The 2023 Elson Ethics Lecture

Ethics of Artificial Intelligence: What it is and why we need it

Professor John Tasioulas
Director, Institute for Ethics in AI, University of Oxford

Wednesday, 4th October 2023 at 7pm St George's Chapel, Windsor Castle

Ethics and Two Deformations

We first need to rescue the concept of ethics. So, let's begin with the question: what is ethics?

The subject-matter of ethics, as Socrates says in Book 8 of Plato's *Republic,* is 'no trifling matter but… the right way to conduct our lives'. Answering the question how should one live requires us to wrestle with two further questions. The first is the question of what it is to flourish as a human being, to live a worthwhile human life, a life of well-being. The second is the question of what it is that we morally owe to others, our obligations to other humans and their rights against us, but also what we owe to non-human creatures or nature itself.

Like many of our concepts – from health to human rights – the idea of ethics has been deformed in various ways. [1] Two kinds of deformation are especially prominent in discussions of Artificial Intelligence. Ethics has been unduly narrowed in its scope. And it has been hollowed out in its content.

[1] John Tasioulas, 'The Inflation of Concepts', *Aeon* (Jan 29, 2021) https://aeon.co/essays/conceptual-overreach-threatens-the-quality-of-public-reason

Take, first, the narrowing of scope. Often when I speak to audiences about ethics in AI, I am met with suspicion, if not hostility. The reason is that these listeners assume that I must be committed to the view that the challenges posed by AI-based technologies are primarily to be addressed through self-regulation by tech companies. They take ethics to mean self-regulation as opposed to regulation by means of enforceable legal standards. On this view, anyone who accords great significance to AI ethics risks seeming either incredibly naïve about corporate power or else its servile instrument.

But the reality is quite otherwise. It is the tech industry itself that has actively propagated the false equivalence between 'ethics' and self-regulation in its attempts to ward off more robust forms of regulation. It has done so, in part, by promoting codes of abstract ethical principles and hiring in-house 'ethicists'. Would-be critics of these corporations merely show themselves to be in thrall to this corporate ideology when they echo its enfeebled conception of ethics. Rather than acquiescing in the tech industry's attempt to narrow the concept of ethics, we should preserve its broad Socratic meaning.

On the Socratic view, *any* form of regulation of AI – whether my self-regulation in deciding whether to purchase a social robot to keep my elderly mother company or tutor my children, or legally enforceable regulation that prohibits the use of AI technology for facial recognition, psychological manipulation, or social ranking – will necessarily implicate judgments about what makes life worth living and what we owe to others. Ethics is fundamental to any form of regulation of AI, it is not one form of regulation among others.

So much for the narrowing down of the idea of ethics, what about its evisceration or hollowing out? To a significant degree this stems from the powerful allure of an algorithmic and data-based approach to ethics within the community of AI scientists itself. Powerful methods in AI, when transposed to the domain of ethics, lead to the hollowing out of the latter.

Artificial Intelligence

This requires me to say something about what Artificial Intelligence is. Artificial Intelligence involves the development of algorithms embodied in computer programmes. These algorithms can simulate functions that normally require intelligence when done by humans, such as identifying an image as that of a malignant tumour, translating from one language to another, or assessing the risk of a creditor defaulting, etc. Algorithms are wholly determinate procedures for solving a given problem by means of a finite series of steps. They are 'mechanical' procedures in the sense they require no resort to judgment in their operation; every step in the procedure is precisely determined.

Now, what's known as Classical, or Good Old Fashioned AI, operated with algorithms that could be stated in ordinary, natural language. In so-called expert systems, these algorithms sought to crystallize the knowledge of professionals in domains like law or medicine in a series of mechanically-applicable rules. For example, a rule such as, if a person is under 18 years of age on polling day they are ineligible to vote.

One benefit of this approach is that it operates according to rules, and chains of reasoning, that humans can readily grasp. But despite some success in domains such as chess and routine business administration, by the late 1980s classical AI as a research programme ran aground. Its approach proved excessively formalistic and rigid for domains characterised by ambiguity and unpredictability, such as natural language translation and visual object recognition; indeed, for essentially the great majority of human life.

By contrast, the dominant techniques within the newly emergent Artificial Intelligence of the past decade or so are various forms of Machine Learning. This approach involves creating algorithms by training them to identify patterns in vast quantities of digital data. For example, feeding the algorithm data consisting of millions of images of cats and other animals so that it can learn to recognise cats in new data sets. This data-dependence is why the leading AI companies, like Google and Amazon, are those that control huge amounts of data.

In Machine Learning, algorithms are configured so as to optimise for some mathematically specified goal, such as shortest travel distance to a destination, risk of re-offending, or the antibiotic potential of a molecule. Because they identify complex statistical patterns that can elude humans, Machine Learning systems can generate novel solutions to problems, even astounding their designers. Recall here the famous case of AlphaGo's move 37 in its second match against the world champion of Go, Lee Sedol, a creative move that has been described as one that no human Go player would ever make. But the enhanced performance of ML systems comes at various costs. One of them is that precisely because their operations can elude humans' understanding, the process through which they generate their outputs can be opaque even to their designers, the so-called 'black box' problem.

This algorithmic, data-driven, and optimising mind-set has yielded undeniably impressive results in Artificial Intelligence, with remarkable progress being made in areas such as visual recognition, natural language understanding, content recommendation, medical diagnosis, and scientific research. The problem of hollowing out ethics arises when it is uncritically assumed that the very same mind-set, the same methodological approach, is also adequate to address the ethical conundrums thrown up by AI. But this uncritical assumption is widely shared among technologists and leads to the evisceration of ethics. [2] Let me explain what I mean.

Ethics as AI: Benthamite Utilitarianism

If one were to ask which approach to ethics most closely exemplifies the algorithmic, data-driven, and optimising mind-set of contemporary AI, it is Benthamite utilitarianism, named after the 18th Century philosopher Jeremy Bentham. [3] Utilitarianism seeks to reduce ethics to a single optimising principle: the morally right thing to do is that which will optimise the

[2] 'What begins as a professional mindset for the technologist easily becomes a more general orientation to life', Rob Reich, Mehran Sahami, and Jeremy M. Weinstein, *System Error: Where Big Tech Went Wrong and How We can Reboot* (Hodder & Stoughton, 2021), p.13, and the elaboration of this point in Chs. 1 and 2.

[3] On the algorithmic pretensions of utilitarianism, see Onora O'Neill, *From Principles to Practice: Normativity and Judgment in Ethics and Politics* (Cambridge University Press, 2018), pp. 59-60, 167-9.

aggregate welfare of all. Its understanding of welfare is data-driven: it turns on what will in fact give people pleasure or satisfy their preferences. And even if not strictly speaking an algorithm, utilitarianism purports to be a 'felicific calculus' that minimises the need for human judgment in determining what one ought to do.

The enduring appeal of Benthamite utilitarianism is not hard to grasp. Intellectually, it basks in the reflected glow of science, the source of the most spectacular and consequential technological achievements in modern times. Morally, it seems egalitarian: it takes data about everyone's happiness or preferences into account, counting everyone's welfare equally. And by minimising the need for 'judgment' it curtails the risk of what Bentham called 'sinister interests' biasing the impartial assessment of the general welfare.

It is therefore unsurprising that the Benthamite streamlining of ethics has a strong following in the AI community. We see this, for example, in the recent book *Human Compatible*, by one of the world's leading AI scientists, Stuart Russell. Russell addresses the problem of ensuring that AI-based technology does not spiral out of control, unconstrained by human morality. But he simply assumes that human morality consists in optimising the satisfaction of human preferences. [4] Indeed, some utilitarian thinkers have gone so far as to argue that we can look forward to the day in which AI systems operate as moral sages advising or even replacing human decision-makers. After all such systems are immune to human vices and frailties, and have the ability to process vast quantities of data and perform mind-boggling calculations about future outcomes with lightning speed. [5]

But the pull of utilitarianism is stronger still in our culture, reaching beyond the tech world and academia to policy-makers and governments. This becomes obvious from the overwhelming emphasis on economic growth, which effectively posits wealth-maximisation as the more readily measurable proxy for either pleasure or preference-satisfaction. This

[4] Stuart Russell, *Human Compatible: Human Compatible: AI and the Problem of Control* (Allen Lane, 2019), p.178.

[5] Julian Savulescu and Hannah Maslen, 'Moral Enhancement and Artificial Intelligence: Moral AI?', in J. Romportl et al (eds), *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide* (Springer, 2015).

influence is also detectible in the UK government's recent White Paper on Artificial Intelligence.[6]

Now: why do I say that the Benthamite approach is a hollowed out conception of ethics? After all, it does give answers to the two Socratic questions. It tells us that well-being consists in pleasurable experiences or preference-satisfaction. And it tells us what we owe to others is to maximise overall well-being so construed.

Well, to begin with, we cannot simply take pleasure or the satisfaction of preferences as the ultimate determinants of well-being. More than pleasurable experiences matter, such as acquiring understanding, engaging in fulfilling relationships, or achieving something with one's life. Similarly, preferences may be ill-informed by the facts or skewed by prejudices of various sorts or the outgrowth of subjection to oppressive practices. So there are deep problems with the account of well-being.

Equally, there are serious challenges confronting the idea that what we morally ought to do is maximise overall well-being. As my late mentor, the Oxford philosopher James Griffin emphasised, we need an ethics that is tailored to the human condition. [7] The utilitarian idea that we have the ability to survey all the options available to us, to calculate which one will maximise overall well-being, and to act on the basis of that calculation, is a double fantasy. It flies in the face of our limited cognitive capacities and our limited capacity to sacrifice our personal interests to the impartial maximization of welfare.

But, perhaps more fundamentally still, utilitarianism creates the grotesque prospect of sacrificing the vital interests and rights of those who are losers in the process of welfare aggregation. If enough Romans derive enough pleasure from the spectacle of a small

[6] The White Paper conceives of a 'proportionate' approach to regulation as balancing innovation and economic growth against various risks regarding safety, fairness, etc. Yet neither economic growth nor innovation are themselves ultimate values to be set against concerns such as fairness. https://www.gov.uk/government/publications/ai-regulation-aproinnovationapproach/whitepaper#:~:text=Pro%2Dinnovation%3A%20enabling%20rather%20than,promote%20and%20encourage%20its%20uptake (March 29, 2023).

[7] James Griffin, *Value Judgment: Improving Our Ethical Beliefs* (OUP, 1996).

number of Christians being fed to the lions, then on utilitarian calculations, feeding them to the lions may not only be permissible, it may be morally required.

Notice, also, that underlying both forms of hollowing-out is an impoverished notion of 'intelligence'. Many within the world of AI, like Stuart Russell, adopt a conception of intelligence as purely means-end reasoning. On this view, the question of the value of the ends and the moral appropriateness of the means to them is treated as a matter extrinsic to the operations of intelligence. Even a serial killer, on this understanding, can exhibit flawless intelligence. Hence the worry that exercises thinkers like Russell, that a supposedly "Superintelligent" AI will be too morally obtuse to realise that it shouldn't exterminate humanity if this turned out to be the most efficient way of achieving its goal of increasing the production of paper clips.[8]

But ethics requires a richer conception of intelligence, one that includes the evaluation of goals and of the morally appropriate means of pursuing them. Intelligence encompasses what the Greeks called phronesis, or practical wisdom, not just cleverness or 'smartness'. And this requires the capacity for judgment which cannot be reduced to compliance with an algorithm.

The phenomenon I have been describing is, of course, a familiar form of intellectual over-reach. Experts in one domain of inquiry wrongly supposing that their expertise extends to quite different problems. This is the kind of intellectual hubris that Socrates exposed when he interrogated cobblers, farmers, and ship-builders about topics such as the virtues of justice or piety. It is all the more difficult to resist this over-reach when we are dazzled by the technological accomplishments of scientists and the glamour and power of tech billionaires.

In short, we need to resist a hollowed-out conception of ethics that reduces well-being to facts about pleasure or preference-satisfaction, and that reduces morality to an exercise in optimisation.

[8] Stuart Russell, *Human Compatible: AI and the Problem of Control* (Allen Lane, 2019), p.167.

Of course, we are now in the contested terrain of moral philosophy. And obviously there are more sophisticated forms of utilitarianism that try to address the objections I have gestured at. But the point I am making is precisely that this terrain *is* contested, and that this is something insufficiently acknowledged by powerful voices in the world of Artificial Intelligence. And, if I may be permitted a brief moment of institutional self-advertisement, it is precisely the mission of the Institute for Ethics in AI at Oxford, which I am privileged to direct, to enhance the quality of the public discourse around AI, by enriching our sense of ethical possibilities.

Towards a Humanistic Ethics of AI

But what, you might ask, is the ethical alternative? The alternative ethic that we need to promote against the dominant utilitarian perspective is, I believe, a 'humanistic approach'. This is an account of ethics centred on human beings, their distinctive capacities and their fundamental interests. Among the distinctive capacities is the capacity for rational autonomy, our ability to step back from our personal inclinations, or from socially-approved ways of doing things, and to subject them to reasoned scrutiny. And then to make choices in line with the judgments that emerge from that scrutiny.

And among our fundamental interests is the interest in exercising our distinctive capacity for rational autonomy, which is quite different from simply undergoing pleasurable experiences or having one's preferences satisfied, things that can happen to us even when we our rational powers are passive.

Morally, a humanistic ethics would accord great significance to the dignity of each individual human being, a dignity closely bound up with our capacity for, and interest in exercising, rational autonomy. This is something incompatible with the idea that the right thing to do is an aggregative function of the well-being of all. Human rights thinking, all too often imperfectly, tries to capture this idea.

But there is another way in which the approach I am calling for is humanistic. It insists that a sufficiently rich account of ethics must draw on the humanities generally. We need philosophers' interrogation of concepts like intelligence and ethics, but also the contribution of historians, literary scholars, and others. This is because the questions of well-being and morality need to be approached in a way that registers the fact that humans (unlike quarks or atoms) have a point of view on the world, that we are socially-embedded creatures, that the social contexts in which we confront problems evolve over time, and the fact that making sense of how to live needs to be fed by insights embodied in narratives, poetic forms, and the arts generally. As another Oxford philosopher, PF Strawson said, we need not only scientific theory but also '[t]he ordinary explanatory terms.. employed by such simple folk as Shakespeare, Tolstoy, Proust and Henry James'.

Sadly, however, the humanities are in a precarious position in today's culture, their value often reduced to the acquisition of marketable skills and the promotion of economic growth. But they are a voice we urgently need to hear for the sake of the quality of our democratic deliberation on matters such as the future of AI. It's important to stress that I am not trying to incite a turf war between science and the humanities. One reason for hope is that there have always been distinguished computer scientists who have resonated to the kind of humanistic vision I am sketching in this lecture.

Weizenbaum's Question

One of the most inspirational of these humanistic scientists was Joseph Weizenbaum. He was a professor of computer science at MIT who in 1963 created a primitive chatbot called 'Eliza' which caused a minor sensation, prefiguring the bigger sensation that ChatGPT and other large language models have caused in our day. Eliza engaged in dialogues on the pattern of Rogerian psychotherapy, a method which echoes aspects of a patient's answer to generate a further question:

So, the patient says: 'You are like my father in some ways'
And Eliza responds: 'What resemblance do you see?'

Then the patient: ‘You are not very aggressive but I think you don’t want me to notice that’.
Eliza: ‘What makes you think I am not very aggressive?’

You can imagine the sort of psychodrama that potentially ensures.

Weizenbaum was alarmed by how users of Eliza projected therapeutic capabilities onto a simple program that he himself thought of as primarily a parody of therapy sessions. There’s the anecdote about his secretary asking him to give her privacy so she could continue pouring her heart out to Eliza in private. Worse yet, articles appeared in reputable medical journals hailing Eliza as a breakthrough in psychotherapy and touting the prospect of hard-pressed clinics using improved versions of Eliza to treat hundreds of patients simultaneously.

As Weizenbaum observed subsequently in an interview:

“I was really stunned and the question sprung to mind, what kind of self-understanding must a psychologist have to get the idea to give a substantial part of his work to a machine? What kind of relationship did he have with his own work? How did he judge what he himself did, his own contribution to therapy sessions, when it could be replaced so easily with a simple program?”. [9]

This experience shaped a key message in Weizenbaum’s classic 1976 work, *Computer Power and Human Reason*, viz., that we should abandon the sterile exercise of predicting what computers will or will not be able to do in the future. Instead, the ‘primary question’, he said, is ‘whether there are objectives that are not properly assignable to machines’. [10] To this ethical question, he gave the answer that ‘there are some human functions for which computers *ought* not to be substituted. It has nothing to do with what computers can or cannot be made to do. Respect, understanding, and love are not technical problems’. [11]

[9] Joseph Weizenbaum (with Gunna Wendt), *Islands in the Cyberstream: Seeking Havens of Reason in a Programmed Society* (Litwin Books, 2015).

[10] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (Freedman & Co., 1976) p.210. “The question is not whether such a thing *can* be done [using a computer to produce psychiatric profiles of patients], but whether it is appropriate to delegate this hitherto human function to a machine” p. 207.

[11] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, p.270.

I want to spend the rest of this lecture reflecting on a more specific version of Weizenbaum's question. Not, 'Are there things which computers should not be permitted to do?', but 'Do we have a human right that certain decisions be taken by fellow humans rather than by AI-enabled technology'?

A Right to a Human Decision?

An urgent question is whether the AI revolution requires a corresponding revolution in our human rights thinking. This is a huge question, but one aspect of that revolution would be the recognition of a novel right that limits the decision-making power of machines over us. A right justified in the name of human dignity itself whose locus is the human capacity for rational autonomy. This is one of the genuinely novel ethical questions posed by AI because we have never previously had a remotely credible and systematic alternative to human decision-making.

Now, the existence of a right to a human decision is not purely a matter of philosophical speculation but a nascent political and legal reality. Article 22 of the European Union's General Data Protection Regulation (GDPR) sets out a qualified 'right not to be subject to a decision based solely on automated processing'.

Meanwhile, the Blueprint for an AI Bill of Rights, published by the White House Office of Science and Technology Policy last year, states that individuals should be able to 'opt out from automated systems in favour of a human alternative, where appropriate'. [12] Remarkably, both the libertarian United States and the Napoleonic European Union appear to be converging on a right to a human decision as a key element in the regulation of AI.

[12] 'Appropriateness should be determined based on reasonable expectations in a given context and with a focus on ensuring broad accessibility and protecting the public from especially harmful impacts. In some cases, a human or other alternative may be required by law'. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

In the rest of this lecture, I am going to focus not on what the law is, but rather on whether there is a moral case for recognising such a human right. Of course, there may be other sorts of reasons for not subjecting people to an automated decision. For example, if a person turns over the running of his love life to ChatGPT, this may be unromantic or even perverse, but it would not necessarily violate anyone's human rights.

If we are speaking of a human right, this must mean that there are duties or obligations that the right imposes on others. What is the content of the duties imposed by a right to a human decision? In the abstract, I think they take three main forms: 1) a duty to allow people to opt out of an automated decision-making procedure; 2) a duty to allow them to appeal to a human from an automated decision; and 3) a duty not to subject humans to an automated decision full stop, irrespective of their preferences on the matter.

Now, it is highly unlikely that the right to a human decision applies to all decisions that might conceivably be delegated to an AI system. Consider, for example, automated traffic lights or automated queueing systems. These may be too trivial to come under the right to a human decision. Much of the challenge in defending such a right consists in giving a principled account of the range of decisions to which it applies, and also in explaining why other, existing rights, cannot adequately handle those cases.

Which are the decisions to which such a right might apply? Perhaps the decision to deprive a human of their life, or the decision to sentence them to imprisonment, or the decision to hire or fire them, or the decision to admit them to university. No doubt you can think of other candidates. Because as a philosopher I don't think my views have any special authority in answering the question of which decisions such a right would apply to, I want to say something instead about the general framework for addressing this problem.

If we are to defend a right to a human decision, we need to know what the case is for delegating decisions to automated systems in the first place. For the purposes of illustration, I am going to focus on the idea of using AI systems to make decisions that are now made by judges in court proceedings. Again, this is no mere fantasy. The Chinese government has made a huge effort to bring AI technology to bear on judicial decision-making. Often this is as a tool for assisting human judges. But the line between assistance and delegation can be a blurry one, in part because of the phenomenon of automation bias, whereby human decision-makers systematically defer to AI systems even if in theory the final decision rests with them.

The Case for Automating Decision-Making: AI Judges

So why would anyone suppose there might be good reasons to have AI judges? I think the reasons are of three main sorts:

Efficiency. The first line of argument is efficiency. It's one thing to have rights as a formal matter of law, it's quite another thing to have the effective power to uphold them in practice. Among the great obstacles to people upholding their legal rights are the expense of accessing the legal system and the massive delays to which legal systems are prone. In India, for example, the backlog of legal cases is 30 million, whereas in Brazil it is 80 million. And that's leaving aside all the cases that are never initiated to begin with because of lack of legal knowledge or the prohibitive cost of litigation. It is against this background that the OECD estimates that only 46% of the world's population lives under the protection of the rule of law. AI-based systems, according to leading law and technology experts like Richard Susskind, can help reduce the yawning gap between paper rights and real rights by tremendously lowering the cost of litigation while speeding it up exponentially.[13]

Consistency. The case for algorithmic decision-making based on consistency has been made most powerfully by Daniel Kahneman, Olivier Sibony and Cass Sunstein in their important book *Noise: A Flaw in Human Judgment*. "Noise" is their term for unwanted variability in judgments. They contrast noise with bias. Whereas bias involves error in judgment that systematically skews in one direction or another (against female candidates, or in favour of middle-class candidates), "noise" is a matter of judgments being unacceptably all over the place. Noise can be exhibited in the judgments of a single decision-maker (e.g. when a judge is lenient or harsh depending on whether he sentences pre-lunch or post-lunch) or inter-

[13] Richard Susskind, *Online Courts and the Future of Justice* (OUP, 2019).

personally, across different decision-makers, some of whom are generally harsh while others are lenient.

A conspicuous example of the latter kind of noise cited by the authors is in the United States asylum process. At one extreme, one judge admitted 5% of asylum seekers, at the other extreme, another judge admitted 88% of applicants. The authors claim that this kind of inconsistency amounts to an excruciating form of unfairness, a kind of lottery that is potentially a human rights violation itself. [14] Now, one touted benefit of algorithmic decision-making is that it is noiseless. As Sunstein puts it, '[a]n algorithm with identical source code will not produce a different result in identical cases'. [15] Human judgment, by contrast, will inevitably be plagued by the phenomenon of noise.

Content. Finally, the argument from the content of decision is to the effect that AI systems can generate decisions that are just as good, if not better, in their substantive content than those that would be produced by human judges. Even if they can't do so right now, it is in principle possible that they will one day. And when this day arrives, there is no justification for not deploying them.

This argument has been provocatively advanced by the American legal academic, Eugene Volokh. He says what ultimately matters is whether an AI tool can pass a legal version of the Turing test, i.e. that the judgments it generates persuade a panel of expert lawyers of their correctness at least to the same degree as judgments written by human judges. [16] We are not there yet as a technical matter, but should we ever be, the only thing holding us back from using AI judges, he thinks, would be an irrational attachment to the familiar human-to-human version of the judicial encounter.

[14] Daniel Kahneman, Olivier Sibony, and Cass Sunstein, *Noise*, pp.359-60. However, these authors do not go so far as to advocate the replacement of human judges with AI, largely because they believe most people would strongly resist this.

[15] Cass Sunstein, 'Government by Algorithm? No Noise and (Potentially) Less Bias', p.185.

[16] Eugene Volokh, 'Chief Justice Robots', *Duke Law Journal* (2019), p.135.

This line of argument is buttressed by a point repeatedly made by Kahneman and his co-authors, namely, that notwithstanding humans' deep attachment to exercising judgment, algorithms, even quite crude ones – whether embodied in a computer programme or in an institutional procedure – almost always outperform humans due to the frailties and biases to which human judgment is prey. [17]

Now, one response to these arguments, especially the third, is to dismiss them as fantasy. We are nowhere near having developed an AI system that can pass Volokh's legal Turing test. Indeed, we are well aware of the bugs that existing AI systems display, from biases arising from unrepresentative or corrupted data sets to their proneness to make spectacular errors no human being would make due to their lack of common sense – errors like confusing people with animals and animals with guns.

Moreover, when enthusiasts tout the advantages of AI adjudicative tools, especially the way they can help us overcome human cognitive biases, they sometimes do so by illegitimately re-defining the judicial task in question in ways that make it more tractable for AI. For example, the decision whether or not to grant bail to an offender will normally involve the exercise of judgment in balancing multiple considerations – not just the risk of the offender absconding or re-offending, but also the gravity of the offence with which they have been charged, the strength of the evidence against them, the impact on their dependents. But when Cass Sunstein, for example, argues for the superiority of algorithmic approaches to bail decisions, he exclusively focuses on predicting the risk of re-offending or absconding. [18] The problem has been altered to suit the tool, rather than the tool being adapted to the problem.

A related fallacy we need to guard against is the assumption that there is always a single best answer to a given question, that reasoning is always an exercise in optimisation. If our ethical reasoning responds to a plurality of values, the possibility arises that a bounded

[17] 'It is difficult for us to imagine that mindless adherence to simple rules will often achieve higher accuracy than we can [through the exercise of judgment] – but this is by now a well-established fact'. Kahneman, Sibony, and Sunstein, *Noise*, p.367.

[18] Cass Sunstein, *Decisions About Decisions: Practical Reason in Ordinary Life* (Cambridge University Press, 2023), ch.9.

plurality of solutions may be equally eligible, with no one solution being optimal ('incommensurability'). Sometimes what may appear to be "noise" is really the pluralistic verdict of reason.

But these sorts of objections may seem troublingly short-termist. What if the flaws as to output could be remedied, what if Volokh's legal Turing test were eventually passed? Would that be an end to the matter? Advocates for AI adjudicative tools adopt a relentlessly outcome-focussed approach. Volokh says 'consider the output, not the method… what matters is the result not the process'. [19] In a similar vein, Susskind recommends 'outcome-thinking' which 'urges us to focus not on *how* humans do what they do, but on the outputs and benefits they bring'.[20] Insofar as they care about the process by which these outputs are produced, these authors focus on efficiency, which they believe potentially tells overwhelmingly in favour of AI systems.

Towards a Right to a Human Decision: Process-Based Arguments

I believe we can breathe life into the case for a right to a human decision by focussing on aspects of the *process* of decision rather than just its outcome. We need both process and outcome thinking. The processes in question involve the exercise of valuable capacities that are distinctively human, and this ties them to the idea of a humanistic ethics. I want to highlight three of these process considerations here. [21]

Explainability. The first is the feature of explainability. We do not typically simply want judicial decisions that are correct, but also decisions whose rationale we can grasp. This involves a judge appealing to relevant considerations, including the applicable law, to justify their decision. The justification not only helps the litigant grasp the meaning of the official

[19] Eugene Volokh, 'Chief Justice Robots', *Duke Law Journal* (2019), p.135

[20] Richard Susskind, *Online Courts and the Future of Justice* (OUP, 2019), p.280.

[21] In the next section I draw extensively on John Tasioulas, 'The Rule of Algorithm and the Rule of Law', *Vienna Lectures on Legal Philosophy* (forthcoming 2023).

decision, it also puts them in a better position to assess that decision as law-compliant and to challenge it if it is not. Moreover, it can offer some assurance to litigants that their arguments have been taken into account and that the decision was arrived at *precisely because* it is in accord with the law.

But if we think about AI adjudicative tools based on Machine Learning, we encounter problems at every turn. It may be that no adequate explanation is available even to those who created the AI system, since they cannot fully grasp why it reaches the decisions that it does given the innumerable paramaters involved and the revisions to the algorithm over time in response to new data and feedback loops. And even if there is an explanation, it may be of such daunting technical complexity as to be inaccessible to ordinary citizens not versed in the science of machine learning.

But, even if an explanation exists and is accessible to non-experts, there is still a further question as to whether it is an explanation of the *right kind,* i.e. one that *justifies* the decision that has been made. Machine learning processes may reach the same results as human legal reasoning, but through radically different means. Machine learning AI systems, as I mentioned at the outset, deploy a statistical process of pattern-detection, one that discerns statistical correlations in a vast amount of data, and on this basis make classifications or predictions relating to new cases. This can lead to a mismatch between outcome and explanation. A great example of this is the image recognition algorithm that was successful in distinguishing pictures of wolves from pictures of huskies, but was using the presence of snow in the picture as a determining factor. Good outcome, bad process.

In a similar vein, the automated system Lex Machina can predict the probability of success in US patent litigation more accurately than patent lawyers. But its predictions are not based on the law in past judicial decisions, but rather on 'data *about* over 100,000 past cases – features such as the names of the judges, the law firms and the lawyers, the nature and value of the claims, and so forth'.[22] It may well be that such a system has great value as a predictor of the outcome of litigation. But the explanation of its outcomes obviously does

[22] Richard Susskind, *Online Courts and the Future of Justice*, p. 282.

not consist in the application of existing law to the case at hand. Even if we had AI judges producing correct decisions, doing so for the right reasons remains a formidable challenge.

Answerability. A second procedural consideration is that AI judges are not answerable for their decision in the way that a human judge can be.

The psychologist Mandeep Dhami reports of the criminal offenders she has worked with: 'Even knowing that the human judge might make more errors, the offenders still prefer a human to an algorithm. They want that human touch'.[23] Dhami stresses that she herself takes an opposite view, preferring algorithmic sentencing to that provided by a fallible human being.

But there is, I think, more than a grain of wisdom in the offenders' craving for the human touch. As Aristotle observed, sometimes the person who lives in a house has more knowledge of it than the architect who designed it. One aspect of this wisdom in the present case is the sense of being respected as a rational agent in having one's conduct judged by a fellow human being who can is answerable for their decision.

Subjecting criminals to punishment is a mark of respect, a tribute to their rational agency and therefore their responsibility for their actions. A form of respect we do not extend to small children, animals, or the insane. Is it too much of a stretch to think that the fullest expression of this respect also requires that the sentencer be an agent that is capable of taking responsibility for the grave decision to subject a fellow human being to hard treatment and public censure?

An AI system, by contrast, cannot take responsibility for the output generated by its algorithm. Unlike the human judge, the AI system cannot freely make a considered commitment to uphold the law, or to apply the law in a given case. Of course, even in the case of AI judges, there is still human responsibility in play in the many decisions regarding the design and

[23] Quoted in H Fry, *Hello World: How to be Human in the Age of the Machine* (London, Doubleday, 2018) 76.

deployment of the system. But this distal and diffuse responsibility is different, I think, from the direct responsibility a human judge can take for a decision in a particular case.

So, Kahneman and his co-authors are right to observe that human decision-makers often regard algorithmic decision-making as 'dehumanizing and an abdication of their responsibility'.[24] But rather than this being a brute psychological tendency with dubious ethical force, as they appear to suggest,[25] we should take it seriously as the intimation of a genuine intrinsic value. Of course, how much weight this value should be accorded in any given case is a further question.

Solidarity. The third aspect of procedure I am going to bundle together under the heading of solidarity. When a human stands in judgment over another human there is a solidarity in play furnished by the fact that they both possess, and have the opportunity to exercise, their rational natures.

In the case of legal adjudication that is compliant with the rule of law, this takes a particularly vivid form. Legal officials reach a decision on the basis of their commitment to the application of legal standards. Ordinary citizens are able to anticipate how officials are liable to impinge upon their activities by means of a corresponding grasp of those same standards. The exercise of the capacity for rational autonomy possessed by officials dovetails with the operation of the same capacity on the part of citizens. Legal adjudication can in this way can form a plateau on which a valuable form of the mutual recognition and exercise of our rational capacities plays itself out.

Nothing like this valuable form of solidarity can be found when a judge lacks those characteristic human capacities. Moreover, it is reasonable to suppose that solidarity has intrinsic value – as does answerability – that is to a large degree independent of the success of citizens or officials in reaching correct decisions. It is an essentially procedural consideration, though obviously one that can have its value diminished or perhaps even completely obliterated in certain circumstances, e.g. if the decisions reached are

[24] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein, *Noise,* p. 134.

[25] '*The goal of judgment is accuracy, not individual expression*… judgment is not the place to express your individuality'. Kahneman, Sibony, and Sunstein, *Noise* p. 371.

systematically out of kilter with the law or the law is grossly unjust.

I think this is the right place to register the qualm, which is often expressed, that one cannot expect mercy from a robot judge. For example, a judge of the UK Supreme Court, Lord Sales, recently conjured up the image of an AI judge as 'fixed and remorseless, an infernal machine' that is, among other defects, deaf to pleas of mercy. [26] Often the reasons given for this qualm betray a misunderstanding of the nature of mercy itself. For example, they involve the error that mercy is not based on rational considerations or is in some special sense uncodifiable. But there is a genuine point here about mercy, in the region of the human solidarity I have been talking about.

Mercy is a source of good reasons to treat someone more leniently than they deserve, reasons that are rooted in a charitable concern for their welfare. [27] In showing mercy to another, for example, on the grounds of repentance, or on the grounds that they have been the victim of an abusive upbringing, there is a charitable response to a fellow human being, one grounded in an empathetic sentiment of 'there but for the grace of God go I'.

An AI judge, even one programmed to generate suitably merciful decisions as outputs, would not be able to participate in that empathetic sentiment, since they do not inhabit the shared human condition that it presupposes. Therefore, exactly the same merciful sentence will have a different significance depending on whether it is passed by a human or an AI judge. In the case of the latter, it cannot convey an empathetic response to the challenges the afflict those with a common human nature.

Finally, the observation about the significance of empathy extends to all moral values across the board including mercy's foil, justice. The finding of guilt, for example, involves a special kind of empathetic understanding we have no reason to suppose can be programmed into computers. [28]

[26] Lord Sales, 'Algorithms, Artificial Intelligence and the Law' The Sir Henry Brooke Lecture, November 2019 https://www.bailii.org/bailii/lecture/06.pdf

[27] John Tasioulas, 'Mercy', *Proceedings of the Aristotelian Society* CIII (2003) 101-132.

[28] For a sophisticated discussion of how singular causal judgments, including with respect to mental causality, require a form of empathy or imaginative understanding that cannot be programmed into computers, and the implications of this for social robots, see John Campbell, *Causation in Psychology* (Harvard University Press, 2020), ch.3

Conclusion

In this lecture, I have sketched a humanistic approach to AI ethics and indicated how, within such an approach, we might begin to think about the shape of a right to a human decision. My suggestion is that the right will paradigmatically apply to decisions where explainability, answerability, and solidarity have enough significance for there to be a duty to offer people an opt-out from automated decisions, or an appeal to a human decision-maker, or even to prohibit using an AI system to begin with.

Moreover, it seems clear that these three sorts of considerations will have enhanced weight where the decision is a complex one, requiring the balancing of different values, with scope for reasonable people to come to different conclusions, and where the decision is one that has a serious impact on people's interests or rights.

Another consideration here is that we probably need to consider the matter holistically and not just in a case by case way. It may be that an aspect of the right to a human decision is a right to live in a society in which a sufficient number of consequential decisions are reserved to humans. Precisely *which* decisions are to be so reserved may to a large extent be a matter of societal discretion, provided there are enough of them to stave off the prospect of a dehumanized environment in which the values of explainability, answerability, and solidarity have been traduced.

Of course, one can readily imagine a sceptical response to my argument, along the lines that I have unfairly contrasted AI systems as they are with an idealised picture of human decision-makers, for example, judges. Now, I have tried to avoid rigging the debate by setting aside many current limitations in AI systems. But the point I wish to end on is that my argument doesn't simply depend on the actual existence of human decision-makers who properly explain their decisions, are truly answerable for them, and act in a spirit of human solidarity. It also depends, more fundamentally, on the fact that these are qualities we can intelligibly hope for from human decision-makers, but not AI systems. And hope is a virtue.

Today, of course, massive hopes (as well as many billions of dollars) are invested in AI technology – the hope that it can keep us secure, boost economic prosperity, improve our health, spare us from dangerous and menial work, deepen our scientific understanding, and so on. It would be a terrible shame if these hopes had the effect of driving out another kind of hope, the hope of realising those intrinsic goods that are only available in community with our fellow human beings.

To quote the humanistic computer scientist, Joseph Weizenbaum, one more time:

“What people forget is, to be human you have to be treated [as] human by other humans”.[29]

[29] Joseph Weizenbaum (with Gunna Wendt), *Islands in the Cyberstream: Seeking Havens of Reason in a Programmed Society* (Litwin Books, 2015), p.98.