# AI ethics and the limits of code(s)

www.nesta.org.uk/blog/ai-ethics-and-limits-codes/

What might be going wrong and how to put it right. Five ways AI ethics needs to be radically reshaped.

Monday, 16 September 2019 In Futurescoping , Government innovation 7 min read

By: Geoff Mulgan

The ethics of Artificial intelligence (AI) matters, obviously. The arrival of hugely powerful technologies, in a wide range of fields, demands urgent attention to questions of right and wrong, power and abuse, bias and distortion.

Autonomous weapons, facial recognition and the use of predictive algorithms in justice are just a few of the intensely ethical issues that AI has thrown up. The sheer scale and range of activity now underway is captured in this excellent recent overview from Nature[1] by Effy Vayena and colleagues at ETH.

That a clutch of new institutions have been created to attend to these issues is something to celebrate, and I am usually very much in favour of philosophy, reflection and attending to ethical issues. I greatly admire many of the people working in this space. It's good that AI ethics is getting plenty of attention (like this recent Time Magazine article[2] or comment in the New York Times[3]). It's good to see governments engaged, like the moves announced at the G7[4] to create an International Panel on Artificial Intelligence (IPAI), or the new European Commission President's promise of EU wide action. So why do I fear that much of what's being talked about, and done, won't achieve a lot?

In this short piece, I suggest five ways AI ethics needs to be radically reshaped if it's to be more useful. I make the case for focusing on: ethics as interpretation rather than code; attending to urgent current problems rather than vague long-term ones; thinking politically; addressing technological complexity; and above all taking outcomes seriously and understanding how ethics intersects with social processes.


## 1. Ethics involve context and interpretation - not just deduction from codes.

Too much writing about AI ethics uses a misleading model of what ethics means in practice. It assumes that ethics can be distilled into principles from which conclusions can then be deduced, like a code. The last few years have brought a glut of lists of principles (including some produced by colleagues at Nesta). Various overviews[5] have been attempted in recent years. A recent AI Ethics Guidelines Global Inventory[6] collects over 80 different ethical frameworks. There's nothing wrong with any of them and all are perfectly sensible and reasonable. But this isn't how most ethical reasoning happens. The lists assume that ethics is largely deductive, when in fact it is interpretive and context specific, as is wisdom. One basic reason is that the principles often point in opposite directions - for example, autonomy, justice and transparency. Indeed, this is also the lesson of medical ethics over many decades. Intense conversation about specific examples, working through difficult ambiguities and contradictions, counts for a lot more than generic principles.

The proliferation of principles and codes may have been a necessary phase to go through. But the real work starts from contextual application not conceptual abstraction.

This is not hard to do. Our own experience at Nesta is that when groups of the public are taken through real questions (such as the dilemmas involved in handling data and AI in health) they can reason in subtle and nuanced ways. They quickly get out of the crude polarisation that often dominates these debates (and so many conferences) which make choices binary - either technological progress or human rights; either AI or privacy. Indeed well-curated conversations turn out to be a lot more useful than mechanistic codes, if the aim is to handle real-world dilemmas. What they generate are best described as 'ethical strings' rather than codes: x is an acceptable use if there are y constraints, oversight from z, and a shared recognition that if trigger q is reached we'll need to reconsider.

Recognising the limitations of code-based approaches to ethics will also be essential if we're to develop viable answers at a global level. Many thinkers in China and elsewhere are now challenging the codes produced by the West, claiming that these are at odds with Confucian, Hindu or Islamic values. When they start with classic Western enlightenment principles they can indeed look quite culturally narrow. But my guess is that once people start working through specific examples, and involving the public, the differences won't be that great.

In short: the proliferation of principles and codes may have been a necessary phase to go through. But the real work starts from contextual application not conceptual abstraction.

## 2. It needs to attend to live dilemmas

So far, there has been too much talk about interesting but irrelevant future questions, and not enough about harder current ones. Fretting about the singularity is one example. The other classic current example is the proliferation of writing and surveys on 'trolley problems': should an AI in a driverless car save the life of the owner or a pedestrian, an 80 year old or an 8 month old? These questions are quite fun. But no designers are really working with them and very few humans have ever had to make these choices.

Meanwhile a host of current AI algorithms are making crucial decisions about medical treatments, probation or mortgages, with very little attention or scrutiny. It would be far better if smart minds could be directed to live questions, like: how should facial recognition algorithms be overseen or made accountable (a topic[7] being addressed by the new Ada Lovelace Institute), or how should regulators treat targeted social media advertising if there's evidence that it fuels compulsive or unhealthy behaviours (one of the issues[8] being addressed by the UK government's Centre for Data Ethics and Innovation). These more fine-grained analyses quickly confirm the limits of deductive analysis; and they also hint at the likelihood that the 'right' answers will change over time, in light of experience.

## 3. Ethics is often unavoidably political.

There is a small industry of AI ethics experts, and significant amounts of funding flowing into it, particularly in Silicon Valley. But what has it achieved? The honest answer so far is: not much. Where there has been serious impact on the ethics of AI, it has mainly come from journalism (like the exposure of Cambridge Analytica), activism (like the many moves against autonomous weapons), bureaucrats (like GDPR), or detailed academic analysis of abuses. Not much has been achieved by the professional AI ethicists. In some cases that was because they were already too co-opted by the big corporates (many of which, ironically, now have AI ethics policies but ethics policies on nothing else).

One main lesson is that ethics is unavoidably political. If you want to influence ethical behaviour you may have to think as much like a political activist as like an armchair philosopher.

---

A few advisers belatedly realised that they could exercise more power by challenging or resigning rather than being coopted. Ann Cavoukian, one of the world's leading advocates for privacy by design, resigned last year from Google's Toronto Sidewalk Labs because she felt she could exercise more influence that way than through remaining involved. But as far as I am aware, none of the AI ethics advisers to Facebook resigned or used any of their influence publicly to shift the company's behaviour. Instead all the serious challenge came from activists.

Looking to the future, it's clear that many of the most difficult issues around AI will be simultaneously ethical and political: how to handle very unequal access to tools for human enhancement; how to handle algorithmically supported truth and lies; how to handle huge asymmetries of power.

One main lesson is that ethics is unavoidably political. If you want to influence ethical behaviour you may have to think as much like a political activist as like an armchair philosopher.

## 4. The field needs to be more self-critical

The first test of any code is whether it can in practice be applied. If it can't, it's no more than signalling. Unfortunately many of the codes fall apart on serious inspection, particularly where AI is drawing on unstructured training data.

In these cases the requirements to reveal training data, identify bias, respect privacy, or respect GDPR, are hard if not impossible. This is even more true for dynamic applications – like Google Search – which are constantly evolving with live data, and are now far beyond the capacity of any human mind to understand.

As with any science, however, facing up to these limits is useful. Discrepancies force harder thinking. In this case, recognising the limits of the codes should encourage more rigorous and more imaginative thinking; and pressure us to go beyond the comfortable homilies of the current codes and lists of principles.

## 5. Ethics needs to connect to outcomes.

My final worry is that the sudden bubble of spending and activity on AI ethics is taking resources away from the, potentially just as important, topic of AI outcomes. I've written elsewhere on why I think the AI industry is repeating past mistakes[9] made by digital industries: in particular focusing far too much on inputs and far too little on outcomes. I wish a bit of the energy and money going into AI ethics could be spent on trying to work out how AI, combined with human intelligence, could best improve health, education, the environment or the economy.

A big lesson of past technologies is that they were socially shaped. They didn't just land on passive societies. Instead societies channelled, constrained, blocked or twisted technologies in numerous ways, and always with an ethical dimension. For example, the many constraints put on cars - from driving tests to speed limits, emission rules to drink-drive laws - were in part ethical, but better understood as part of a social conversation to work out how societies could get the advantages of a new technology without too many harms. None of the rules that resulted could be deduced from abstract ethical principles or codes. But all of them had an ethical dimension.

I would love to hear from people involved in the AI ethics field in response to my five suggestions. None of my arguments is a reason for ditching AI ethics. Quite the opposite. Serious, reflective investigation of the rights and wrongs of AI will become ever more vital.

But we need a shift of approach. The moral of the story is in some ways quite simple (and one mirrored in the conclusions of the recent Nature overview[10] mentioned earlier). Ethics is more a habit or muscle, than a code (in either sense). It's a way of thinking and reasoning, not a rigid framework. It's nurtured in real life contexts and built up more like case law than a constitution. We need – all of us – to become much more fluent in this. But the current approaches risk getting in the way of, rather than helping, these vital conversations.

*[This note on ethics was written to complement other recent pieces I've written on AI, including proposals on how AI could be regulated[11], proposals on data trusts[12]; an overview of how governments should use AI[13]; another on AI and civil society[14]; and why I think AI research should focus more on outcomes rather than just inputs.[15] An overview of Nesta work on AI can be found in this post[16]. And my broader argument for a new discipline and practice around intelligence is set out in my book Big Mind[17] which is out in paperback this autumn]*

## Appendix of Links

1. https://www.nature.com/articles/s42256-019-0088-2
2. https://time.com/5659788/ai-good/
3. https://www.nytimes.com/2019/09/06/opinion/ai-explainability.html
4. https://www.nature.com/articles/d41586-019-02491-x
5. https://link.springer.com/article/10.1007/s11023-018-9482-5
6. https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/
7. https://www.adalovelaceinstitute.org/beyond-face-value-public-attitudes-to-facial-recognition-technology/
8. https://www.gov.uk/government/publications/interim-reports-from-the-centre-for-data-ethics-and-innovation
9. https://www.nesta.org.uk/blog/intelligence-outcome-not-input/
10. https://www.nature.com/articles/s42256-019-0088-2
11. https://www.nesta.org.uk/blog/machine-intelligence-commission-uk
12. https://www.nesta.org.uk/blog/new-ecosystem-trust/
13. https://www.nesta.org.uk/blog/roadmap-ai-10-ways-governments-will-change-and-what-they-risk-getting-wrong
14. https://www.nesta.org.uk/blog/civil-society-and-fourth-industrial-revolution/
15. https://www.nesta.org.uk/blog/intelligence-outcome-not-input/
16. https://www.nesta.org.uk/blog/nestas-work-artificial-intelligence-maximising-public-benefit
17. https://www.amazon.co.uk/Big-Mind-Collective-Intelligence-Change/dp/0691170797