Ethics for AI – way forward or false dawn?

Brian Williamson

Revised January 2020

Abstract

This paper considers the proposed application of regulation based on ethics for machine learning-based algorithms (referred to as Artificial Intelligence or AI) and contrasts this approach with a counterfactual whereby markets, subject to public policy designed to correct market imperfections such as externalities and information asymmetries and subject to political choice, largely determine technology and market choices. Both approaches have ethical underpinnings, and proposals for regulation based on ethics for AI *per se* should be assessed relative to the counterfactual. Indeed, it remains unclear why AI *per se* should be the focus of ethical debate and proposed new rules. Placing a higher regulatory hurdle for machine-based versus human-based approaches to prediction and decision making would result in foregone opportunities for productivity growth, and potentially foregone health and safety benefits. Operationalisation of ethics for AI may also centralise decisions over what technologies are good, rather than relying primarily on decentralised innovation by entrepreneurs coupled with selection pressures, and error correction by venture capitalists and consumers. An explicitly ethical or rights-based approach to regulation may reduce scope for consideration of policy trade-offs in terms of economic welfare and broader political choice. An alternative approach, given the promise of AI, would be to forebear from setting higher standards for machine than human algorithms (individual, institutional and coded software based) whilst focussing on identifying and removing unnecessary barriers to the adoption and use of AI. Proposed new rules should be subject to a cost-benefit test, be made on a politically accountable basis and applied to existing as well as machine-based algorithms where justified. Change is likely required, but regulation based on ethics for AI *per se* arguably represents a policy false dawn.

What is meant by AI in the context of this paper?

AI in this paper refers to machine learning based algorithms i.e. algorithms that ‘learn’ from data. The dawn of deep learning, a powerful technique that transformed the field of artificial intelligence, is considered by many to be the October 2012 ImageNet victory (image recognition accuracy for a standard set of images) over competing algorithms.

Humans also rely on algorithms – genetic and learned. Human intelligence – our algorithms – are more flexible than any existing machine; but are imperfect and inferior in some domains, for example, calculators have long outperformed humans at arithmetic.

This paper focusses on AI which involves the use of algorithms and learning to automate tasks and open up new possibilities for machines; much as automation during the industrial revolution did in relation to manual work and crafts.

Why focus on AI?

An underlying question in this paper is why focus on AI *per se*?

One reason may be fear of a superhuman AI, though it is unclear what superhuman intelligence might mean (Kelly 2017). A further reason for focussing on AI has been framed by Yuval Harari (2015):

> *'What will happen to society, politics and daily life when non-conscious but highly intelligent algorithms know us better than we know ourselves?'*

However, these questions are not the primary focus of work on ethics for AI, at least not in relation to current proposals that may be translated into regulation. The focus of regulation in relation to ethics for AI, and of this paper, is on what is referred to as 'narrow' or 'weak' AI which performs domain-specific or specialised tasks.

Grounds for focussing on AI, which apply to narrow AI, are provided by Whittlestone *et al* (2019):

> *'AI can often be used to optimise processes and may be developed to operate autonomously, creating complex behaviours that go beyond what is explicitly programmed.'*

Yet complex behaviour of systems is not something new or unique to AI, and individual human and institutional behaviour often goes beyond 'what is programmed'. Indeed, going beyond what is 'programmed' is typically seen as a desirable feature of human decision making and institutions, and is the motivation for offering incentives to go beyond what is required.

Another ground for focussing on AI identified by Whittlestone *et al* (2019) is that it has dual characteristics:

> *'ADA-based [algorithms, data and AI] technologies are dual-use in nature: the purpose to which they are initially developed can easily be changed and transferred, often radically altering their moral valence'.*

Yet this is true of technology more generally and general-purpose technologies in particular, from stone adzes to fire, steam power, electricity and computers.

Whittlestone *et al* (2019) also conclude with a policy prescription:

> *'These features, together with the remarkable speed with which powerful private companies have pioneered new applications of ADA-based technologies in the recent decade, explain the increase in focus on the need to regulate and guide the ethics of ADA in the right direction.'*

Yet the speed with which applications have been pioneered is not a reason to regulate and guide AI in the 'right direction'. Rather, reliance on decentralised innovation, prediction and selection/error-correction. Hubris is required in the face of profound uncertainty about what will work and what investors and consumers will judge most beneficial.

Explanations for a focus on ethics and regulation of AI *per se* do not therefore appear convincing. Indeed, the focus on AI is reminiscent of the fear in the late 1980s that nanotechnology might turn the world into 'grey goo' (Drexler, 1986). Focussing on a broad technology class arguably represents a category error, too broad to reflect a specific concern and too narrow as a basis for framing general 'horizontal' law and regulation such competition, consumer or data protection law.

Nevertheless, AI has the potential to offer economic and social benefits, and its promise rather than dystopian fears arguably do justify a specific focus on enabling AI. However, this suggests a different agenda which priorities identifying and removing unnecessary barriers to the application of AI.

Further, if on re-examination new rules or higher standards are justified, they should apply to existing as well as new approaches.

The counterfactual in appraising ethics for AI

Policy appraisal of proposals for ethics and regulation for AI should take into account the human and institutional counterfactual, which is far from perfect. Human decisions are beset by a range of limitations including inconsistency, noise and bias (Kahneman 2011). If AI is held to a higher standard than human decisions opportunities for improvement may be foregone. As the draft White House (2020) regulatory guidance noted:

> *'Agencies must avoid a precautionary approach that holds AI systems to such an impossibly high standard that society cannot enjoy their benefits.'*

The comparison should also take into account market-based prediction and decision making which involve a contestable process of innovation by entrepreneurs, selection by consumers and error correction. In contrast, proposals for the translation of ethics for AI into law and regulation tend to involve reliance on decisions by 'elite' groups about what is good, and would almost certainly involve a far slower process of error correction.

Finally, the comparison should take into account economy wide and application specific (but non-AI specific) policy interventions focussed on reducing market imperfections and harms. These tend to be designed mindful of trade-offs, both political and economic (via policy advice that may include cost-benefit analysis and political decision making which may take into account wider considerations mindful of the preferences of citizens), rather than a view regarding what applications are good or a framework based on rights which makes consideration of trade-offs difficult.

Proposals in relation to ethics for AI

Thinktanks (AINOW 2019, Hofheinz 2018), States (House of Lords 2018 in the UK, Villani 2018 in France and the White House 2020 in the US) and regions (Floridi *et al* 2018, European Commission 2019) have put forward indicative proposals in relation to ethics for AI.

The European Commission President (Ursula Von der Leyen, 2019) has indicating that new law, with explicit ethical underpinnings, will be introduced to regulate AI.

In contrast, The White House (2020) draft guidance on AI does not mention ethics explicitly and frames the issue in terms of conventional regulatory appraisal grounded on an assessment of costs and benefits of specific intervention taking account of a market and general regulatory counterfactual.

Voluntary approaches versus regulation

Companies and other institutions may voluntarily adopt ethics and guidelines in relation to AI to guide development and signal their approach to employees and others as a means of promoting a particular internal culture and external brand.

Yet it has been argued that the adoption of voluntary principles for the ethical use of AI by may not be sufficient to align developer and user interests (Brent Mittelstadt 2019).

However, voluntary approaches can be powerful in creating a focal point for conduct and for informing consumer choice. For example, a company's consistent commitment to privacy may better align employee conduct and inform consumer choice than the necessary, but detailed, terms and conditions for services.

A virtue of voluntary approaches is that they are contestable with innovation and rival approaches subject to selection pressures. However, the approach also has its limits where a voluntary approach alone cannot deliver an efficient collective response, for example in relation to mitigation of greenhouse gas emissions.

The focus of this paper is on evaluating the merits of application of ethical principles for AI *per se* via law and regulation where the approach is no longer voluntary and contestable, but mandated.

Objectives

Objectives for AI have been defined by the various groups considering ethics for AI. Examples include benefit society as a whole, promote the common good, a fairer or more efficient distribution of resources, more societal cohesion and collaboration, elimination of discrimination, global justice and equal access to the benefits of AI.

What precisely is meant by terms such as the good of society is typically left unclear, though specific aims are set out in a number of indicative proposals. For example, Floridi *et al* (2018) propose that the goal in relation to ethics for AI is to guide us towards a 'good AI society' and set out the following objective:

> *'…AI should be designed and developed in ways that decrease inequality and further social empowerment, with respect for human autonomy, and increase benefits that are shared by all, equitably.'*

The European Commission (2019) has proposed ethics guidelines for 'trustworthy AI', putting trust in AI at the centre of the proposals. According to the guidelines, trustworthy AI should be: lawful - respecting all applicable laws and regulations; ethical - respecting ethical principles and values; and robust - both from a technical perspective while taking into account its social environment.

The White House (2020) focusses on innovation and trust:

> *'The importance of developing and deploying AI requires a regulatory approach that fosters innovation, growth, and engenders trust, while protecting core American values…'*

Means

Proposals for ethics for AI include procedural elements and preferences for the way in which AI operates. These include, for example, explicability, a right to explanation and a right not to be subject to a decision based solely on automated processing of data - the latter incorporated in Article 22 of the European General Data Protection Regulation (GDPR) – a regulation developed largely before possible implications for the development of AI were a consideration.

Further, proposals set out by European Political Strategy Centre (EPSC 2018) include a 'human in the loop' principle that AI should augment human abilities but not substitute for them, and periodic tests and retraining to ensure that humans would still be able to perform the task in question in case of a technology breakdown (principles that would not be met for many existing technologies and applications).

Another recurring focus is potential for bias in AI systems, with EPSC (2018) noting that 'While augmenting humans' capabilities, AI can also exacerbate existing power asymmetries and biases.'

Floridi *et al* (2018) propose financial incentives for the development and use of AI technologies that are 'socially preferable' and potential certification for 'deserving' products and services. A European oversight agency for the evaluation and supervision of AI products, software, systems or services is proposed.

The European Commission (2019) ethics guidelines for trustworthy AI put forward a set of key requirements, and a detailed (pilot) assessment list to promote trustworthy AI.

The White House (2020) emphasises reducing barriers to deployment of AI, notes that AI may be deployed in already regulated industries where it 'presumably offers economic potential'. The White House also notes that application of existing law to questions of responsibility and liability for decisions made by AI could be unclear in some instances leading to a need for agencies to evaluate the benefits, costs and distributional effects of any proposed change in terms of accountability.

Evaluation of proposed ethics for AI taking account of the counterfactual

The counterfactual to specific regulation based on ethical principles for AI involves a process of entrepreneurial experimentation and selection via the allocation of capital and consumer choices coupled with general (for example competition law) and specific regulation (for example antidiscrimination law).

The counterfactual, which has ethical underpinnings (the Pareto principle), permits innovation without permission by default, subject to law and regulation and the selection pressures of consumer choice. The counterfactual for government services is different, since these tend not to be contestable and involve limited consumer choice.

Ethical underpinnings of the counterfactual versus specific ethics for AI

Markets place weight on individuals' decisions in deciding what is 'good'. Entrepreneurs and those making capital allocation decisions seek, in order to profit, to predict what will satisfy consumers, whilst minimising the resources required to do so. Consumers select the services they prefer.

Markets therefore have ethical foundations rooted in individual preferences and a decentralised evolutionary process for discovering good outcomes and for error correction. The market process draws on decentralised information and distributes the power to decide widely.

The market counterfactual is also underpinned by focussed intervention in recognition of the moral limits of markets (Tirole 2017). For example, where the pursuit of individual preferences involves externalities i.e. benefits or harm to others not reflected in market prices and decisions.

AI developed and deployed in the absence of specific regulation designed to promote ethical AI would therefore nevertheless be subject to selection pressures founded on ethical principles.

More formally, under idealised conditions, a competitive market leads to a Pareto-efficient outcome - an allocation of goods where there is no possibility of redistribution in a way where at least one individual would be better off while no other individual ends up worse off (Arrow and Debreu 1954). The Pareto principle is, however, silent on the distribution of resources.

In welfare economics and public policy evaluations potential Pareto improvements are considered, namely if one state provides an improvement for one party but causes deterioration in the state of the other, it will be chosen if the winner can compensate the loser' losses until the situation is at least as good as in the initial situation.

Ethics and what constitute 'good' are not new considerations, but features of existing law, institutions, markets and interventions. If an alternative specific ethical standard is proposed in relation to AI, it should be assessed against the ethics embodied in markets, law and regulation.

Productivity growth

> *'Productivity isn't everything, but in the long run it is almost everything. A country's ability to improve its standard of living over time depends almost entirely on its ability to raise output per worker.'* Paul Krugman 1994

Productivity growth - getting more for less - is the driver of growth in real income per hour worked. In 12 Western European countries from 1870 through to the year 2000 productivity increased 10-fold, split between 5-fold real income growth and increased leisure, with no appreciable change in employment per capita despite successive waves of automation (Maddison 2000).

Income and leisure are not the only things that matter to people and society, but productivity growth - getting more for less – has expanded the scope to pursue other goals as well, for example improved sanitation and health care leading to increased life expectancy and increased productivity of industry and renewables contributing to decarbonization of output.

Whilst the details of an environment conducive to innovation and productivity growth is debated, it is widely accepted that productivity growth rests on innovation and resource reallocation, a process of 'creative destruction' (OECD, 2018).

Productivity growth also tends to come in waves with so-called general-purpose technologies including stream, electricity and computing - which can be applied widely - driving successive waves (Jovanovic and Rousseau, 2005). AI has the characteristics of a general-purpose technology, and freedom to experiment is likely to be particularly valuable in discovering fruitful applications of AI.

Yet the checklist proposed by the European High-Level Expert Group on Artificial Intelligence (April 2019) frames job losses related to AI as a risk:

> *'Q52 - Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?'*

Past gains in productivity would not have been possible without the loss of jobs and job reallocation due to automation and new ways of doing things. Proposals for regulation based on ethics for AI should be assessed in relation to the risk that they diminish productivity growth rather than in relation to the risk of job losses; since productivity growth underpins growth in income, leisure and other societal gains.

Trust

The High-Level Expert Group on Artificial Intelligence (April 2019) aim to promote 'trustworthy AI'. Yet other general-purpose technologies including steam, electricity and the internet were widely adopted without prior across-the-board trust, indeed their successful adoption and use arguably depends on a degree of caution regarding who and what to trust.

Consumers need to be able to discriminate between trustworthy and untrustworthy applications and providers (O'Neill 2013), and governments may be able to help ensure consumers have the information they need to discriminate. But trust is not an absolute and uniform requirement.

Trust is contingent, and mechanisms that help us sort the trustworthy from the untrustworthy evolve over time and include legal accountability for outcomes, brands, knowledge of the identity of those you are dealing with and ratings. For example, peer-to-peer online platforms are founded on mechanisms for creating trust on a distributed basis, rather than via top down regulation (Botsman 2017). There is slso evidence that such distributed mechanisms can be superior to top-down regulation (Farronato e*t al* 2020).

Trust is important but cannot simply be imposed. Consumers need to be able to discern who and what to trust and market mechanisms and regulation can both contribute to ensuring they have the information required to do so, but across-the-board trust in a general technology such as AI is neither necessary, nor perhaps even desirable for its adoption and use.

Explicability

Explicability - understanding how something works - may not be necessary, provided it works. Clinical trials illustrate that we can validate an approach even though our understanding of the underlying system is incomplete. For example, the mechanism by which anaesthesia works has been poorly understood, but we gladly embraced it (Perkins 2005).

Explicability requirements in relation to AI, including the general provisions in the GDPR, therefore appear particularly ill-conceived. There is no reason to rule out 'black box' AI (Holm 2019), particularly if it produces better results or is cheaper. A pioneer in the development of machine learning, Hinton (2019), has also cautioned against explicability:

> *'One place where I do have technical expertise that's relevant is [whether] regulators should insist that you can explain how your AI system works. I think that would be a complete disaster.'*

Illustrative examples of applications of AI that are not necessarily explicable include AlphaFold (DeepMind 2018) in relation to protein folding and navigation for Google internet balloons (Wired 2017).

However, for some applications a degree of explicability may be valuable and feasible. An example is the use by DeepMind (2018) of AI to interpret eye scans:

> *'For most AI systems, it's very hard to understand exactly why they make a recommendation. That's a huge issue for clinicians and patients who need to understand the system's reasoning, not just its output – the why as well as the what.'*

The system combines two different neural networks. The first neural network analyses the scan to provide a map of the different types of eye tissue and the features of disease it sees, such as haemorrhages, lesions, irregular fluid or other symptoms of eye disease. This map provides insight into the system's "thinking." The second network analyses this map to present clinicians with diagnoses and a referral recommendation expressed as a percentage, allowing clinicians to assess the system's confidence in its analysis.

Explicability can be an important service attribute; but should in general be left to the market and consumers to assess the trade-offs involved.

Government services are however an exception since they are not subject to a market test. The government therefore needs to decide what is appropriate in terms of explicability, and the trade-off with costs, for machine and human algorithms involved in government service provision.

Prejudice, discrimination and corruption

Prejudice against 'out-groups' is widespread, and a sizeable number of people self-report their own prejudices in surveys, whilst many are subject to unconscious bias (Charlesworth and Banaji 2019). A reminder that we are a 'black box', even to ourselves.

Laws against discrimination are also widespread (though what is regarded as unacceptable discrimination may differ, and some forms of discrimination are considered desirable in some countries yet outlawed in others).

The possibility of AI bias should be compared and contrasted with the risks of bias from continued reliance on human decisions and institutions which are known to exhibit bias. A higher hurdle for the development and use of AI versus the non-AI based recognition, screening and decision making could perpetuate rather than reduce bias.

There do not appear to be grounds for assuming *a priori* that algorithms will make the problems of bias and discrimination worse. Indeed, there are grounds for thinking that AI will help identify and reduce discrimination relative to existing decision making. For example, a customer with experience of bias and the personal costs of long-standing behavioural adjustment to bias, made the following remark in relation to the automated Amazon Go store:

> *'…everyone is just a shopper, an opportunity for the retail giant to test technology, learn about our habits and make some money. Amazon sees green, and in its own capitalist way, this cashierless concept eased my burden a little bit.'* (CNET 2018)

Whilst AI could reflect and amplify existing discrimination, algorithms, in combination with appropriately adapted application of existing anti-discrimination law, also offer the prospect of greater scrutiny of decisions in ways that are not possible with human decision-making, for example, to ask precisely what data was used for training and to experiment with different objective functions and training data. 'Re-educating' AI is also likely to prove easier than re-educating humans, and easier to verify.

The application of existing nondiscrimination law to algorithms without careful consideration of the way in which it is applied could however accentuate bias in some circumstances. The scope to reduce bias can, for example, be greater if an algorithm has access to information about disadvantaged group membership. Kleinberg *et al* (2019) illustrate this possibility:

> *'Consider a firm that is trying to decide which sales people to steer towards its most lucrative clients based on a prediction of their future sales level. Candidate predictors include (1) past sales levels and (2) manager ratings. Suppose that for men, managers provide meaningful assessments that include useful signals about employee performance that are not fully captured in the past sales data. But suppose that for women, the managers discriminate and give the lowest possible ratings. (This is of course disparate treatment and therefore unlawful.) An algorithm that is prohibited from knowing gender might well use manager ratings as a predictor, because it has a useful signal for half the sample. And because the algorithm in this case does not know who is male and who is female, it has no choice but to assume that manager ratings mean the same thing for all workers. The resulting predictions would understate the future productivity of women and hence contribute to gender gaps in earnings.*
>
> *But what happens if we instead allowed the algorithm to be aware of gender? With adequate training data, the algorithm could detect that manager ratings are predictive of future sales for men but not for women. Since the algorithm is tasked with one and only one job – predict the outcome as accurately as possible – and in this case has access to gender, it would on its own choose to use manager ratings to predict outcomes for men but not for women. The consequence would be to mitigate the gender bias in the data.'*

In this example allowing the algorithm to access protected group status information can help reduce bias and prohibiting such access on the grounds that it would contribute to bias (as it might for a prejudiced human decision maker) would perpetuate bias. Flexibility as to how bias is reduced is therefore desirable.

Finally, AI itself is not corruptible, whereas humans may bias their decisions in return for reward. For example, a London based peer-to-peer car service driver mentioned that he preferred the service to previous work for a taxi company, because pick-ups were algorithmically determined based on proximity and willingness to accept a request, rather than kickbacks to call dispatch operators (Pers comm 2018).

Society has rightly sought to address prejudice, bias and discrimination via social change, legal requirements and sanctions. It is also the case that AI trained based on data which incorporates bias may reflect bias and potentially amplify it. However, automation can also remove opportunities for discrimination and corruption, and AI may prove easier to check for bias and to 'retrain' than humans and human institutions.

The application of law, regulation and policy may however need to be adapted to be effective in relation to machine versus human algorithms. For example, hiding the identify of individuals may help reduce discrimination by humans but hinder non-discrimination for algorithms which may utilise identity to identify and correct for discrimination revealed in training data. The details matter and a one size fit all approach to is unlikely to be optimal.

Concentration of power and the challenge of prediction

Expert proposals for ethics for AI, and rules applicable to AI, involve a concentration of power (subject to a degree of political accountability) versus a counterfactual in which decisions are decentralised i.e. left with entrepreneurs and consumers. Whilst market power is also a possibility, it held in check via the prospect of market entry and application of anti-trust law.

Markets help ensure that multiple judgements and bets regarding the future emerge and are tested, that decentralised information is harnessed, and rapid error correction occurs. Regarding prediction, there is a long history of people, including experts and the inventors of technologies, making wildly wrong predictions, as this illustrates:

> *'The coming of the wireless era will make war impossible, because it will make war ridiculous.'*
> *Guglielmo Marconi, 1912*

One should be cautious about vesting too much power with an individual or committee and ensure that prediction is contestable, and that predications are ruthlessly tested. Both the scientific method and markets embody this principle, with tests by other scientists against nature; and by investors against their expectations and consumers against their preferences respectively. These are decentralised processes in which there is no single authority. Political decision making may also be subject to contestability via periodic elections.

Evolution by natural selection also illustrates how even a blind process of trial and error, provided there is a correction mechanism, can achieve good design. Markets more closely mimic this process than centralised judgement, since the set of innovations is larger and the feedback process resulting in growth or elimination is swift and ruthless.

A benevolent committee of experts, no matter how well intentioned and informed, cannot replicate this process. Some stakeholders may be consulted regarding their preferences, however this is no match for the power and diversity of individual choice in revealing preferences. Some predictions and decisions must be made centrally with limited contestability, but the power of decentralised prediction, innovation and selection is a reason to limit centralisation of 'authority' where possible.

From a liberal perspective, the rights of the individual, individual preferences and competing ideas are central to a dynamic process that brings about better outcomes (The Economist 2018). Further, a key element of liberalism is a distrust of power, particularly concentrated power. Regulation based on ethics for AI would tend to entrench a particular view of what is good, slow experimentation and error correction and entrench the *status quo*.

Policy synthesis

At a general level it is not clear why a general-purpose technology such as AI should be the focus of ethics or regulation. AI is hardly unique in terms of its wide application and potential power. As Tom Standage (2018) put it:

> *'…given how widely applicable AI is—like electricity or the internet, it can be applied in almost any field—the answer is not to create a specific set of laws for it, or a dedicated regulatory body akin to America's Food and Drug Administration. Rather, existing rules on privacy, discrimination, vehicle safety and so on must be adapted to take AI into account.'*

Further, problems that may warrant intervention tend to either be broader (data protection or discrimination more generally) or narrower and not necessarily specific to machine algorithms (use of facial recognition by machines or humans). We need greater clarity as to why we are focussing on AI at all.

If new regulation is proposed, whether for AI, algorithms more generally or some broader issue motivated by consideration of AI; then it should be tested against a clear counterfactual. An appropriate counterfactual is markets with a default of innovation without permission coupled with orthodox targeted public policy intervention designed to reduce market imperfections or harms and assessed on a cost benefit basis augmented by political decision making and accountability. Both markets and cost-benefit analysis are grounded in ethical underpinnings, namely the Pareto principle; and whilst policy decisions may depart from the Pareto principle, they do so in a manner that is mindful of citizens preferences and involves a degree of political accountability and contestability.

An ethics or rights-based approach to AI regulation would tend to deny the possibility of trade-offs (other than between competing rights or ethical principles). Such an approach would also tend to substitute centralised for decentralised prediction and error correction i.e. it foregoes the benefits of a more Darwinian process. Specific ‘ethical’ principles, for example, a right to explanation, could also deny citizens the opportunity for improved ‘black box’ medical diagnosis.

Given that the counterfactual has well developed ethical underpinnings any proposals for regulation based on ethics for AI should be evaluated against this counterfactual. Further, the implications for innovation, productivity growth and the concentration of power versus markets with decentralised consumer choice and policy developed with comparatively direct political accountability should be assessed.

Given the uncertainty involved in assessing costs and benefits, upside opportunity needs to be weighed against downside risk. The White House 2020 tend to favour an innovation bias, whereas Europe tends to favour a regulatory approach which is pre-cautionary with regard to innovation.

Whilst the foregoing is a caution in relation to regulation of AI, and regulation based on ethics for AI in particular, the potential of AI to offer economic and social benefits, rather than dystopian fears, may be a reason to focus on AI. An agenda focussed on opportunity could be organised as follows.

Remove unnecessary barriers to AI

Existing law and regulation tend to reflect existing technology and market structures and may unintentionally impede innovation. Given the promise of AI a systematic and focussed search for, and removal of unnecessary barriers, may be justified. An example might be aspects of ‘data privacy’ regulation such as the right to explanation which could impede valuable applications of AI (Chivot and Castro 2019).

Do not apply a higher standard to AI alone

In a number of areas machine algorithms offer the prospect of not only lower costs than human algorithms, institutions and code but also higher quality outcomes, for example in relation to road safety. Yet the tendency, if we focus on AI *per se,* is to propose new rules and standards for machines alone, standards we do not expect of humans. We should not let perfection become the enemy of the good.

Further, were the performance of AI to justify a higher mandated standard on cost benefit grounds, adopting a common higher standard may rule humans out in relation to performing particular tasks, for example, driving motor vehicles on public roads. This highlights the possibility that some of key decisions regarding regulation motivated by AI are likely to be about people and what they are allowed to do rather than machines, and are therefore likely to be intensely political, rather than rights-based or ethical and technocratic.

Adapt existing law and regulation to AI rather than vice versa

Incumbent businesses recognise the ways in which existing rules can limit innovation, market entry and competition; and are therefore likely to lobby for extension of existing rules – whether fit for purpose in the new environment or not to impede competition. This has played out, for example, in relation to the taxi industry and messaging applications in the face of peer-to-peer services and internet-based messaging respectively.

To counter this tendency a clear focus on outcomes is required, reform of law and regulation to accommodate new ways of doing things. This may require different rules to achieve the similar outcomes, when account is taken of different technology and market structures (a level playing field in terms of outcomes rather than in terms of the rules *per se*). Reform may have to be driven at the political/policy level rather than relying on regulators - who may also be invested in the *status quo* and, in any case, lack the power to legislate to fundamentally adapt the rules.

There are two broad categories where adaptation may be required. First, the way in which existing rules are verified and enforced may need to differ between machine and human algorithms, for example, to achieve a desired reduction in discrimination. Second, AI itself – given its power to process information and support bottom-up market governance – may alter the need for centrally provided regulation.

The latter is an extension of the argument that online platforms have in part grown by offering superior governance to that offered by regulation (Cohen and Sundararajan 2015 and Williamson and Bunting 2017). AI is utilised to enhance market governance, for example, image recognition may be used to verify the identity of a peer-to-peer transportation driver (Lyft, 2019) and to interpret smartphone sensor data to detect accidents or other anomalies (The Verge, 2019). AI changes what the market can achieve in terms of information processing, can help reduce information asymmetries and may therefore reduce the optimal extent of centrally provided regulation in some contexts.

Recognise that AI may highlight problems that, whilst not novel, become material

By raising productivity and lowering costs AI may open up opportunities at scale that justify attention, even though there is nothing novel about the application. Examples include the potential for an increase in congestion if autonomous vehicles lower the costs and improve the convenience of personal mobility; and the potential loss of privacy and liberty if facial recognition by machines is applied at a scale not previously practical using human-based facial recognition (there are also alternative means of identification at a distance which might operate at scale including gait and heartbeat, The Economist 2020).

We therefore need to be clear about problem definition, and where new remedies are proposed consider their application to humans as well as machines. For example, if traffic congestion increases to the point where new interventions are justified such interventions should apply to all road users. In relation to facial recognition greater clarity over our unease is required prior to deciding what if anything should be done, and to the extent there is a trade-off between security and privacy and/or our sense of liberty (which relates to security) the question is political and unlikely to be resolved via the technocratic application of ethical principles. It should also be clear that there is no realistic prospect of global agreement over such trade-offs.

Conclusion

First, to reap the benefits of AI, we should focus on identifying and removing unnecessary legislative and regulatory barriers to AI. Second, we should adapt existing law and regulation as required in relation to AI mindful of the objectives of legislation rather than simply extending existing rules. Third, we should be cautious about holding AI to a higher standard than competing human algorithms, lest perfection becomes the enemy of the good.

In assessing proposals for regulation of AI we should be clear and focussed about problem definition, assess any proposal relative to a market and general regulatory counterfactual, which has ethical underpinnings, and have regard to the benefits of decentralised prediction, innovation and error-correction for a general-purpose technology with uncertain but promising application.

Change may be required, but regulation based on ethics for arguably AI *per se* represents a policy false dawn.

References

AINOW, AINOW 2019 Report, December 2019.
https://ainowinstitute.org/AI_Now_2019_Report.pdf p

Arrow and Debreu, Existence of equilibrium for a competitive economy, *Econometrica*, Volume 22, 1954.

Chivot and Castro, The EU needs to reform the GDPR to remain competitive in the algorithmic economy, Centre for Data Innovation, May 2019. *https://www.datainnovation.org/2019/05/the-eu-needs-to-reform-the-gdpr-to-remain-competitive-in-the-algorithmic-economy/*

Tessa Charlesworth and Mahzarin Banaji, Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016, *Psychological Science*, Volume 30 Issue 2, February 2019.
https://journals.sagepub.com/doi/abs/10.1177/0956797618813087

Cohen and Sundararajan, Self-Regulation and Innovation in the Peer-to-Peer Sharing Economy, University of Chicago Law Review Online, Volume 82(1), 2015.
https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1039&context=uclrev_online

Cooper and Kovacic, Behavioral economics: implications for regulatory behaviour, Journal of Regulatory Economics, Volume 41, No. 1, February 2012.
https://www.law.gmu.edu/assets/files/publications/working_papers/1313BehavioralEconomicsImplications.pdf

Council on Foreign Relations, Defending America From Foreign Election Interference, March 2019. *https://www.cfr.org/report/defending-america-foreign-election-interference*

CNET, In Amazon Go, no one thinks I'm stealing, 26 October 2018. *https://www.cnet.com/news/amazon-go-avoid-discrimination-shopping-commentary/*

DeepMind, AlphaFold: Using AI for scientific discovery, December 2018. *https://deepmind.com/blog/alphafold/*

Drexler, Engines of Creation: The Coming Era of Nanotechnology, 1986. Drexler.

European Commission communication, Artificial intelligence for Europe, April 2018. *https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe*

European Commission, High-level Expert Group on Artificial Intelligence – Ethics guidelines for trustworthy AI, April 2019. *https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai*

European Political Strategy Centre, The Age of Artificial Intelligence - Towards a European Strategy for Human-Centric Machine, March 2018. *https://ec.europa.eu/epsc/sites/epsc/files/epsc_strategicnote_ai.pdf*

Facebook, Expanding Our Efforts to Protect Elections in 2019, January 2019. *https://newsroom.fb.com/news/2019/01/elections-2019/*

Facebook, First Grants Announced for Independent Research on Social Media's Impact on Democracy Using Facebook Data, April 2019. *https://newsroom.fb.com/news/2019/04/election-research-grants/*

Farronato, Fradkin, Larsen, Brynjolfsson, Consumer Protection in an Online World: An Analysis of Occupational Licensing, NBER Working Paper 26601, January 2020.

FDA, Breakthrough Device Programme – Final Guidance, December 2018. *https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/ucm441467.htm*

Floridi *et al*, An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, November 2018. *https://www.researchgate.net/publication/328699738_An_Ethical_Framework_for_a_Good_AI_Society_Opportunities_Risks_Principles_and_Recommendations*

Forbes, Google's DeepMind has an idea for stopping biased AI, March 2018. *https://www.forbes.com/sites/parmyolson/2018/03/13/google-deepmind-ai-machine-learning-bias/#7efb56b46829*

Yuval Harrari, Homo Deus, Harvill Seeker, 2015.

Hinton, Google's AI Guru Wants Computers to Think More Like Brains, Wired, December 2018. *https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/amp*

Elizabexth Holm, In defense of the black box, Science, Vol 364, Issue 6435, April 2019. *https://science.sciencemag.org/content/364/6435/26*

Paul Hofheinz, The Ethics of Artificial Intelligence: How AI Can End Discrimination and Make the World a Smarter, Better Place, May 2018. *https://www.lisboncouncil.net/publication/publication/148-the-ethics-of-artificial-intelligence-how-ai-can-end-discrimination-and-make-the-world-a-smarter-better-place.html*

Daniel Kahneman, Thinking fast and thinking slow, 2011.

Kelly, The myth of a superhuman AI, Wired, April 2017. *https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/*

Paul Krugman, The Age of Diminishing Expectations, 1994.

Jon Kleinberg and Sendhil Mullainathan, Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability, August, 2018. *https://arxiv.org/pdf/1809.04578.pdf*

Hosuk Lee-Makiyama, Briefing note: AI & Trade Policy, 2018. *https://euagenda.eu/upload/publications/untitled-189277-ea.pdf*

Kleinberg, Ludwig, Mullainathan and Sunstein, Discrimination in the age of algorithms, February 2019. *https://www.nber.org/papers/w25548*

House of Lords, AI in the UK: ready, willing and able?, April 2018. *https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf*

Jovanovic and Rousseau, General purpose technologies, 2005. *https://www.nyu.edu/econ/user/jovanovi/JovRousseauGPT.pdf*

Lyft blog, Lyft’s Commitment to Safety, April 2019. *https://blog.lyft.com/posts/2019/4/14/lyfts-commitment-to-safety*

Maddison, *The World Economy – A Millennial Perspective*, OECD Development Centre Studies, 2000.

Rachel Botsman, Who can you trust?, Penguin, 2017.

Perkins, How does anesthesia work?, Scientific American, February 2005. *https://www.scientificamerican.com/article/how-does-anesthesia-work/*

Brent Mittelstadt, Principles alone cannot guarantee ethical AI, November 2019. Nature Machine Intelligence. *https://www.nature.com/articles/s42256-019-0114-4*

Onora O'Neill, TED Talk, June 2013. *https://www.ted.com/talks/onora_o_neill_what_we_don_t_understand_about_trust*

Pilat and Criscuolo, The future of productivity – what contribution can digital transformation make? Policy Quarterly, Volume 14(3), August 2018. *https://www.victoria.ac.nz/__data/assets/pdf_file/0009/1686141/Pilat_Criscuolo.pdf*

Richards, Thorlby, Fisher and Turton, UK cancer survival rates are poor, not because of poor treatment but because of delays in identification and referral for treatment. Unfinished business - An assessment of the national approach to improving cancer services in England 1995–2015,

November 2018.
https://www.health.org.uk/sites/default/files/upload/publications/2018/Unfinished-business-an-assessment-of-the-national-approach-to-improving-cancer-services-in-england-1995-2015.pdf

David J. Robertson, Eilidh Noyes, Andrew J. Dowsett, Rob Jenkins, A. Mike Burton, Face Recognition by Metropolitan Police Super-Recognisers, February 26, 2016.
https://doi.org/10.1371/journal.pone.0150036

Suleyman, A major milestone for the treatment of eye disease, 2018.
https://deepmind.com/blog/article/moorfields-major-milestone

The Economist, The Economist at 175 - Reinventing liberalism for the 21st century, 13 September 2018. https://www.economist.com/essay/2018/09/13/the-economist-at-175

The Economist, People can now be identified at a distance by their heartbeat, January 2020.
https://www.economist.com/science-and-technology/2020/01/23/people-can-now-be-identified-at-a-distance-by-their-heartbeat

The Verge, Uber is now using your smartphone (and your driver's) to detect vehicle crashes, September 2019. https://www.theverge.com/2019/9/17/20868466/uber-ridecheck-detect-vehicle-crashes-smartphone-driver-gps

Tom Standage, Regulating artificial intelligence, December 2018.
https://worldin2019.economist.com

Jean Tirole, Economics for the Common Good, Princeton, 2017.

Cédric Villani, AI for humanity, March 2018. https://www.aiforhumanity.fr/en/

Ursula Von der Leyen, A Union that strives for more - My agenda for Europe - Political Guidelines for the Next European Commission 2019-2024, 2019. https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf

Williamson and Bunting, Reconciling private market governance and law: A policy primer for digital platforms, June 2018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3188937

Wired, Machine learning invades the real world on internet balloons, February 2017.
https://www.wired.com/2017/02/machine-learning-drifting-real-world-internet-balloons/

Whittlestone et al, Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research, 2019. https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf