s43681-020-00017-0

OPINION PAPER

# AI and moral thinking: how can we live well with machines to enhance our moral agency?

**Paula Boddington**[1]

Received: 16 September 2020 / Accepted: 18 September 2020 / Published online: 6 October 2020


**Abstract**
Humans should never relinquish moral agency to machines, and machines should be 'aligned' with human values; but we also need to consider how broad assumptions about our moral capacities and the capabilities of AI, impact on how we think about AI and ethics. Consideration of certain approaches, such as the idea that we might programme our ethics into machines, may rest upon a tacit assumption of our own moral progress. Here I consider how broad assumptions about morality act to suggest certain approaches in addressing the ethics of AI. Work in the ethics of AI would benefit from closer attention not just to what our moral judgements should be, but also to how we deliberate and act morally: the process of moral decision-making. We must guard against any erosion of our moral agency and responsibilities. Attention to the differences between humans and machines, alongside attention to ways in which humans fail ethically, could be useful in spotting specific, if limited, ways that AI assist us to advance our moral agency.

**Keywords** Ethics · Moral agency · Autonomy · Responsibility

My starting point is this: that humans are agents. This agency is a central feature of our humanity and of what makes each one of us both interesting and valuable individuals. It is also central to the value of humanity viewed collectively. And this agency includes, importantly, our moral agency. This we must not lose. Computers, even those with artificial intelligence, are our tools. They should not diminish our agency; ideally, we should use them to enhance our agency.

Hence, we can see the validity of the strategy of AI alignment. It is right and proper that attention is paid to ensuring that AI does not control us, but that we control AI, and that AI does not produce decisions or results which are at odds with the moral judgements of those human who use AI and of those humans who are on the receiving end of an AI's decisions and actions: we must ensure that AI does what we want it to do. This can be seen as a negative strategy, of trying to ensure that disaster does not occur, or less dramatically, of fine-tuning AI to keep on track with our wishes and values. It can also be seen as a more positive strategy, of ensuring that AI works for us, to produce beneficial outcomes. This is naturally to be welcomed.

A further goal might be to develop and use AI that could make moral decisions for us. I would strongly argue against this on grounds of the importance of our moral agency. And there is a sense in which this is simply not possible, because if we decide to outsource our moral decision-making to a machine, it is we who have taken this outsourcing decision, and it is we who are ultimately responsible, we who have decided to let go of our moral responsibility to automation. But, perhaps, AI could assist us in our moral decision-making. It will all depend on how.

Note carefully that these strategies of AI alignment assume that we have a clear idea of what our moral goals and values are. I will come back to this.

And note that if we want to meet this goal of AI alignment, we need to ensure that AI and its habitual use by individuals and by society, does not warp or eclipse our values and our goals, and does not distort or obscure our view of the world. This is especially critical given our growing dependence upon technology which uses AI, given its power to control the information presented to us and in turn to nudge or manipulate our responses.

Behind this lies a feature of human beings that is so obvious we had better point it out in case we miss it—we are

✉ Paula Boddington
paula.boddington@nchlondon.ac.uk

[1] New College of the Humanities, 19 Bedford Square, London WC1N 3HH, UK

flawed. However, central to humans our agency might be, this also is central to who we are: our agency can fail, our moral agency can break.

There are many different detailed accounts of human nature, whether grounded in philosophy, theology, spirituality, or science. Despite major differences, there are some commonly recurring themes, and our weakness, in particular our moral weakness, our lack of moral insights, our tendency to fail when the going gets tough, is one of these.

We are flawed in multiple ways (and there are disagreements, naturally, on how to describe, understand, and account for these flaws). One reason why we ought to be wary of developing AI that takes value decisions in place of us, or erodes our moral agency in any manner (which could happen in many subtle ways), is the human failing that produces our tendency towards the abnegation of our own moral responsibility, a shortcoming which has been present in human beings at least since Adam tried to blame everything on Eve and Eve tried to blame everything on the serpent. Another flaw is our ever-present capacity for backsliding, for interpreting the world in ways that favour ourselves and make our lives easier. We are also very liable to be manipulated and tricked to think that all that glistens is truly gold.

So one problem is that we often willingly give up our agency, hand it over to others. Another problem is that it is often wrested from us, as when we become duped by others, by machines, when for instance, an AI might 'understand' us so well it can manipulate our emotions. There is of course a widely held view that our agency is an illusion and that we are simply the product of deterministic forces. Putting that debate to one side, its certainly the case that we can become the product of deterministic forces around us, that we may even welcome this, and that some AI may offer a particularly powerful means of subduing our moral agency.

So to summarise so far: although our moral agency is a given, and retaining our moral agency is a sine qua non, it is incredibly unreliable. Thus, the ethics of AI needs to consider, not simply how to align AI with our current values, but how we are to work well together with AI, in ways that enhance our moral agency, and taking into account our tendencies to weaken or lose this agency.

The fast progress in AI, and the exploding attention to the ethics of AI, ought to make us stop and take pause about our values. We assume that our progress in technology will continue into the future. But its very tempting to assume that our progress in ethics has reached some kind of pinnacle—we can look back to the past and 'tsk, tsk' at views which considered women were not capable of voting, or that the death penalty was justified for stealing a sheep, and so on and so forth. But the very speed of some recent changes in attitude should indicate that its unlikely that we are now living in Peak Moral Times, as it were. That would be too much of a coincidence.

So in addition to working to ensure that AI is aligned to our values, we ought also to work very carefully to consider the processes by which we form those values. And just as importantly, given the human tendency, even when we know exactly what needs to be done, somehow to fail entirely to do it, the ethics of AI needs to consider how moral judgements lead successfully to actions, finely tuned to the concrete circumstances of application.

We must always be aware that we could be wrong, that we might have overlooked something. One way of illustrating the point is this: there is good reason to argue that there should be ultimate human control of AI. But who controls the humans?

Thus, the ethics of AI needs also to be the ethics of the human. We need to look not simply at our moral values, but also at our natures as moral agents, our strengths and weaknesses in discerning that moral questions arise in any given situation, at analysing the issues, at coming to moral judgements, and at implementing these in practice; both individually and collectively. If we are to be the best moral agents we can be, we need always to consider that we may be in the wrong.

When we consider process of moral judgement and action, when we consider the human weak spots where moral judgement is liable to crack or break, we need to do this both abstractly and in the concrete details of how we interact with AI. So as well as requiring thought about the nature of morality, about our capacities as moral agents, which needs to draw on disciplines such as philosophy, anthropology, psychology, sociology, and theology, this project has to be done in minutely detailed partnerships with experts in technology and the real-world applications of AI. Hence a journal such as this is to be welcomed.

The ethics of AI is sometimes presented in terms of fear: of the existential risks many argue we are facing. It is also often presented in terms of hope: of working to maximise the enormous potential for good of AI. And I believe that here we have an additional reason for hope: to use what might be called the recent Cambrian Explosion of interest in the ethics of AI, to consider not just how to align AI with human values, but how to align humans with human values: to do moral work on ourselves, to advance discussions, debate, and action in regard to that hallmark of our humanity, our agency, and our ability to realise that we have weaknesses, and strive to overcome them.

There are some grounds for optimism, as well as grounds for fear that our use of technologies may already be eroding our moral capabilities. There are moves to extend ethics education for computer science students, for example, but this needs to be more than a cursory examination of codes

and regulations. This needs to be a wide conversation about how human beings are going to flourish into the future.

Such a conversation about human flourishing has always taken place, in some form and at some level. The very fact that it needs to happen now, and that it needs to happen urgently, is as a result of the very dangers of AI that can be turned to a great benefit. What is happening now with the rise of interest in the ethics of AI is that precisely because we are understanding that some forms and applications of AI may threaten our agency, may radically change our lifestyles, may radically and subtly change how we relate to each other, to the world, and even how we understand ourselves; precisely because of such dangers, we are confronted with having to address some really fundamental questions about human agency and how to address human weaknesses, and some fundamental conceptual questions at the foundations of ethics.

Some of these questions can be illustrated as follows. Some have argued that we can take medical ethics as a model for work in the ethics of AI. Yes, of course, we must help ourselves to whatever we can learn from other areas. But the ethics of AI, again and again, will force us to dive deeper. Take the goal of producing benefit for humanity. Medical ethics deals with health—an obvious benefit (although even here achieving this benefit may have certain costs which may on occasion be weighed against the goal of health above all else). But what about all the possible benefits accrue from AI? How do we address this? For the areas of application of AI are so many and various that they reach into every area of life, so that we are forced to consider more deeply what kind of lives we wish to live. And this is especially so given that all the while we are considering this, AI and various applications that use it are changing how we act, how we relate to each other and to the world, how we think.

Then take autonomy. It is possible to paint the history of medical ethics in terms of a growing emphasis on protecting patient autonomy from the powers of the medical profession. Within discussions of medical ethics then, autonomy is thus conceptualised within the large, but nonetheless specific, arena of health, and seen as counter to the powers of the medical profession and health authorities. But not only is the range of application of AI potentially more or less any area of life, the question of autonomy of AI, how to understand it, and how to respond to it, is one of the central difficulties! It is true that work in medical ethics has indeed examined closely different ways of understanding autonomy and problems associated with this. But work in AI needs to go even further. And this means not simply understanding autonomy as it might apply to machines, but also as it applies to humans.

AI should be a tool. We need mastery over our tools. We use tools to complement our strengths and weaknesses. We use hammers to extend our strength—try bashing in a nail with your bare hands—but we have to have a certain strength, and a certain degree of precision, a certain discipline and skill, to use one. Work in the ethics of AI needs to explore ways in which our moral weaknesses and strengths can be helped, or hindered, by AI and its use, using whatever approaches might shed light on these questions.

If you seek moral advice from a friend, its best to pick a friend who is rather different from you in certain ways—a calmer personality, perhaps, someone who is not quite so timid, perhaps, and with some different life experiences. But you would diminish your moral agency, and diminish your chances of developing your moral judgement, if you just blindly followed your friend's advice. And we may have some friends whom we know full well should never be consulted about certain moral dilemmas! Likewise, we should consider how the particular strengths of AI might lend support to our moral judgements in ways that enhance our moral agency and develop our moral skills. Here is just one general thought: one of the big differences between humans and AI is the ease and speed of sharing information. Humans take far longer to share information, and there are many stumbling blocks, both personal and interpersonal, to effective communication. Yet, many great preventable disasters, many great moral failings in institutions, have as a prominent cause the failure to communicate. Is there any possible way that we can use AI to assist us in overcoming this perennial and often catastrophic shortcoming that we have as a species? Bearing in mind, that of course communication is no good, if nobody is listening.

We could think of our journey into the future with AI as a journey of learning a new form of craftsmanship. Both Plato and Aristotle used crafts as an analogy for virtue. Much could be said in explication of the use of such analogies, but here are a few thoughts to end. The skilled craftworker understands their tools, and in developing and fine-tuning their use of these tools, will produce work which nonetheless, is fully their own. In doing so, they will also be developing their characters—patience, attention to detail, focus, perseverance, and so on. Simply getting bigger and better tools is no substitute for the acquisition of such personal skills. Skill at crafts involves aiming to produce something of great merit; but at its best, it also hones and refines the humanity and agency of the craftworker. Craftwork is a fine-tuned blend of humanity and tools.

Hence, in moving forward with AI, we do not need simply a wide-ranging approach to its ethics, but dialogue and integration between such approaches. This journal promises to contribute to that task.