Global Digital Foundation
The Digital Policy Network

Making Rules for AI

The errors and fallacies regulators must avoid

Revised (October 2020) edition

Paul MacDonnell, Executive Director

Global Digital Foundation

25th October, 2020

Contents

Abstract

Policy analysts and policymakers are responding to AI as a threat to human rights and as a potential savour of humanity from discrimination. These responses originate in two places: the first is the widespread belief that inequality reflects malevolent intentions embodied in historical social hierarchies; the second is the belief that, unchecked, AI will instrumentalise these intentions. These responses explain why many people advocate that 'fairness' must be programmed into AI. This paper takes no issue with the view that AI should be designed so as not to treat people unfairly. It argues that while safety—e.g. in transport or healthcare—can, and should, be designed into AI prior to implementation, fairness defined as the achievement of equality of social outcomes, cannot be designed into AI in the same way. The likely causes of bad social or economic outcomes following the use of AI will be both multivariate and exogenous to AI technologies in a way that threats to safety will not; and, in fact, these 'outcomes' will usually predate the use of such technologies. For this reason, arguments to design or mandate for AI 'fairness' *ex ante* to achieve equality of outcomes rely on poor reasoning and, if successful, will result in threats to human rights arising from certified 'fair' AI passing unnoticed.

1. Introduction

Many people believe that the 'intelligence' in Artificial Intelligence (AI) is like human intelligence. It is not. In its current form, and for the foreseeable future, AI is a mental prosthetic that does things humans find difficult or time consuming. An AI that recognises pre-cancerous symptoms in retina scans more quickly and reliably than medical specialists has been trained, using other scans, to spot anomalies. It's more sophisticated than a pocket calculator but not more intelligent.

The intuition that AI has a mind finds its ultimate expression in the Singularity. A metaphor borrowed from the science of black holes, the Singularity is a hypothetical point in time when AI becomes powerful enough to leave humans behind or, even, take over the world. It captures well the idea of exponential growth in AI's capability (Silver and Hassabis, 2017). However this is misleading. AI's powers are inductive. It answers questions by finding correlations within masses of data. But the questions themselves must be supplied by

humans. Human intelligence, on the other hand, has a non-inductive side which AI lacks and without which we could not ask—and AI could not answer—any question. Our ability to ask questions both depends upon and supports our capacity to hypothesise, to form theories of reality, in other words, to *explain*. Karl Popper (1963: 105) describes science as embodying both conjecture and refutation—I take the abilities to conjecture and refute to better represent the workings of true human-like intelligence than the standard problem-solving models used within the computer science community (McCarthy, n.d.). All would-be rulers of the world need both of these faculties. Without them AI will only ever rule within a domain of tools and instruments.

The intuition that AI has human-like intelligence and a mind manifests an ancient habit of treating natural and social phenomena as if they had intentions (Searle, 1980; Hayek, 1973: 26-29; Dennett, 2013: 159-160, 419). It is an example of how we don't just use metaphors as descriptors. We use them to think about, and to explain 'reality' to ourselves and others (Lakoff and Johnson, 2003: 3). That is to say, our belief that AI has real intelligence is, at bottom, metaphorical thinking. Often our metaphorical explanations ascribe intentions to complex phenomena. This tendency is one of the most important elements of our culture. It is the basis for animism, for classical texts such as Ovid's *Metamorphoses*, and for getting people to agree with us in matters of public policy. It is why, for example, governments seek to motivate us to support measures to counter COVID-19 by declaring the disease to be 'the enemy' (CNA, 2020).

Rousseau and Voltaire rejected ancient social and ethical hierarchies as oppressive by challenging the intentionalist idea that God had ordained them (Dennett, 1996: 25-26; Hayek, 1973: 10, 25-27). Social theory—from Marx, to Gramsci's cultural hegemony, to post-Marxist 'critical' theories such as Critical Theory and Critical Race Theory—inherited the Enlightenment impulse to challenge traditional social structures in this way (Gramsci, 1971: 416-418; Matsuda 1993: 1). Along the way they redefined them in terms of dominance in (successively) economics, culture, gender, and race; but they left intact the intuition that 'intention'—man's, rather than God's—was their final cause. As with the Creationists' defence against Darwin, modern social theory partly understands these re-defined structures in terms of the 'intentions' that gave rise to them and which they, in turn, are supposed to serve. The contribution of this approach to the social sciences is its use of scepticism as a tool to dismantle complacent assumptions about the permanence, usefulness, and moral standing of established and powerful socio-economic and cultural systems of authority and

thought. But though such methodological scepticism has rendered services to social science and to policy, modern social theory's insistence on identifying 'intention' as a univariate cause behind social phenomena such as poverty or inequality leads us astray.

The EU's proposals to regulate AI

This paper was originally conceived as a response to EU proposals for the regulation of AI (European Commission, 2020). In the course of preparation for this relatively narrow analysis it has become clear that a wider examination of the background to the anxieties, understandings, and arguments surrounding AI is necessary in order to supply perspectives that are missing from the current debate in Europe and without which sound public policy is unlikely to emerge. By way of a down-payment on the paper's original promise the following is a summary analysis of the EU's response to the challenge of AI. A more detailed discussion can be found in 7. below. For those with particular interest in the EU position the rest of this paper should serve as an analysis of the perspectives and, sometimes mistaken, assumptions that have led to some of the EU's policy ideas and which, if these ideas became law, will need to be considered by a wider audience.

The EU's response combines policy ideas developed against the philosophical background briefly sketched out above with proposals for the development of a government-led AI industrial complex aimed at making Europe a centre of AI excellence where 'excellence' is defined as meeting the highest standards of safety and human rights as well as serving the continent's economic needs. Specifically, the EU proposes: 1) a single regulatory framework of 'excellence and trust' for 'high-risk' AI that would protect both the safety and the human rights of Europeans; 2) to invite into the same 'high risk' framework all developers and users of low-risk AI; and 3) to require that some imported (into the EU) AI be retrained using European datasets. Furthermore, the EU is considering the extension of strict liability for injury or breaches of human rights resulting from AI to all stages of its development and use. Overall, the proposed regime would include detailed and ongoing scrutiny of the key components of AI systems, in particular, datasets and algorithms that would require regulatory approval prior, and ongoing review after, deployment. The white paper's most notable feature is the proposed *de facto* treatment of human rights as indistinguishable from safety.

The EU proposal treats human rights as if they were the same as safety

The relationship of cause to effect in bad AI that leads to injury or death has, outside of simple errors, no counterpart in the relationship between AI and human rights. In its rules and in the wider discourse on safety and liability the EU tends to locate risk within the technology itself (Council of the European Union, 2001, 2006; and European Commission, 2020b). A 'cause' of a breach of human rights, on the other hand, will only be found there in the simplest of cases—usually as a result of an error. Accordingly, and in contrast to its recommendations to regulate AI for ‘fairness’, the European Commission (2020) locates the source of unfairness in AI outside of the technology—within society—and cites, as supporting evidence, a study that applies the standard of group fairness in a discussion of AI’s bias (Tolan et. al., 2019).

Though safety risk is endogenous and human rights risk is exogenous to AI technology, the EU proposes to treat them both as equivalent for the purposes of regulation. For safety the European Commission (2020) proposes that AI training data should cover ‘all relevant scenarios needed to avoid dangerous situations’. For human rights it says that data sets should be ‘sufficiently representative, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected…’. This latter is just as precautionary a regime as those for AIs affecting, for example, the safety of driverless cars. But the relationship between ‘representative’ datasets to the 'fairness' of decisions made by an AI is by no means as direct, certain, or even, for that matter, discoverable as the relationship between a driverless AI’s data and training to safety.

This takes us into the white papers’ notion of causality in AI. The risk presented by certain AIs to safety, for example in the field of transport or healthcare, lies with the technology itself, viz. an engineering failure—of design or function—could cause injury or death. Here an AI could be both the *instrument* and the *cause* of an accident. When I say 'cause' in this context I mean that one can say, without risk of being misunderstood, that the cause of an accident, for example a (driverless) bus crash, lies within, or concerns directly, the technology itself. Of course investigators may identify several ‘causes’ including management failures within the firm that designed the AI. But it is clear that to grasp that the technology has malfunctioned is to grasp the problem or its salient manifestation.

Other than unintended, easily identifiable, and easily repairable errors, the risk to human rights associated with AI, on the other hand, exists outside of the technology itself and remains in the hands of its users whose actions are already subject to EU or US human rights and privacy laws. An AI that has been used to abuse human rights is an *instrument* not a *cause* of this abuse. In seeking to apply rules directly to it in order to protect human rights the EU is falling into an anthropomorphic or intentionalist view of AI. Furthermore, this view inevitably encourages proposals that governments *instrumentalise* AI to protect human rights. This is already visible in the thinking of those policy analysts who go further than the EU to assert that society is structurally unfair and that AI can be programmed for 'fairness' to identify and eliminate bias (Hofheinz, 2019; Clegg, 2019; Montgomery, 2019). The EU's position effectively invites this opinion to stake its claim to AI and to advance proposals to instrumentalise AI along these lines.

The EU proposes a vast and mostly-unnecessary system of oversight

The consequence is that the EU proposes a regulatory system of pre-approval certification and post-deployment review and revision that treats all 'high-risk' AI in the manner of critical aviation technology or the necessity to establish the safety of new drugs. Such a regime is not valid for AI that could threaten human rights.

The EU proposes that all AI be invited to join the high-risk regime

The EU proposes that developers and users of AI that is not high-risk—i.e. *all* other AI—be invited to join the high-risk regime. This proposal is motivated by the desire to establish the EU as a centre of excellence in AI. However it simply ensures that an unnecessary level of oversight will apply to all AI.

The EU proposes imported AI be retrained for use in the EU

The EU proposes that AI imported into Europe be retrained using European data that conform to the Union's principles of social diversity. Because the decision-making of most AI will not be dependent on such data sets, for example data sets that contain information about credit rating or educational attainment only, this measure could present a barrier to entry into Europe for non European AI and deprive European organisations of access to perfectly functional, fair, technologies.

Conclusion

On a broader level, behind the EU's approach are anxieties (listed in the European Commission's white paper), that AI threatens access to jobs, privacy, and freedom of expression (i.e. democracy). The Council of Europe (2017) along with philosophers and commentators (Vinge, 1993: 11-12; Curiel, 2019; Kurzweil 2005) have also indicated these and other threats from AI. But they are almost entirely misconstruals—identical to those that have led some scientists to predict the technological Singularity itself; and they explain why the EU wishes to address AI's human rights, safety, and economic implications within a *single* framework, irrespective of the inherent heterogeneity of the factors that give rise to them.

The white paper is the response. It is a call to construct, like President John F. Kennedy's (1962) proposal that the U.S. should go to the Moon, a shared scientific endeavour that will embody not just European technical prowess, but European values. But if the future importance of AI is as great as Europe's policymakers believe then their proposals to protect human rights in an AI context would represent the most ambitious and far-reaching state intervention in the economic and social life of Europeans since World War II. Meanwhile, influential commentators feel free to propose measures that go even further, including measures to ensure that AI can, itself, be harnessed to create a more just society. But the proposal to regulate and instrumentalise AI in the cause of human rights risks the very outcome that these measures seek to avoid.

This paper is both a response to the EU's proposals for the regulation of AI and a more general engagement with ideas and assumptions that underpin the current European (European in the broadest political sense) understanding of the relationship between AI and socio-economic policy. It is structured as a series of brief discussions of the following six topics:

1. The response to AI considered as the latest in a history of anxieties triggered by rapid technological change.

2. How the failure to appreciate the differences between AI and human cognition distorts the thinking of policymakers.

3. How our answer to the question: *What do we mean when we say that AI is 'biased'?* determines whether we take a *minimalist* (AI should avoid harm) or a *maximalist* (AI can fight endemic bias in society) position on AI's proper relationship to human rights.

4. How the maximalist view of AI treats it as a causal instrument of social inequalities while ignoring the multivariate causes of such inequalities; and how this error makes it more likely that 'unbiased' AI that has been instrumentalised to fight 'bias' will, itself, be used to cause human rights abuses.

5. The role played by three epistemological errors, *essentialism*, *inductivism*, and *intentionalism* in attempts to construct relationships between AI and future policy.

6. The details of the EU's proposals for an AI regulatory framework and some of the practical implications of these for European AI.

2. Technology, anxiety, and the Singularity as a bad thinking tool

2.1 Anxiety about technology before AI

New technology has always sparked anxiety. Queen Elizabeth I refused a patent to William Lee for his stocking frame, a knitting machine she thought would deprive the kingdom's hand knitters of their livelihoods. The Luddites of the 19th century attacked the machinery they saw as a threat to skilled jobs. The introduction of the Kodak camera in 1888 was followed by widespread concern that snap-happy enthusiasts were invading the privacy of citizens by photographing them without permission (McQuinn, 2015). Since the invention of the printing press in the 15th century governments have worried about losing control of information every time a new and disruptive communications technology comes along (Emord, 1991: 27). In the 20th century European social scientists and philosophers—notably Theodore Adorno, Max Horkheimer, and Herbert Marcuse—worried that a capitalist culture-industry nexus was using technology to entertain and distract citizens while it reduced them to components of an automated industrial system (Marcuse, 1964: 163, 172 ; Adorno and Horkheimer, 1956 p.39; Antonio, 1983).

More recent anxieties focus on technology platforms, which are said to threaten jobs, privacy, and political sovereignty by out-competing bricks and mortar rivals, by selling personal data to third parties, and by targeting disinformation and extreme political content at social media users. There is no escaping it. Technology is disruptive and the anxiety it causes should not be a surprise. It bestows greater competitiveness on some thereby disadvantaging others. That, overall, its benefits may outweigh its costs is small comfort to the worker made redundant from a legacy industry, to the victim of identity theft, or to the target of online incivility. Technology brings change and it is natural that we respond to change with anxiety.

2.2 Anxiety about AI in particular

Now these anxieties are being intensely focused on AI. Technology writers have, since the early 1990s, been discussing a ‘technological Singularity'—a metaphor for a runaway AI which will render human beings superfluous (Curiel, 2019; Vinge, 1993; Kurzweil 2005). The metaphor borrows from astrophysicists' explanation of black holes. A black hole results when gravity around a body is so strong that nothing, not even light, can escape it. Astrophysicists

believe that at the centre of every black hole is a Singularity, a point with no dimensions where the structures of spacetime and matter break down—along with the laws of mathematics and physics. The point of proximity, beyond which we are unable to observe what happens close to the black hole, is known as the event horizon. This is also the point beyond which no object can escape being sucked in by the gravitational collapse. These concepts are worth understanding because their metaphorical use in relation to AI explains both the anxiety surrounding this technology and the ways in which we can be misled when we use them them as tools for thinking about it.

The earliest use of the Singularity metaphor is said to have been by the mathematician and physicist John von Neumann who was reported to have commented on 'the ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue' (Ulam, 1958). The metaphor of the Singularity is useful and misleading in the following ways. It usefully conveys the sense that AI can, by using reinforcement learning, improve its own cognitive power exponentially. An example is Google Deep Mind's Alphago Zero, a game-playing AI which, from a standing-start of never having played the game, became one of the world's most powerful Go players in just 3 days (Silver and Hassabis, 2017). The metaphor of the Singularity also conveys the unknowability of how a very powerful AI might arrive at a good decision. In a match it played with one of the world's leading players, Lee Sedol, on 10th March, 2016, an earlier version of Deep Mind's Go programme, Alphago, made a move which top-ranked Go players assumed to be a mistake but which turned out not to be (Metz, 2016). Go is a long-term strategy game and the move's significance only became clear at a later stage in the play. The Singularity metaphor is misleading, however, in one very important respect. It assumes that AI's cognitive brute-force 'intelligence' is like human intelligence. But this is not the case (Dennett, 2015).

If we consider the EU's response to AI we can see that it stems, in general, from anxiety about a loss of European agency within a globalised economy whose infrastructure appears to belong increasingly to American corporate giants and, in particular, from an earlier and still lingering philosophical reaction to the impact of technology that gripped a number of influential European philosophers during the 1960s and 70s. Herbert Marcuse's *One Dimensional Man* is an important example (Marcuse, 1964: 163; Whitfield, 2014). Marcuse described a future where 'the liberating force of technology—the instrumentalization of

things—turns into a fetter of liberation; the instrumentalization of man'. He explains further (Italics are the author's):

> Only in the medium of technology, man and nature become fungible objects of organisation....technology has become the great vehicle of *reification*…The social position of the individual and his relation to others appear not only to be determined by objective qualities and laws, but these qualities and laws seem to lose their mysterious and uncontrollable character...the world tends to become the stuff of total administration…' (Marcuse, 1964: 172).

When it comes to AI Marcuse's sentiments must surely resonate with the current anxieties of European policymakers (e.g. EDPS Ethics Advisory Group, 2018: 16) who, for good measure, resent the fact that, as in previous waves of IT development, it is U.S. technology companies that are in the vanguard of AI deployment (Satariano and Pronczuk, 2020).

In fact, the Singularity metaphor tells us more about those who use it than the reality of AI itself. The technological Singularity seems less a valid homologous analogy to its astronomical counterpart than a reflection of hidden emotions and anxieties. Like a 'shadow' Universe which some physicists think explains the patterns caused, on an opposite surface, by 'shadow' particles passing through the second slit in the famous two slit experiment (Deutsch, 1998: 88), the technological Singularity reflects a shadow *emotional* reality—a Singularity of technological anxiety.

In this parallel reality AI is cast as a unique threat to human rights: to privacy, justice, and, even, democracy. Here, we look with dread on a future where personal economic autonomy, human dignity, and democracy break down as each individual becomes a node within a techno-capitalist administrative network. This Singularity is a point of convergence and intensification of anxieties that predate AI but which appear to be accelerating through history over an event-horizon—in this case the point in time beyond which governments' ability to exert control will be forever lost. (Council of Europe, 2017). The Technological Singularity seems to rely on qualities of self-similarity to the Astrophysicists' Singularity. David Deutsch (1998: 97) has commented on the fact that 'science and other forms of knowledge are made possible by a special self-similarity property of the physical world.'

It seems that we can view metaphors such as the technological Singularity as a form of knowledge in this light (Lakoff and Johnson, 2003: 147-155). But some forms of knowledge can mislead us.

3. What AI is, and what it is not

3.1 Introduction

There are two kinds of intelligence and AI has only one of them. Human intelligence is the ability to do two things. First, humans can solve problems, such as how to find the square root of 20, where to find the nearest Asian restaurant while walking down the street, and calculating the average distance between Earth and Mars during Earth's solar year. Second, humans develop, think about, and use explanations as frameworks within which they think and act. These include the Theory of Relativity, religions, and the philosophy of human rights. AI is good at the former but not the latter. It has no theories about what is real or important. It can prioritise a 'best' move on a Go board or which felons are most likely to reoffend. But this is merely an instrumental application of a system of rules to data. The rules and data may exist within a theoretical framework but this will not be the *AI's* theoretical framework. AI has no intentions.

AI is an extension of computational ability, no different in principle from other technologies humans have used over millennia, such as the abacus, the slide rule, the pocket calculator, and the smartphone. In its current forms AI explains nothing and, in fact, could never arrive at an explanation, or theory, no matter how much data or self-teaching it used. Part of the reason for this may be because explanations are never *true* in any computational sense. Explanations answer *why?* questions. The question '*Why did this happen?*' can attract any number of answers which in turn can attract further '*Yes, but why?*' questions and the answer to that question can attract the same, '*Yes, but why?*', answer and so on in infinite regress. Nothing has a *final* explanation, yet we need explanations in order to identify and prioritise problems we wish to solve. An AI that predicts a felon as likely to reoffend doesn't know if this means he or she should be given a longer jail term any more than the tin opener you are using to open a can of beans knows how hungry you are.

Daniel Dennett's (1996: 24) suggestion that the best way to derail a series of *why* questions in infinite regress is with a *how* question, though it is a useful way to stop fruitless speculation, potentially misleads us as to the value of *why?* questions that genuinely seek an explanation. AIs are potentially very good at helping us to answer *how?* questions.

But without our own, albeit always provisional and imperfect, answers to *why?* questions we could never decide to do anything and would not know what questions to ask an AI or how to use it within any reliable moral framework.

Much of the dystopian account of the technological Singularity and some proposals about how to regulate AI do not reflect this distinction between explaining and problem-solving. Anyone intending to profit from AI regardless of the impact on human rights, or, even using AI to justify actions to achieve 'fair' outcomes would likely also disregard it. We could expect them to assert that an AI morally justifies a particular outcome after they had imputed explanatory powers to it that it does not have and, or, defined it as an instrument for social good. Policy analysts', regulators', and even philosophers' failure to make this distinction creates the danger that regulators will not only overreact but that they will attempt to impose an Augustinian conversion on AI in an attempt to make it into an instrument for 'good'. This is the central problem with which this paper is concerned.

3.2 Induction machines

AI's ability to see relationships across masses of data promises improvements in intellectual work of all kinds. For example, an AI that sees connections across multiple health data-points will improve diagnoses, the timing of interventions, and the targeting of treatment. But AI's powers go beyond outpacing traditional expertise such as reading medical scans faster and more accurately than hospital consultants. Its ability to identify relationships across multivariate and disparate datasets could grant it a much broader role. This could include, for example, informing or even enabling the development of policies aimed at promoting social, educational, and economic development among the disadvantaged. Hence, as well as healthcare, AI could transform public policy in many other areas, including education, housing, and criminal justice (Stevenson, 2018). But here we can see that its role would be the interrogation of data to help policymakers create better explanations for social problems as a necessary prelude to creating better public policy. This is very different from using AI technologies as machines whose sole role is to confirm pre-existing ideologically-motivated explanations for, or responses to, social deprivation.

This goes to the heart of how AI uses inductive rather than deductive methods. The problems that AI solves and the new data that it generates are the products of inductive learning. In other words an AI searches for examples of things, such as correlations between two or more data points or patterns across datasets and, depending on the instructions

encoded into its algorithms, suggests or takes certain actions based on what it finds (Sweeney, 2018). A computer, or a person, that uses such inductive methods will have a worldview that derives from the data they have been trained to see as relevant (*Mind Matters News*, 2019). However, an AI programmed with Newton's Law of Universal Gravitation, while it could predict the movement of planets around any star system, would never be able to suggest that Newton's is the best explanation for the motion of bodies in the Universe. To do this it would need to consider alternatives. But it wouldn't have all the necessary data, much of the data it did have would have the wrong values, and if it had not been asked in the first place it would not even begin to consider other explanations. It would certainly never arrive at Einstein's explanation of gravity as the curvature of spacetime around large objects. All of this is because AI in its current machine-learning form uses inductive reasoning.

Because of its ability to interrogate data AI will be invaluable to scientists with explanations that need testing. Its current role in the sciences is as a super laboratory assistant. It can increase the speed of testing and experimentation with data and help generate new insights that may debunk old theories and prejudicial habits of mind (Royal Society, 2019). Thus it could help scientists in their quest to develop new explanations for the mysteries of quantum physics. However, though it could play a similar role in addressing social problems, as with the 'hard' sciences, AI could only do this if its data were handled according to theoretical frameworks that themselves had been tested to the point where they qualified as valid foundations for research.

Social science is distinctly lacking in this regard. When we consider the question of identifying and eliminating bias in AI we find that its inductive nature is the very thing that gets in the way. Leaving aside straightforward errors or intentionally prejudicial design, a finding of bias within an AI will depend on standards that have been applied by its users and, possibly, retroactively by those who are seeking to confirm 'bias'. These standards will invariably constitute theories or problems-to-be-solved to which only an inductive evidential framework will be applied. In other words, only data that seems to confirm the original theory will be deemed relevant. The overwhelming number of examples of 'bias' will be of indirect, unintentional, and ultimately, *interpreted*, effects. If, for example, differential social outcomes are to be used as evidence of bias then any interaction between members of disadvantaged groups and an AI could be identified as 'biased' if the AI 'affirms' their socio-economic status. Such *post hoc* reasoning already affects much of the discourse around AI and human rights.

3.3 AI's limitations in applied public policy

The implications for using AI to apply public policy are, therefore, stark. AI systems that behave in a way that confirms, or aligns with, social inequalities, such as offering higher interest loans to the less creditworthy or recommending that repeat-offending convicted felons spend longer in jail, are unlikely to be biased—in their *design*—against ethnicities, religions, or genders. Discoverable bias, if there is any, will originate with users' *intentions* in using such technology. For example, proving that black and white residents in a city with the same credit scores are, nonetheless, offered different interest rates on loans, would not—other things being equal—require examination of an AI system used to make these offers. The bias would be manifest in the outcome of lending decisions across the two groups.

To understand the challenges of AI's relationship to public policy in more detail we can consider COMPAS, an analytics tool used in the United States to predict the likelihood that a felon will reoffend. COMPAS can identify patterns in human behaviour, family background, and living conditions that allow it to predict criminal behaviour with at least as much accuracy as the human predictions of law-enforcement authorities.[1] But COMPAS' insights are purely the product of inductive processes. That is, they use data that inform but do not confirm a pre-existing policy/hypothesis—a hypothesis (viz. *locking up felons who we think will reoffend is a good idea*) that is not encoded into the system's design. COMPAS could support a range of policies besides that for which it is primarily used, reaching sentencing decisions for convicted felons (MacMillen, 2019). It could just as easily support interventions by social workers, economists, health professionals, and potential employers to reduce the risks of reoffending and increase the chances of rehabilitation without resort to longer prison sentences (Bertelsmann Stiftung, 2020). To be clear, COMPAS informs, but does not validate, the chosen policy of keeping felons who are most likely to reoffend off the streets. Therefore, it does not support any assertion that this policy is the way to keep the public safe. It doesn't even support any assertion that such a policy is not harmful and immoral. The same will apply to any assertions about an AI system that reflects inequalities in social outcomes between people categorised according to race or gender.

[1] Though, as we shall see in 5.3 and 5.4 below, COMPAS is designed to overestimate the risk of re-offending to serve a policy of minimising the risk of crime.

3.4 A brief discussion about AI and intentions: Searle vs. Dennett

The impulse to think about AI as if it has intentions may explain the origin of our fear about its dangers. It is a paradox, therefore, that the argument to make AI 'good' relies less on this anxiety than on the idea drawn from modern social theory that it is impersonal *collective* intentions within historical social structures, and not *individual* intentions, that direct our lives. Furthermore, the seemingly unrelated neo-Darwinian view that because evolved human intelligence doesn't need intentions then an AI with human-like intelligence won't need them either is a fortuitous corollary to this view. This has been advanced by one of the most influential of contemporary philosophers, Daniel Dennett, and it may help us to understand why the gateway to the *ex ante* regulation of AI to achieve equality of social outcomes is standing wide open.

In 1980 the philosopher John Searle (1980: 3, 417-457) published a rebuttal of 'strong' or truly intelligent AI (i.e. AI whose intelligence is indistinguishable from human intelligence) which used the following thought experiment. Searle, who knows no Chinese, is locked in a room and given: 1) a batch of Chinese writing; 2) a second batch of Chinese writing; 3) rules in English (a language he knows) for correlating the first and second batches, and; 4) a batch of Chinese symbols. The people outside the room call the first batch a 'script', the second batch a 'story' and the third 'questions'. Searle shows that, by following the rules to return correct answers in Chinese to the questions in Chinese about the story in Chinese, the system inside the room could fool a person outside the room into thinking that it understood Chinese. Hence the experiment would pass the Turing Test (Turing, 1950). That is to say, it could fool a Chinese speaker reading the replies emerging from the room into the belief that they were dealing with a human intelligence. Searle holds the fact that neither he nor any other component of the experiment embodies an understanding of Chinese as proof that strong AI is impossible and that passing the Turing Test is, therefore, not a demonstration of any kind of intelligence.

Responding to Searle, Daniel Dennett (2013: 324-326) argues that the 'system' (i.e. the room + Searle + three batches of paper with Chinese characters on them + instructions) is what is doing the 'understanding'. Furthermore, he points out that Searle ignores Alan Turing's observation that the 'competence' of an intelligent AI is in the software and that, for this purpose, the 'system' *is* the software. Dennett defines 'competence' here as equivalent to intelligence and further argues that Searle has ignored Darwin's 'strange inversion of

reasoning' which has overcome our traditional idea that competence can only come from comprehension. In other words Searle's reliance on comprehension (which is what his intentionalism amounts to) is a relic of the Cartesian superstition that intention must be behind phenomena in the biological and natural worlds.

This criticism of Searle is persuasive, but only in so far as it targets what seems to be a weakness in his argumentation. Searle appears to rely on the fact that human and non-human sentient creatures are 'made of similar stuff' that uniquely constitutes the basis for intelligence, the essential ingredient of which is intention. But Searle's Cartesian error is not fatal to a much more important point which, it must be admitted, he manages to conceal, as the following shows.

Searle regards the absence of intentions in AI as manifest and contrasts this with its manifest presence in humans and, indeed, in other mammals. What Dennett means by 'intention' on the other hand is what he sees as an axiomatic truth, following Darwin, that neither God nor any other intending agent can account for the natural world, and that includes intelligence of *any* kind. Two circles in a Venn diagram representing, respectively, Searle's view that AI cannot have intentions and Dennett's view of AI as another phenomenon within an intentionless evolutionary Universe would overlap. But Searle and Dennett mean different things by their respective overlapping assertions about the 'intentions' of AI. Searle's view is that intentions are required for real intelligence while Dennet has no such view. Dennett just doesn't like Searle's conclusion that AI is, therefore, stupid. This is because Dennett regards intentions as superfluous to any discussion about biological or natural phenomena, including humans and, therefore, as immaterial to explaining the evolution of biological complexity, including that of the brain. Dennett asserts that intelligence doesn't need intentions because in evolutionary terms it is nothing more than the ability to solve complex problems. But Searle isn't talking about how the human brain evolved. He's talking about how it works.

If for 'agency' or 'intention' we substitute 'ability to hypothesise' or 'ability to explain' Searle's argument becomes much stronger. Dennett may be correct in his assertion that Searle's thought-experiment fails in its objective of proving that strong-AI is impossible. But by equating it with mere problem-solving he casts intelligence in purely instrumental terms. This overlooks that aspect of our intelligence—notably scientific or, even, religious thought— that seeks explanations about the world. And indispensable to such speculation are human

intentions or agency that manifest in the form of hypotheses or conjectures. Popper (1963: 100-103) explores this divergence between instrumentalist ideas about what constitutes human knowledge and the idea that human knowledge proceeds from hypotheses or proposed explanations ('conjecture'), via refutation or attempts at refutation, to explanations which, in turn, are subject to further experimental challenge and so on. It seems as if Dennett has extrapolated his instrumentalist idea of intelligence from a Darwinian evolutionary framework and simply applied this to AI. The key here is that understanding how intelligence evolved is not the same thing as understanding how it works. A truly intelligent AI would be able to both hypothesise and solve puzzles. Dennett doesn't even raise this as a possibility.

Searle's argument is weakened because in pointing out that intention is a constituent element ('stuff') of intelligence he fails to explain why it is so. This is why turning to Popper helps us to grasp the problem. Dennet's concept of intelligence obliterates any distinction between explaining and problem solving. He sees intelligence as just about problem solving.

By now it should be clear that Dennett and Searle are at cross purposes. Popper's (1963: 97-118) account of what constitutes true science—as opposed to 'pseudo science' such as astrology or the Baconian, inductive, collecting of facts—is that it comprises both conjectures (hypotheses or explanations) and refutations (tests aimed at disproving these conjectures). Without agency there cannot be conjectures and without conjectures there can be no questions. In the case of the Chinese Room experiment the conjecture belongs to the questioner who seeks confirmation or disconfirmation of an explanation for an account of reality that must exist outside of the frame of reference of the questionee—that is, outside the room so that the actual system, i.e. the room and its contents, can either pass or fail the Turing Test.[2] But once you remove the questions, what we have left looks much more like Searle's (1980: 419) explanation that the Chinese room produces correct answers through the application of intentionless, syntactic, formal rules.

Dennett is right that intentionality is not necessary to explain anything in biology, including the development of human intelligence. But this is not the same thing as saying that true intelligence, human or otherwise, can do without intentionality or, at least, that aspect of it that seeks to solve particular problems. Dennett's instrumental definition of 'intelligence' as problem solving would define AI as (potentially) having human-like intelligence based solely

[2] Searle had earlier also rebutted the same 'system' explanation in Turing, 1950.

on its competence—the more complex the better. But we need to retain intention, in the sense of *agency*, as a concept in human intelligence not out of a Cartesian superstition about locating the *essential causes* that will explain the vicissitudes of human affairs but in order to: 1) explain how we can hypothesise and (current) AI cannot and; 2) assign agency to individuals in order to assert and protect their human rights. Therefore, it is reasonable to draw a distinction between AI and human intelligence on the basis that AI lacks the ability to hypothesise and on this point (assuming that this point is a corollary to his argument) Searle's view is, for now, closer to the truth and we should not be distracted by the easy Cartesian target he provides to Dennett (and, presumably, others) as he gets there.

The idea that agency is not a necessary component of either human or artificial intelligence, which is the thrust of Dennett's argument, is also a premise at the heart of the maximalist view of AI—that it should be given the ability to make decisions in favour of human rights based on data science. This maximalist approach seems to be inchoate within a number of proposals to instrumentalise AI to fight bias (Hofheinz, 2019; Chivot and Castro, 2020). Furthermore, Dennet's (2017: 57-58) competence-without-comprehension view of intelligence is consistent with—though it doesn't support—the widespread idea that human agency is an illusion that masks the instrumentalisation of people by deep structural and 'unconscious' biases within society.

To summarise—the argument made here, that AI cannot defeat bias and in favour of the view that our intentionality is what distinguishes humans from machines, must be distinguished from our need to reject intention as a reliable explanation for social as well as natural phenomena. Rejecting intention as a component of *individual* human intelligence has dangerous implications for human rights. Rejecting the intention of individuals is consistent with the idea that they are mere extensions of broader forces that operate at the level of social categories such as class, race, or gender. This is the view that has opened the door to the instrumentalisation of AI as a weapon to fight 'endemic bias' in society. But intention has a habit of creeping back into philosophical arguments that purport to have expelled it (Hayek 1973: 27-28). In this case Dennett may have successfully expelled intention from human actions at the level of the individual but this may simply have cleared the ground for, and resistance to, those who assert that 'intention' is a decisive factor operating at the level of social categories such as race or gender—a Cartesian error that could have profound consequences.

4. AI and human rights

4.1 Can AI really be ‘biased’?

It’s time to look at the question of bias and AI in more detail. Well-documented mistakes in AI alert us to the dangers of granting unsupervised powers to such far-reaching technologies (Zunger, 2019). It is often pointed out that the components of building an AI will typically reflect the world-view of its developers. This includes framing the problem an AI is being asked to solve, collecting and using data that reflect past practices and assumptions, and selecting attributes to use in generating outcomes (Hao, 2019). Consider the following three examples:

1. A lender’s AI defines ‘creditworthiness’ in terms of how likely an applicant is to repay a loan. It offers fewer loans to people on low incomes, a disproportionate number of whom are black.

2. Another lender’s AI defines ‘creditworthiness’ not in terms of how likely an applicant is to repay a loan, but which applicants will be most profitable. It offers a greater number of more-profitable, high-interest, subprime loans to people on low incomes, a disproportionate number of whom are black (Hao, 2019).

3. A firm uses data reflecting past hiring practices to reject female applicants to a male-dominated profession, such as computer engineering (Dastin, 2018).

The criterion upon which the AI in the first example reaches its conclusion—capacity to repay a loan—is undoubtedly the result of the lender’s prudential policy which requires that it not become insolvent through making bad loans. The fact that this policy appears to negatively impact black borrowers simply makes visible an underlying social situation. I say ‘appears’ here because it is not clear that refusing to lend someone money they can’t repay is unfair. The AI that makes this ‘decision’ is not racially biased. Even if a new government—having been elected on a campaign that the ‘financial system is racist’—mandates more loans to a greater number of black borrowers with low credit scores that still doesn’t make the original lending policy racist. There will be plenty of poor white residents in the same situation and the government’s intervention will not only not help them but could be said to discriminate against them.

In the second example the lender can be accused of bad lending practices which impact more heavily on black individuals because black individuals are, in this context, more heavily represented amongst those with low credit scores. It is possible that the lenders' executives behind this policy are racists who would not have used it if most of the loan recipients had been white. However, even in this case, the reason why their policy of subprime lending to black individuals is bad is not because the decision to do so relied on racial prejudice but because subprime lending is of itself a bad policy. The decision to implement the policy may have been biased but the bias was in the decision and not in the AI. This takes us back to the difference between regulating AI for safety and human rights. A faulty AI in use on a driverless bus could cause an accident. Investigators will, in all likelihood, identify the origins of the accident within the quality control processes of the AI's developer. The cause of the accident can be narrowed down to, and proven by, examining the technology itself. We can reasonably say that the faulty AI 'caused' the accident. Outside of obvious errors that could lurk in its data or algorithms, the 'bias' of an AI that impacts negatively on an ethnic group will, on the other hand, most likely reflect an existing social condition, or a wider policy framework. The bias will not be found in the AI because it will not exist there. The AI will simply be an instrument of policy.

In the case of the third example, the firm's hiring algorithms could reflect assumptions that ignore attributes more common among women that make for, or do not preclude, strong engineering skills. Recruitment and promotion practices tend to rely on the heuristics of 'hiring people like us'. This is genuine bias. Oftentimes it is what firms are doing when they talk about 'cultural fit'. There is evidence that, for the most part, it impacts on class and background (Friedman and Laurison, 2020: 17-27). Even here the bias is likely within the organisation's own culture and, therefore, there is no reason to single out an AI as the 'cause' of the problem. There is, though, every reason to challenge the firm's assumptions about whom it should be hiring. The firm could use AI as a tool to circumvent its own biased human recruitment heuristics. But we should be careful not to treat this as an indictment of human bias in contrast to an AI that is 'fair'. The decision to use a 'fair' AI here would require that the firm eliminate bias from its own decision-making in the first place. The 'unfairness' of a putative legacy recruitment AI could be described as a reflection, evidence, or even, an instrument of bias; but it should not be described as a cause of unfair recruitment practices. Accordingly, the 'fairness' of a firm's replacement 'fair' recruitment AI would be a reflection or evidence, or an instrument of its new practices.

4.2 Responses to bias and AI

Responses to the concern about bias within AI include both technical and governance initiatives. On the technical side, Google has developed the What-If Tool (webpage) to inspect machine-learning models that enables developers to alter the model's parameters and visualise the likely consequences for sectors of the population. At the same time companies, trade associations, legislators, and non-profit organisations have proposed a range of governance models aimed at preventing bias in AI.[3]

Such tools and initiatives are undoubtedly useful in avoiding errors or unintended outcomes in the use of AI. They are necessary assistants to our judgement of how organisations should conduct themselves. But they do not tell us what constitutes bias itself. For example, is setting the standard for applicants for a job at a higher level than the role needs—resulting in fewer women or members of ethnic minorities being hired—necessarily biased? Google's What-If Tool won't tell you. It may tell you of the unintended effect of such a decision. But the decision to set the standard higher than is 'necessary' is not *per se* biased. It may, in fact, be a good idea. It could be a way to ensure that new employees are easier to move to another role or easier to promote in anticipation of rapid growth in the company that, itself, will result in more women and minority hires. If the AI were deployed by racist and sexist managers who simply wanted to exclude minorities and women from the firm then we could say that the AI is 'biased'—but only in the figurative or metaphorical sense (Searle, 1980: 419) that racist and sexist managers conspired to use the AI as an instrument to exclude minorities and women from the firm.

Even if we had proof that this was their intention would this make the AI *itself* biased? If, instead of spending money on an AI, the executives had put the applicants' names into column A of a spreadsheet, added checkboxes under columns B, C, and D, respectively headed 'white', 'minority', and 'women', to categorise each candidate, and then exported a shortlist comprising the names of the white men only to a word processor's mail-merge utility generating a letter inviting them to an interview—would *this* make the spreadsheet biased? And if so, should we, in consequence, devise human-rights regulations that would apply to Microsoft Excel or Google Sheets? Or to ask the same question in a more extreme way: in the aftermath of crimes against humanity during World War II did we need special regulations for trains and barbed wire? Thus, enabled by metaphorical thinking, the

[3] See Appendix for a survey of these.

anthropomorphic intuition—that of identifying the bias as endogenous to the AI—leads us to seek regulations to address ‘biased AI’.

Therefore, considering the EU White Paper on AI, we have no reason to believe that AI can circumvent U.S. or European anti-discrimination legislation and therfore no reason to support additional human-rights rules for AI, over and above this legislation. The evidence of bias will be easily found by regulators who could spot anomalous hiring outcomes and, even, submit job-applications to test outcomes. This would be sufficient to impose sanctions and alter behaviour deemed illegal without recourse to micro-regulating the organisation’s AI back-end.

4.3 Two views of bias and AI

The anxiety about AI and ethics centres on the fear that AI has the capacity to embed bias within its datasets or algorithms. In the current debate the argument that AI can be biased takes two forms that are roughly analogous to John Searle’s (1980) acceptance of ‘Weak’ AI that functions as an instrument or tool and his rejection of the possibility of ‘Strong’ AI that could have human-like intelligence. The first *minimalist* argument regards bias in AI as likely an outcome of carelessness when designing algorithms, compiling datasets, or teaching systems during the machine learning stage of development. The second *maximalist* argument regards bias as endemic in society, responsible for social deprivation and inequality, and ready to be instrumentalised and expanded in AI. Those who hold this second view agree with the minimalist argument that we must not allow biased assumptions (e.g. ‘women don’t make good engineers’) to contaminate AI but they go on to argue that AI itself can be deployed to reduce bias in society. The overwhelming majority of AI policy discussions take the minimalist view but a minority blend this with elements of the maximalist view without distinguishing between them.

The minimalist view

This minimalist view employs a straightforward, somewhat technical, argument. AI will increasingly support interactions with important sectors like healthcare, finance, and criminal justice, and poor design (including biased datasets or algorithms) could lead to unfair outcomes for individuals. The solution is to introduce procedures to ensure that its design is good in the first place and, failing this, procedures that will not only rapidly identify and correct problems but which will provide redress to individuals who have suffered loss as a result. The minimalist view of bias and AI will usually regard negative effects of AI as

unintentional. It does not rely on a belief that structural bias exists in society. Its reactions to unintended effects are contingent on what is understood to be socially acceptable (Jobin et. al., 2019 pp. 8, 15; Access Now, 2018: 48-49; Perrault et. al., 2019). The minimalist view holds that AI must not hold back social progress being made by ethnic minorities, women, and other potentially marginalised groups. It does, however, not view AI as an instrument to advance marginalised groups.

The maximalist view

The maximalist view believes that society is structurally and endemically biased (Delgado and Stefancic, 2017: 11-13; Donnelly, 1999). It goes beyond the imperative that it should do no harm to take the view that unless AI is instrumentalised to counter this bias then it will inevitably be used to amplify it (Hofheinz, 2019). Proposals include the suggestion that algorithms can be created to eliminate bias if they are written for values such as 'non-discrimination, social inclusion and fairness' (Hofheinz, 2019: 4; Clegg, 2019; Montgomery, 2019). Another suggestion is that algorithms created by teams that are diverse in terms of both gender and ethnicity are more likely to be fair (The Royal Society, 2018: 3; European Commission 2020). Another is that AIs should be programmed to deliver outcome and error parity. That is to say designated protected groups should receive the same rate of positive as negative outcomes from AI decision systems (ICO, 2019). Still another extols the use of AI to identify 'bias' within culture (Chivot and Castro D, 2020).

5. *Diabolus ex machina*—the maximalist view of bias and AI

5.1 Introduction

In this section we leave the minimalist view of bias and AI to one side to examine the maximalist view. But before we do we can point out that the minimalist view of bias and AI is open to criticism on the grounds that, within its framework, an AI can be found to be 'biased' solely by the retroactive application of a norm of social acceptability. We can, however, be broad-minded, understanding 'biased' here as a figure of speech that unconsciously reflects social theory, while at the same time accepting that interventions to mitigate or prevent unintended effects can be justified as situational contingencies. For example, the response to an AI that recommends no ethnic minority candidates for shortlists to many open positions in a company could include intervention to ensure that just-below qualified ethnic minority candidates are shortlisted and, or, deeper long-term intervention aimed at improving the qualifications of marginalised individuals. The minimalist view is closely related to the principle of 'do no harm'. Those who hold it may consciously or unconsciously believe, like holders of the maximalist view, that bias is endemic in society. But they have no need to rely on this in order to justify their position.

5.2 The maximalist view of bias and AI relies on the *post hoc* fallacy

The maximalist view of AI is open to more serious challenges. Those proposing the maximalist view tend to rely on the *post hoc* fallacy to assert that negative social outcomes are themselves evidence of 'bias' even when no causal relationship can be shown between historical biases such as racism or sexism and the outcome in question. Dr. Brandie Nonnecke (2019), Director, CITRIS Policy Lab, at the University of California, Berkeley, uses a version of this argument when she asserts that the act of recording by police of high crime rates in minority neighbourhoods, itself, both constitutes and generates bias. In an article posted on the OECD website, she argues 'Since minority and low-income communities are far more likely to have been surveilled by police than prosperous white neighbourhoods, historical crime data at the core of predictive policing will provide a biased picture, presenting higher crime rates in communities that have been more heavily patrolled. As a result, predictive policing may amplify racial bias by perpetuating surveillance of minority and low-income communities'. This could, indeed, be the case. However Nonnecke wrongly implies that the recorded incidence of crime is a product of policing, rather than the incidence of crime itself, something that is manifestly untrue. Police have good reasons to

concentrate resources where armed criminals are active—often neighborhoods where young men use gun violence to enforce rules within markets and supply-chains for illegal drugs. In the United States many of these are black neighbourhoods where the overwhelming majority of victims of violent crime are also black and from the same communities (Center for Health Progress, 2018).

5.3 Hard cases that support the maximalist view of AI

In 2016 Florida law enforcement authorities were accused of mistakenly assigning a higher risk of reoffending to some black criminals than was borne out by subsequent conviction rates (Larson et. al., 2016; Young, 2018). Because such risk scores form part of courts' sentencing guidelines the likelihood of unjust sentences for black offenders was, accordingly, said to be higher. A widely-reported study published in *Science* in 2019 found that an algorithm used extensively in the U.S. healthcare system gave a lower priority to black than to white patients with similar illnesses (Obermeyer et. al., 2019; Evans, and Wilde Matthews, 2019). AI-driven bias has also been alleged in cases where advertisements for higher-paying jobs were served up to more male than female web users (Datta et. al., 2015; Kay et. al., 2015).

These and other cases have been used to argue that because bias is widespread in society we need rules to prevent the perpetuation of such bias in AI and that AI can be programmed to fight bias (Silberg, and Manyika, 2019; Obermeyer et. al., 2019; V. H., 2018). None of these arguments asserts that AI systems use *race* or *gender* to decide that black people should be locked up, or left to die, or that women are less suitable for senior positions. Rather they reach conclusions of bias based on information about outcomes that confirms an already held social theory about structural bias in society.

There is no doubt (see 4.1 above) that AI could help eliminate bias from human decision-making. But the conclusion that an AI is itself 'biased' because decisions that rely on it confirm already-held theories about structural bias sidesteps the real issue. This is that AI is frequently used to support choices that organisations want or have to make. The suggestion that the AI is somehow responsible avoids what may be a necessary, if uncomfortable, discussion about policy. AIs that reflect the higher proportion of black Americans who can't afford good healthcare or who commit firearms offences are mirrors both to underlying social conditions and to policies intended to address (or avoid) them.

The conclusion (e.g. Kosof, 2019) that unequal social outcomes prove 'bias' in an AI or, for that matter, in society, relies on a fallacy that can, itself, be explained in terms of cognitive bias. Its advocates assume that bad outcomes can *only* be evidence of bias even when other explanations are available. For example because it is risk-averse to allowing criminals the opportunity to reoffend, criminal justice policy in the United States bears down more heavily on repeat offenders, a disproportionately large percentage of whom are young black American men (Alper and Durose, 2018: 6). If the United States were an institutionally racist country there is no doubt that young black Americans would suffer disproportionately (to their share of the overall population) within its criminal justice system. But it does not follow from the fact that because young black Americans do suffer disproportionately (to their share of the overall population) within this system that the U.S. government is, therefore, racist. The route that commentators and activists take to reach this conclusion is a form of *post hoc* inductive reasoning that manifests as the following circular argument: 'Historically, racial bias caused bad outcomes for young black men, therefore, bad outcomes for young black men today are evidence that racial bias is alive and well in the United States' (Schagrin, 2019). The cognitive bias that leads to such reasoning can be explained as an example of the availability heuristic. The availability heuristic suggests that an individual will reach conclusions about an event or outcome based on the salience of information available to him at the time (APA Dictionary of Psychology, n.d.). For example, to those who are exposed to newspaper articles and social discussion which lean toward the view that U.S. society is endemically racist, the shooting of any armed black individual during an interaction with law enforcement officers may, no matter what the circumstances are, be deemed as evidence that the United States criminal justice system is inherently 'racist'. However white police officers are actually less likely to shoot black suspects than non-white officers (Johnson et. al., 2019). Furthermore, black Americans make up 14% of the country's population but account for over half of all gun-homicide victims. And nearly 90% of black homicides are perpetrated by black criminals (Center for Health Progress, 2018; National Crime Victims' Rights Week Resource Guide, 2017). Therefore, police are bound to interact disproportionately with armed black suspects. The high homicide and conviction rates for young black men have many causes all of which may justify state intervention. There is no reason to single out bias as a univariate cause.

5.4 Other explanations

COMPAS, the predictive system used by Florida law enforcement authorities, did not use data about ethnicity, which would be illegal, but it did use data which included prior convictions, gang membership, rates of arrests amongst friends, use of drugs, whether or not the offender came from a one or two parent family, and the history of criminality in the offenders' family. A disproportionate number of black felons are likely to tick these boxes and, hence, are more likely to be deemed at risk of reoffending (Siegel, 2018).
Krishna Gummadi, head of the Networked Systems Research Group at the Max Planck Institute for Software Systems in Saarbrücken, Germany is closer to the truth when he argues that Northpoint—which developed the COMPAS system designed to assign measurements of risk of reoffending to convicted criminals—was no more right or wrong than its critics who accused the system of racial bias (Larson et. al., 2016). In simple terms, he argues, the system's designers and customers placed more weight on the objective of minimising the risk of repeat offences and thus caused the system to show a higher rate of false positives than if their aim had simply been accuracy. As black felons are already over represented in the category of felons who are more likely to reoffend and as the system serves a wider criminal-justice policy of reducing the overall risk of crime this inevitably leads to the incarceration of blacks at higher rates than whites (Alper and Durose, 2018). In short, the system was disinclined to give convicted felons the benefit of the doubt. This caused it to show more false positives for black offenders. If it were aiming for accuracy it would likely have shown more false negatives, and some criminals would have been released early to reoffend (Spielkamp, 2017). The outcome was not a result of biased AI but of a policy objective that was biased against the possibility of repeat offences.

The (Obermeyer et. al.) study suggesting racial bias in the U.S. healthcare system is found, upon examination, to be a cost and not a race bias (Evans and Wilde Matthews, 2019). A main goal of the system is to control the cost of providing healthcare and, hence, the ability of patients to pay is a key factor in prioritising treatment. One effect of this is that the system recommends patients who cannot afford healthcare be treated as lower priority for certain conditions. More of these patients are black. As with COMPAS the system is used to inform a policy. In neither case does the use of these decision-support systems support the assertion that the policy in question is fair. But it is the policy, not the AI, that must be judged.

The exact reason why fewer senior executive jobs were served up to women web users has not been made clear but likely results from search settings that act as unintended proxies for gender. They may also partially arise from the fact that more of those searching the web for such roles are likely to be male. If web search engines have such a bias then either it is unintended or has been encoded into them by their users, including their female users who had not shown the same alacrity as men in searching for such roles.

Leaving aside hypothetical situations where datasets and algorithms are deliberately created to discriminate against people of specific ethnicities or genders, artificial intelligence shows a mirror to a world in which life is harder for the poor and for those from difficult family circumstances. It also shows a mirror to a world where ethnicities and genders display some tendencies to sort. More Asians run Asian restaurants (Buettner, 2008). More men work in construction. More women become carers (Eurostat, 2018). As Montgomery (2019) and Clegg (2019) imply, if algorithms that reflect these tendencies are 'biased' it is because our decision-making systems, individual and organisational, are also biased. But is inequality (or just difference) of outcome really manifest evidence of discrimination? Making such an argument to support incorporating fairness into AI design begs questions that have not yet been answered in the non AI-world. Burdening AI with the task of addressing what are, in the end, unresolved social issues is both technologically Utopian and a dishonest deflection of hard questions.

5.5 Using AI to cure bias in society

If identifying bias in a closed AI system is problematic then using AI as a tool to identify and eliminate bias in society presents an insurmountable challenge. For example, anomalies in the treatment of women and minorities may be uncovered by specially-designed AI systems but, as with the challenge of framing the problem that AI is trying to solve, it is unlikely to be possible to determine if bias is the cause.

Removing errors from an AI that could lead to unfair treatment of individuals is necessary. It is also desirable—in defence of the good intentions of those who hold the minimalist view of AI and bias—that we advance the interests of the disadvantaged though it is preferable that such decisions should be made, not on grounds of ethnicity, but of economic and educational status. Even without evidence of bias there are situations where making such changes will be acceptable. But the problem with identifying 'bias' as a univariate explanation for undesirable social outcomes isn't that it goes too far. It's that it cannot go far

enough. Causes of chronic social deprivation, for example, are likely to include many factors, such as drug abuse, imprisonment, family breakup, and untreated mental illness (Bourguignon and Chakravarty, 2019: 83-107; Brady, 2019). The focus on bias anthropomorphises AI and neglects its inductive power to examine the many factors that lead to deprivation and other social problems.

This is where the reasoning behind the idea that AI should be used to identify bias as a decisive cause of differential social outcomes breaks down. In addition to the intentionalist fallacy—that AI has agency—a second error bedevils those who believe in the maximalist view of AI. This is the view that hunting down and removing 'bias' from our society and culture is just a more effective version of the necessary and laudable task of ensuring that faulty AI does not cause unfair outcomes. It is a category error: a mistaken belief that the way we understand one situation permits us to understand a second situation that is syntactically similar (Ryle, 1949: 20-25). The difference is very great, however. The principle of 'do no harm' is very different from the fantasy that we can remake the world anew. Those who hold the maximalist view of AI regulation will, likely, make no categorical distinction between the two views.

5.6 Conclusion

Removing errors from Artificial Intelligence (AI) that could lead to unfair treatment is both necessary and achievable prior to its widespread adoption in areas that could affect the rights of individuals such as financial services, healthcare, education, social services, and criminal justice. Also, AI could and should be used to identify anomalies in public policy outcomes—whose explanations could include bias—with a view to intervention.

There is no doubt that AI can help us to arrive at a better understanding of the nature of social disadvantage which, always and everywhere, affects some groups more than others. But inferring, from data about social inequality, 'bias' as a univariate cause will distract from AI's potential to employ multivariate analysis to directly address social problems like crime and poverty. It may also encourage the misuse of data to 'scientifically' justify discrimination against supposedly privileged social groups.

The task of identifying, or even defining, bias within AI becomes very hard once our focus shifts from obvious errors in technology to questions of impact that cannot easily be known in advance.

This is made clear when we understand that AI which is ‘unbiased’ could still be misused—depending upon how its purpose is framed. Such ‘unbiased’ AI will likely be the instrument of the vast majority of future unfair decisions perpetrated with the assistance of technology in the future.

6. Three errors behind the maximalist view of AI

6.1 Introduction

For policymakers, for industry, and for legislators, AI's properties seem like those of a black hole. Most of us do not understand the processes that govern its interior. Not just human anxiety but thought processes that we use to arrive at conclusions about public policy are disappearing over its event-horizon towards a Singularity in which much that was formerly understood only in terms of human give and take seems to be infinitely compressed. Armed with social theory some advocates and policy analysts are keen to superimpose on AI an agenda to address society's ills which, if it is flawed in a non-AI world, is untenable or even potentially harmful in an AI world. Policymakers, social scientists and political activists are prone to three epistemological errors which encourage them to promote this maximalist view of AI. These are: *essentialism*, *inductivism*, and *intentionalism*. We have already, briefly, discussed the impact of the last two of these errors, inductivism and intentionalism. Inductivism leads to the belief that AI and human intelligence are similar which, in turn, encourages the intentionalist belief that AI must be treated as if it has a mind of its own or as if it instrumentalises the 'intentions' of structural social-bias. So far we have used the concept of *anthropomorphism* which, for our purposes, is the same as intentionalism. Here, we subject these errors to closer examination. This will enable us to understand not just that the maximalist view of bias and AI is wrong but *how* it is wrong. These three errors lead to the maximalist view of AI—the attempt to burden AI with the task of creating a just society—which may be better described as a Utopian proposal to create a constructivist singularity.

6.2 Essentialism

An ancient fallacy is that scientists can find the essence of things which often consists of a 'true reality' that is hidden behind mere appearances. When the Church encountered Galileo's heliocentric theory it refused to accept that it contradicted the essential 'truth' that the Sun and other planets went around the Earth (Popper, 1963: 97,108.; Deutsch, D., 1998: 74). It viewed Galileo's theory as being merely an appearance that was of instrumental use—good for astronomical calculations (Popper, 1963: 97). But it would not allow Galileo's claim that his theory described reality. In fact, insofar far as he believed his theory to be *essentially* true, Galileo was also mistaken. Against this the Church was determined to

uphold its doctrinal edifice that, behind the ‘appearance’ that the planets revolved around the Sun, there existed a hidden essential truth that the Earth was at the centre of the Universe.

Likewise, those who hold the maximalist view of bias and AI regard bias as the irreducible reality that explains poor social outcomes. They believe their assertion that society is endemically biased to be essentially true—in other words, an irreducible truth that cannot be further broken down into, or explained through the use of, other concepts (Popper, 1963: 104.). It is probably safe to say that their view of AI’s ability to identify multivariate factors in poor social outcomes is that it does not contradict bias as the best explanation for inequality but that it has an instrumental role in public policy. Likewise, the instrumentalist view of science is that it is the basis for engineering, i.e. problem solving and has no role in explaining reality. To give examples of what this means in 20th century science history: Einstein’s engagement with atomic theory was explanatory. While he worked on the Manhattan Project, Robert Oppenheimer’s was instrumental. Social scientists and advocates under their influence posit the interactions between essences of power, race, gender and bias as the explanation for social inequality. Meanwhile, corporations and governments have received and subsequently instrumentalised these concepts within governance and public communications frameworks.

However, recent scholarship has begun to undermine the idea of bias as a univariate explanation of differential social outcomes (Wynn, 2019; Bourguignon and Chakravarty, 2019; Friedman and Laurison, 2020). For example an examination of different rates of promoting women and men in information technology companies in Silicon Valley found factors other than straightforward sexist bias against women—though, as we can see, it did also find this. Operating on an organisational level these included: ‘referral hiring that leads to narrow pipelines of candidates from similar backgrounds’, ‘subjective evaluation criteria that open the door to bias during performance evaluations’ and, ‘lack of transparency and accountability in pay decisions that leads to unfairness’. While this points to gender bias as a factor it appears to be one of several. It seems to be a broader reflection of the heuristics of ‘we hire people like us’ something that is determined by educational paths and social background at least as much as gender.

The focus on ‘bias’ as an essence is closely related to the idea that gender or race are, also, essences. In the case of actual racism, we can see that it, too, is essence-mongering. However, many of those who oppose racism and other forms of bias have come to rely on

the idea that ‘race’ and ‘gender’ are, indeed, irreducible essences. Thus their concept of justice is shaped by concepts they oppose—racism and sexism (Appiah, 2018: 132). The property of self-similarity that links our verbal thinking tools to those things they are used to grasp could be one reason for this seeming paradox. Relationships of analogy between opposing concepts may lead to the spread, and not reduction, of bad ideas (Deutsch, 1998: 97; Lakoff, Johnson, 2003: 147-155). It’s worth asking whether the new doctrine of anti-racism could be an adaptation of the racist meme to a new environment?

The doctrinal fervor supporting essences within the social sciences as irreducible explanations of reality and their instrumentalisation by corporations and governments means that anyone who argues for a broader multivariate set of explanations for social outcomes—explanations that do not exclude bias—risks being cast in the same position as Galileo when he dared to suggest that his heliocentric theory was a better explanation for the visible movement of the planets. For example, Weiner’s (2014: 731-744) assertion that arguments against racism as a factor in social outcomes merely prove its existence seems to be the response of a theologian defending a sacred doctrine and not that of a social scientist engaging with an argument. A broader multivariate view, such as that provided by Friedman and Laurison (2020: 17-27) should prompt governments and corporations to resist the temptation to instrumentalise AI in the service of policies based on social essences.

6.3 Inductivism

Inductivism holds that scientists or, in our discussion, social scientists can infer explanations from observations. Given that policy proposals need explanations about the conditions they are intended to ameliorate mere observations, no matter in what quantity, of these conditions are not sufficient to create explanations. Therefore, inductivism holds that instances of a phenomenon can serve as evidence for, and that their accumulation will lend increasing support to, an explanation which, at some point, can be treated as true or axiomatic and, thereby, valid for decision-making (Popper, 1963: 64; Deutsch, 1998: 84, 90).

A moment’s thought will make it clear that an explanation or hypothesis must already be in the mind before one can make sense of any evidence. For example the inference made by Douglas Murray in his book *The Strange Death of Europe* (2017), that Europe is existentially threatened by Muslim migrants, relies on a preexisting, conservative, theory of its society. Against this he sets examples of criminal behaviour by a very small number of migrants while ignoring their gender, age, and socio-economic status, all of which likely offer better

explanations for their actions than their country of origin or their religion. Most importantly, Murray discounts the long-term likelihood that European Muslims and their descendents will contribute to and enrich European culture in proportion to the economic and educational opportunities available to them. No doubt European culture will evolve as a result of immigration but the hypothesis that it is 'threatened' relies on an inductive argument.

Another example of the fallacy of inductivism is the supposed epidemic of 'hate crime' that has been said to be engulfing Europe and the United States since the 1990s and which partly explains the difficulty in contesting essentialist doctrines that are common within the social sciences. In the early 1990s the idea of a 'hate epidemic' began to take shape in scholarly work.[4] A famous account, *The Rising Tide of Bigotry and Bloodshed: Hate Crimes* by Jack Levin and Jack McDavitt (1993), two leading scholars in the field of hate crimes, argued, without evidence, that such crimes were caused by 'resentment' of white Americans who feared economic competition from black Americans (Jacobs and Henry, 1996: 372). The impact of self-reporting of 'hate' incidents by groups with an interest in promoting the idea that there was an 'epidemic' of hate can also be seen in academic law journals during this time (Jacobs and Henry, 1996). Between 1992 and 1995 thirty-one articles in a leading database of articles by legal scholars supported the idea that the U.S was suffering a 'hate crime epidemic' (Jacobs and Henry, 1996: 374; Jacobs, and Potter, 1997.). These accounts largely relied on self-reporting by advocacy groups and were not reflected in police reports, prosecutions, or the outcomes of criminal trials.

Yet another example is provided by the Anti-Defamation League (ADL), a Jewish rights organisation, which has been collecting statistics on incidents of anti-Jewish bigotry or hostility since the 1970s. The ADL relies on data collected from over 25 regional offices and these, in turn, rely on victim and community group reporting, news reports, and reports by local police for their information. Incidents include reported crimes and non-criminal incidents of hostility or bigotry, some of which would, its unpleasantness notwithstanding, be protected speech under the First Amendment of the U.S. Constitution. Some of the other reported incidents are likely the result of simple misunderstanding (Jacobs and Henry, 1996: 378).

Newspapers, meanwhile, are particularly prone to misrepresenting events as motivated by bigotry when, in fact, they may just be plain incivility or common criminality (Civitas, 2016;

[4] This and the following paragraph are taken from MacDonnell P. (forthcoming), *History's Wrong Lessons: Why Our Response to online Disinformation and Hateful Speech Must Change*;

Reilly, 2019). Organisations that have followed the ADL's data collection model include Klan Watch and the Gay and Lesbian Anti-Violence Project. By 1996 more than half of U.S. states had modeled at least one section of their anti-hate laws on the ADL model (Jacobs and Henry, 1996: 380).

The widespread acceptance of inductive arguments in support of one policy or another likely helps to explain both the fear that AI will be misused and the hope that it can defeat bias. AI could certainly identify phenomena that may be markers for bias. But the attempt to use it to identify bias as an *essence* or, even more implausibly, ensure equality of social outcomes would instrumentalise the inductivist fallacy. Inductive reasoning has channeled the essentialist ideas that bias, race, and gender are irreducible categories and that interaction between them is a final explanation of poor social outcomes. An AI that was, following such principles, put to work as an instrument to achieve equality of social outcome would rapidly become—as Karl Kraus said of psychoanalysis—'that...illness for which it regards itself as the cure'.

6.4 Intentionalism

The final cognitive error that advocates and policy analysts make is the assumption that social and cultural phenomena, be they markets, marriage, epic poetry, or religious practice, have been intentionally constructed for identifiable purposes and that these purposes can be determined by scholars. The conservative reaction to Darwin's theory of evolution was intentionalist. Darwin's detractors could not accept that the rich array of complex life on Earth could be anything other than the product of a single supreme creator (Dennett, 1996: 19). The temptation to keep asking for an underlying cause to what Darwin observed formed a teleological infinite regress which, as with Descartes, ended with 'God' as the final answer. The emotional stakes were also high. Darwin's detractors felt that there had to be an underlying 'final' (in the Aristotelean sense) cause and that a Universe that no one intended was too dreadful to contemplate (Dennett, 1996: 23-24).

Similarly in the social sciences there has persisted, since Decartes, the view that behind all cultural and social phenomena and institutions lie intentions. This view has an emotional appeal similar to creationism and is the counterpart to creationists' response to Darwin in that it holds that we owe to purposeful design both social progress and forces that prevent social progress (Hayek, 1973: 9). It explains the intuitive appeal of 'bias' as the intentional smoking-gun, 'unconscious' or otherwise, behind social inequality. It is related to the first of

our fallacies, essentialism, in that it posits essential 'real' reasons for the march of history and roots these essences within the intentions of power structures like colonialism or capitalism such as we find in the writings of Edward Said (1978: 3, 7).

As an answer to the question, *why are some people socially disadvantaged?* 'bias' is often an exercise in deflection or, worse, scapegoating. It aims at an essential reality which, as a univariate explanation, doesn't exist. It is also an intentionalist attempt to short circuit the discussion. It would be better to follow Daniel Dennett's (1996: 24) example and, instead, ask, how does it come about that some people are socially disadvantaged? This invites a step-by-step answer with context—exactly the sort of problem where AI could serve as a useful assistant.

Descates' expression of radical doubt in *Meditations* concluded by placing the deity as the true basis for his certainty that both he and the world really existed. The insistence that intention is the final cause of phenomena studied by the social sciences, and within the liberal arts, are a consequence of Descartes' solution to this conundrum. However, all that has been achieved by rejecting God is the reversion to an alternative anthropomorphic idea that what is not designed by God must have its origin in human intentions (Hayek, 1973: 10-11).

This way of thinking also partly explains why many policymakers and commentators find it easier to locate AI as both a potential source and potential cure for all 'bias'. AI has designers and users and, thanks to the availability heuristic, these can be pressed into service as the 'intenders' behind unacceptable social outcomes whose real causes are multivariate and too difficult to face. Hence EU policymakers find it intuitive to consider affixing strict liability to all participants in the creation of AI systems.[5] We should remind ourselves, however, that conspiracy theories, such as the idea that Jews control the financial system, rely on the intentional fallacy for their explanations and on the inductivist fallacy for their 'evidence'. Here, I have borrowed the term 'intentional fallacy' from literary criticism where it means basing the interpretation of a work of literature on the imputed intentions of its author. My meaning is clearly wider though I believe that the implied definition is related to the original meaning. Intentionalism allows well-meaning social scientists and activists, as well as bigots to freely ascribe guilt to groups whose 'ancestors' can, depending upon the occasion, be associated with colonialism and slavery, or with the death of Christ. Thus Old

[5] See section 7.3.

Testament justice, the visiting of the sins of the fathers upon the sons, is legitimised in political discourse (Steele, 1990: 497-506).

6.5 Conclusion: a better way

Those who advocate the *maximalist* view of AI regulation—that it can identify and set right all bias in society—are guilty of all three of these epistemological errors: *essentialism*, *inductivism*, and *intentionalism*. Their cumulative effect is a view of the world that mistakes social outcomes for 'evidence', treats categories such as 'race', 'gender', and 'bias' as irreducible essences, and believes outcomes must be treated as intentional. Thus they believe that AI will give humans unlimited powers to remake society. They are mistaken.

A better approach is to treat theories, including theories in the social sciences, as conjectures and to ensure that they are falsifiable—that is realistically testable. While many of the contours of social inequality follow ethnic and gender distinctions, enough of them do not to show us that race and gender cannot themselves be the driving forces behind such inequality. The insistence that social intervention should be guided by racial or gender considerations ignores that the greatest social divides are defined by class, education, and skills (Friedman and Laurison, 2020: 17-27). Therefore, such essences as racism, or sexism may have no more role in addressing problems of social deprivation than Newtonian gravity has in Einstein's General Theory of Relativity.

7. Europe's response

7.1 Introduction

Europe's response to AI is, of course, partly influenced by these wider considerations of anxiety, philosophy, and civil rights, and partly by the Union's particular circumstances. This latter includes the concern that AI will worsen a perceived loss of economic and, even, democratic sovereignty that many believe is already well underway thanks to the US technology giants (von der Leyden, 2019: 13; Fleming et al., 2019; Swartz, 2019). For European policymakers, AI is both an opportunity for economic and social development, and a threat to the Union's 'technological sovereignty' and 'societal wellbeing' (European Commission, 2020). In its recent white paper (2020) the European Commission calls for a close synergy between investment in AI, and rules to prevent it from causing harm.

The Commission acknowledges that for AI to be deemed a risk (to either safety or human rights) it must be both deployed in a sector where such risks could arise and, itself, be their potential trigger. However the Commission proposes to combine the regulatory requirements of high-risk and non-high risk AI within a single framework that aims to be a quality standard—a counterpart to industry standards such as ISO 9000. This departs from a risk-based approach and would expand the scope of AI regulation beyond what is necessary.

With regard to protecting human rights the Commission proposes a mix of technical and governance solutions. One would be to mandate that the gender and ethnic profile of datasets match those of the EU country where the AI is to be used on the assumption that such segmentation will correlate with fairness despite the fact that biased treatment of individuals is easily discoverable and, therefore, easily deterred whether or not AI has supported decisions about individuals.

Beyond this the white paper proposes an expansive regime of micro-regulation requiring AI developers and users to justify every aspect of their technology—not just according to tried and trusted principles of human safety but according to the nebulous idea of group rights. Compounding this error is the proposal to require imported AI technologies to conform to EU design standards, including the need, if necessary, for retraining using European datasets. These proposals are a recipe that could both poison Europe's nascent AI industry and ignite

trade wars with regions of the world whose AI technologies are found not to meet Europe's sweeping standards. Finally the EU is considering a regime of strict liability at all stages of the AI development and deployment chain, something that could create insurmountable barriers to innovative development and deployment of AI.

7.2 Work before the white paper

Just as the EU put its powerful data protection regulation, the GDPR, to bed in April 2016 Big Data and Artificial Intelligence were becoming the hottest topics in technology (European Commission, 2016; Allied Market Research, 2018; Liu, 2019). The result is that the GDPR offers ample protection to individuals whose data may be used by an AI to treat them unfairly. The GDPR does not address AI directly but its principles are relevant to AI as is manifest in chapters two, six, and recital 71 of its text (European Commission, 2016). In simple terms—and excepting caveats which allow data to be collected and stored for scientific, statistical, or public interest purposes—the GDPR requires that: only the minimum quantity of personal data needed can be collected and this must only be used for the purpose for which it was collected; personal data must only be used in a way that is fair and understandable to the data subject; personal data must be kept in a form that does not allow its subject to be identified once the purpose for its collection no longer applies; personal data must be protected from theft or unauthorised processing. And if this isn't enough recital 71 of the regulation on 'profiling' encompasses almost all conceivable unfair practices that could take place through the use of an AI system. Recital 71 protects data subjects from automated decisions, without the right to secure human intervention, by systems which '...analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements'.

In 2018 Europe's Article 29 Working Party (which, in May of that year became the European Data Protection Board) published guidelines that, in effect, map how these GDPR principles would apply to AI (Article 29 Data Protection Working Party, 2018). In particular the guidelines call for data subjects to be notified if they are being profiled and for data protection impact assessments to be undertaken prior to the use of systems with evaluative and decision-making attributes. 'Inferred' data will also be treated as personal data for the purposes of data protection. Data protection impact assessments (DPIAs) are required for the use of 'sensitive data', or in situations when data is combined and used for 'new purposes' or, 'innovative use or applying new technical solutions' (Access Now, 2018: 12).

So in practice this would mean that most entities seeking to use sensitive data in AI would have to conduct DPIAs prior to deployment. To help this the guidelines set forth best practice for DPIAs with detailed criteria and methodology. The Working Party also published guidelines on consent and purpose limitation which will impose strict limits on the use of AI systems in areas that affect sensitive decisions about individuals (Article 29 Data Protection Working Party, 2018). These guidelines note that consent is only lawful if a user has a real choice about accepting a given term.

In 2018 the EDPS (European Data Protection Supervisor) Ethics Advisory Group report: *Towards a Digital Ethics* pointed out that some concepts in the GDPR, such as purpose limitation, data minimisation and data retention, are challenged by AI and that resolving these challenges through established data protection rules will be difficult (EDPS Ethics Advisory Group, 2018: 7). This paper recognizes the far-reaching potential power of AI and argues for 'ethical foresight' in its deployment (EDPS Ethics Advisory Group, 2018: 9). The paper also raises the prospect of AI posing a risk to democracy, the right to a fair trial and negative consequences 'when individuals are treated not as persons but as mere temporary aggregates of data processed at an industrial scale' (EDPS Ethics Advisory Group, 2018: 16). It argues for 'a foundation for the ethical assessment of a range of next generation algorithmic profiling techniques which are increasingly deployed in most sectors of activities and government'. Finally, in April 2019 the European Commission High-Level Expert Group on AI published Ethics guidelines for trustworthy AI, which stated that 'trustworthy AI should be: (1) lawful – respecting all applicable laws and regulations; (2) ethical – respecting ethical principles and values, and (3) robust – both from a technical perspective while taking into account its social environment' (European Commission, (2019b).

7.3 The European Commission's *white paper on Artificial Intelligence*

Of all of the policy papers published to date on AI the European Union's white paper is the most ambitious. It is aimed not only at protecting EU citizens from harmful AI but also at kickstarting a European AI industrial revolution. Based around European commercial and public sector data, this would draw on 'data pools' to achieve European AI dominance through the use of data 'based on European values and rules' (European Commission (2020).

The EU proposals correctly identify AI as high-risk when it can affect either safety or human rights outcomes within specific sectors such as transport, and healthcare. It then extends this definition to include AI systems in any sector that could affect workers' rights, consumers' rights, or recruitment practices. However, because the European Commission's document aims simultaneously at both the regulation and promotion of Europe's AI industry it contradicts this risk-based approach by then proposing that developers and users of low-risk AI be invited to join its proposed 'high-risk' regulatory framework in the interests of building 'excellence' and 'trust' in AI. This, effectively, extends its proposals to regulate high-risk AI to *all* AI in Europe, thus creating a regime that could apply to the most harmless AI applications a regime comparable to rules for aviation safety. EU regulators would apply the regime through 'conformity assessment'—which would include testing, inspection, and certification, of algorithms and datasets in 'development phase' together with ongoing oversight to identify any change in the design or application of AI systems. Also, the EU proposes AI originating outside the EU be retrained using EU data before permission is given for its deployment within the Union. Finally, the EU indicates that it is considering the expansion of strict liability for AI failures to all participants in its development and use.

In summary these measures—1) the inclusion of human rights and safety requirements within the same high-risk regime; 2) the expanded definition of 'high risk' to include whole sectors of activity; 3) the open invitation for *all* AI to take a non-returnable path into the high-risk regime; 4) the need AI entering Europe to be 'retrained' using EU data and; 5) the expansion of strict liability to the entire development and use chain of AI—would present significant barriers to innovation, to the transfer to Europe of AI technologies from outside the Union and, more than anything else, to the beneficial application of AI in the EU.

The EU's proposals are in line with other commentary that regards bias in AI as a danger to human rights and pledges itself to its outright removal. The problems with this approach have already been extensively discussed in this paper. But they can be summarised as follows: First is its lack of any substantive empirical basis. More often than not arguments against 'bias' in AI rely on anecdotal examples that are either the teething problems of, otherwise ineffectual, internet-search algorithms, or case examples of AI that is being used to support politically controversial policies (Simonite, 2018; Larson et. al., 2016; Northpointe, 2016). Second, in order to locate the problem of 'bias' in AI *itself* policy proposals rely on anthropomorphising AI as embodying biased *intent*, thus injecting an emotionally charged accusation of intentional malice into what may simply be a technical glitch (Simonite, 2018).

Third, the false analogy between engineering for safety and engineering to eliminate bias encourages the belief that ‘bias’ can be *technically* removed from AI and, further, that AI itself can help eliminate bias from society altogether (Hofheinz, 2019: 4; Clegg, 2019; Montgomery, 2019).

Finally, there are three broader arguments against the idea of using AI to remove bias from society:

1. There is little evidence that bias is a credible univariate explanation for social disadvantage (Friedman and Laurison, 2020: 17-27). There is evidence that, for example, racial bias in the wider society is a somewhat mild effect, rather than a cause, of inequality (Connor et. al., 2019).

2. The proposal to use AI for this purpose ignores its real potential to help generate powerful solutions to problems of social disadvantage through the use of multivariate data analysis (Stevenson, 2018; Kleinberg et. al., 2016).

3. It will be much easier to use ‘unbiased’ AI to abuse human rights and it is likely that this will make up the majority of such abuses. AI that is approved under the new regulations could be an even greater threat to human rights. For example ‘unbiased' AI systems could use parameters which—though they do not follow the protected characteristics of any ethnicity, gender, or faith—would unfairly affect individuals. Then there is the question of the wholesale use of AI surveillance by authoritarian governments. This is because AI that is ‘approved’ as ‘unbiased’ may, in fact, become unquestioned instruments of socially divisive policies. This is because it is impossible to discriminate in favour of one group without discriminating against another.

7.4 Conclusion

The EU proposals for regulating AI in Europe are ambitious and, given the extent of AI’s likely future role, could represent one of the most far-reaching regimes of economic and social intervention in the world. Proposals to regulate AI for safety, on the other hand, are reasonable and reflect an extension of an established and proven regime. However, the proposal to regulate AI to protect human rights, as if human rights and safety were the same thing, misconstrues that protecting human rights with AI should only be approached in a

negative sense: AI should be designed to avoid harm. AI itself should not be treated as a cause of human rights abuses much less a cure. To treat it in this way anthropomorphises it and opens the door to the maximalist view of AI regulation, which seeks to instrumentalise AI to fight ‘endemic bias’ in society.

The EU proposals to regulate AI mix the objectives of safety and human rights within one framework without acknowledging the important distinctions between them. While regulating AI for safety is an extension of an existing European regime that spans product liability and aircraft safety, protecting the vulnerable from unfair outcomes following the use of AI can easily be achieved through existing EU equality and privacy legislation.

Finally, inviting ‘no-high risk’ AI into a high-risk regulatory framework modeled on rules to protect safety will likely slow investment in European AI and will certainly slow the uptake of AI within the European economy.

8. Conclusion

Threats to safety from AI will likely arise from technical errors, in design, development, or in the application of the technology. They will always be due to an identifiable technical ‘smoking-gun’. AI that ‘threatens’ human rights on the other hand will almost never be identifiable in the same way. The widespread belief, expressed also in the EU white paper, that the risk to human rights from AI originates from its design is an error. The risk will come almost always from the circumstances of its application and, especially, the intentions of its users. 'Unbiased' AI will almost certainly be a greater instrument of social damage when it is misapplied.

End

Views expressed in this paper are those of the author and not those of Global Digital Foundation, which holds no corporate views.

Contact: Paul MacDonnell

Tel: +44 75 3458 0976

Email: paul.macdonnell@globaldigitalfoundation.org

About the author

Paul MacDonnell is executive director of the Global Digital Foundation and leads the organisation’s programme of technology-policy analysis, as well as its interaction with policymakers, legislators, and industry. Prior to joining the Foundation he was the head of European policy at the Center for Data Innovation, an affiliate of the Information Technology and Innovation Foundation. He previously represented Insurance Ireland, the insurance industry trade association, in Dublin and Brussels. In 2001 he co founded an economic policy forum in Dublin, Open Republic, which, as well as hosting events aimed at policymakers and politicians on such issues as pensions, taxation and the EU, was the Irish publisher of the *Economic Freedom of the World Report*. He has extensive broadcast and print media experience and has written for *Economic Affairs*, and the *Wall Street Journal*. He holds a degree in medieval English literature and philosophy from Trinity College Dublin and an M.B.A. from University College Dublin.

Appendix

The global reaction to AI

Focusing both on the development opportunities and on the possible social costs an avalanche of comment as well as governance and regulatory proposals has flowed from universities, consortia, legislators, regulatory bodies, think tanks, and technology Companies (Perrault et. al., 2019: 14,148-151; Access Now, 2018). Thanks to the Human-Centered AI Institute at Stanford University, we can get a sense of both AI's development and of the ideas about policy that are responses to it. Since 2016 the Institute has published an annual global survey that scopes the growing significance of AI using metrics that range from the level of investment to media awareness (Perrault et. al. 2019). The 2019 edition, The AI Index 2019 Annual Report, includes details of: the volume of peer-reviewed AI papers published; the performance of AI-related technologies; investment in AI startups and AI investment; the rate of hiring of AI professionals; the most commonly-reported ethical challenges of AI; and mentions of AI and Machine Learning (ML) during debates in the U.S. Congress, the UK Parliament, and the Parliament of Canada. The report shows the following:

Volume of peer-reviewed AI papers published

During the 20 years between the late 1990s and 2018 the number of peer-reviewed papers and conference publications that referenced AI-related issues grew from 1% and 3% to 3% and 9% respectively, a 300% increase (Perrault et. al. 2019: 14).

Improved performance of AI-related technologies

The ability of image classification algorithms to accurately identify items, including animals, within an image improved from over 84% in 2014 to about 97% in 2019 (Perrault et. al. 2019: 207). Meanwhile the time taken to train an image classification algorithm on cloud infrastructure fell from about three hours in October 2017 to less than 90 seconds in July 2019. The cost of such training also fell significantly (Perrault et. al. 2019: 208).

Investment in AI startups and AI investment across the world economy

Between 2010 and 2018 global investment in AI startups rose from $1.3bn to $40.8bn—with $37.4bn for 2019 as at 4th November—a global average increase of 48% per annum

(Perrault et. al. 2019: 88).[6] In addition to the $37.4bn startup-investment figure as at 4th November 2019, global private AI investment across the rest of the economy for the year was $34bn for mergers and acquisitions, $5bn for Initial Public Offerings (IPOs) and $2bn for minority stake private investment (Perrault et. al. 2019: 94).

The rate of hiring of AI professionals

Countries where the rate of hiring of AI professionals has more than doubled since 2015-16 include: the United States, Germany, the Netherlands. Singapore, Brazil, Australia, Canada, Turkey, and South Africa (Perrault et. al. 2019: 73).[7]

The most commonly-reported ethical challenges

The most commonly-reported ethical challenges covered in documents dealing with the ethical principles of AI were: fairness, interoperability and explainability, transparency, accountability, and data privacy. Fairness, and interoperability and explainability were mentioned in more than 80% of these documents while accountability and data privacy were mentioned in over 60% (Perrault et. al. 2019: 148-149).

Response of global news media

A survey of 60,000 global English news sources and over 500,000 blogs on AI ethics between August 2018 and August 2019 revealed that 32% of all AI-related articles covered ethical guidelines proposed or discussed by governments and supranational bodies like the EU and the OECD. Eleven percent included commentary from advisory groups attached to tech giants such as Google, Facebook, and Microsoft. Concerning specific AI technologies, 13% of articles discussed the ethical implications of facial recognition (Perrault et. al. 2019: 150-151).

AI-ML related legislation

In the U.S. Congress AI or ML were mentioned fewer than 30 times in congressional research service reports, committee reports, and legislation during each of the years between 2002-2017; more than 90 times in 2018; and over 100 times in 2019 (Perrault et. al. 2019: 139). AI or ML were mentioned in the UK parliament fewer than 25 times during each of the years 1995-2015; more than 25 times during 2016; more than 85 times during 2018;

---

[6] This figure includes only AI companies that received more than $400k.

[7] Because it relies on statistics gained from LinkedIn *The AI Index 2019 Annual Report* understates figures for China and India. However see Shen J. (2019), and, *The Economic Times* [of India] (2018).

and more than 100 times during 2019. In the Parliament of Canada AI or ML were not mentioned at all during the years 2002-2013; fewer than 5 times during 2014; more than 20 times during 2017; more than 35 times during 2018; and (as at September of that year) more than 10 times during 2019 (Perrault et. al. 2019: 139).

Everyone agrees: we need rules

As the above survey shows, apart from the need to seize the economic gains from AI, the primary concern of legislators, regulators, policy analysts, industry generally, and the technology industry in particular, is to find ways to prevent AI from being used to attack human rights. Because technologies that support AI are complex and its likely use and consequences are respectively deep and far-reaching such proposals are, for now, often general and principles-based. Overwhelmingly they focus on the need for AI systems: to be interoperable, explainable, transparent, and fair; not to violate privacy; and for its developers and users to be accountable.

A good example of the broad nature of the discussion about AI can be found in a room document for the 38th International Conference of Data Protection and Privacy Commissioners held in 2016 in Marrakesh in which delegates were asked: 'How can DPAs [Data Protection Authorities] support the right to information from the data subject when confronted with big data, artificial intelligence and machine learning? How to evaluate the bias in automated decisions when artificial intelligence and machine learning is used? How can DPAs supervise appropriately an organisation using intensively big data, artificial intelligence and machine learning? Should DPAs create their own pool of artificial intelligence experts and resources to be able to re-create and analyse the models used by the organisations under supervision?' (International Privacy Conference, 2016).

It is notable that even when focusing on the ethical challenges nowhere do organisations address the question of using AI to address the question of bias in society generally, much less the difficulty of identifying it within closed IT systems. This suggests that, for now, the maximalist view of AI is a relatively recent development.

Bibliography

Reports, Articles, Papers, Blogs

Access Now (2018), *Mapping Regulatory Proposals for AI in Europe*, Vodafone Institute for Society and Communications, November, 2018, available at: https://www.accessnow.org/cms/assets/uploads/2018/11/mapping_regulatory_proposals_for_AI_in_EU.pdf.

Adorno T. and Horkheimer M. (1956), 'Towards a New Manifesto?', Discussion in Spring 1956, published in *New Left Review*, Sep-Oct 2010.

Allied Market Research, (2018), webpage, available at: https://www.alliedmarketresearch.com/artificial-intelligence-market.

Alper M. and Durose M. R. (2018), '2018 Update on Prisoner Recidivism: A 9-Year Follow-up Period (2005-2014)' U.S. Department of Justice, May 2018, available: https://www.bjs.gov/content/pub/pdf/18upr9yfup0514.pdf.

Antonio R. J. (1983), 'The Origin, Development, and Contemporary Status of Critical Theory', *The Sociological Quarterly*, Vol. 24, No.3, Summer: 325-351.

APA Dictionary of Psychology (n.d.), webpage, 'availability heuristic', available at: https://dictionary.apa.org/availability-heuristic.

Article 29 Data Protection Working Party (2018), *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, Revised and adopted on 6th February 2018, available at: http://ec.europa.eu/newsroom/article29/document.cfm?doc_id=49826.

Article 29 Data Protection Working Party (2018), *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679*, available at: A http://ec.europa.eu/newsroom/document.cfm?doc_id=44137.

Bourguignon F. and Chakravarty S. R.In: Chakravarty S. (eds) (2019), *Poverty, Social Exclusion and Stochastic Dominance. Themes in Economics* (Theory, Empirics, and Policy). Springer, Singapore: 83-107.

Brady D. (2019), ‘Theories of the Causes of Poverty’, *Annual Review of Sociology*, 2nd April, 2019.

Buettner E. (2008), ‘Going for an Indian’: South Asian Restaurants and the Limits of Multiculturalism in Britain’, *Journal of Modern History*, Chicago University Press, Vol. 80, No. 4: 865-901., available: https://www.jstor.org/stable/10.1086/591113?seq=1#metadata_info_tab_contents.

Center for Health Progress, (2018), ‘Patterns of gun violence’, Center for Health Progress, 24th July 2018, available: https://centerforhealthprogress.org/blog/patterns-of-gun-violence/?gclid=Cj0KCQiA4NTxBRDxARIsAHyp6gC9knjAt55z91IuFV7hDuU1TX4m_qy-z6FEVWU91nNNEZbtMXVmbuoaAtveEALw_wcB.

Chivot E. and Castro D. (2020), ‘Contribution to the European Commission’s Public Consultation on Gender Equality Strategy 2020-2024’, Center for Data Innovation, Washington DC, 13th February, 2020, available: https://s3.amazonaws.com/www2.datainnovation.org/2020-eu-gender-equality.pdf.

Civitas (2016), *Hate crime: the facts behind the headline*s, Civitas, London, October, 2016. Available: http://www.civitas.org.uk/reports_articles/hate-crime-the-facts-behind-the-headlines/.

Clegg A. (2019), ‘Will AI bring gender equality closer?’, *Financial Times*, 8th March 2019.

CNA (2020), 'WHO chief calls COVID-19 'enemy against humanity', CNA, 19th March, 2020, available: https://www.channelnewsasia.com/news/world/coronavirus-covid-19-who-enemy-against-humanity-12554256.

Connor P. et. al. (2019),‘Income Inequality and White-on-Black Racial Bias in the United States: Evidence from Project Implicit and Google Trends’, *Psychological Science*, Feb. 2019, Vol. 30, Issue 2, p205-222.

Council of Europe, (2017), *Algorithms and human rights: study on the human rights dimensions of automated data processing techniques and possible regulatory implication*s, Council of Europe, Strasburg, available at: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5.

Curiel E. (2019), ‘Singularities and Black Holes’, *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), available: https://plato.stanford.edu/archives/spr2019/entries/spacetime-singularities.

Dastin J. (2018), ‘Amazon scraps secret AI recruiting tool that showed bias against women’, *Reuters*, 10th October, 2018, available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Datta A. et. al. (2015), 'Automated Experiments on Ad Privacy Settings', *Proceedings on Privacy Enhancing Technologies*, Vol. 2015, Issue 1, available: https://content.sciendo.com/view/journals/popets/2015/1/article-p92.xml.

Dennett D. C. (2015), 'The Singularity—an Urban Legend?', in *What to think about machines that think*, ed. J. Brockman, Harper Perennial, New York.

Donnelly R. (1999), ‘Racism endemic to British society, not just to police, says Ashdown’, *Irish Times*, 24th February, 1999, available: https://www.irishtimes.com/news/racism-endemic-to-british-society-not-just-to-police-says-ashdown-1.156206.

The Economic Times (2018), ‘Over 4,000 artificial intelligence job roles vacant on talent shortage: Report’, *The Economic Times*, 17th December 2018, available at: https://economictimes.indiatimes.com/jobs/over-4000-artificial-intelligence-job-roles-vacant-on-talent-shortage-report/articleshow/67131803.cms.

EDPS Ethics Advisory Group (2018), *Towards a Digital Ethics*, EDPS, Brussels. Available at: https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf.

European Commission, Ethics Guidelines for Trustworthy AI (2019a), European Commission, Brussels, 8th April, 2019, available, https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

European Commission, High-Level Expert Group on AI (2019b), Ethics guidelines for trustworthy AI, European Commission, Brussels, 8th April, 2019.

European Commission (2020), *White Paper on Artificial Intelligence: a European approach to excellence and trust*, European Commission, 19th February, 2020, available: https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

European Commission (2020b), *Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics*, Report from the Commission to the

European Parliament, The Council and the European Economic and Social Committee, Brussels, 19th February, 2020.

Eurostat (2018), 'Jobs still split along gender lines', webpage, European Commission, 7th March, 2018, available at: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180307-1.

Evans M. and Wilde Matthews A. (2019), 'Researchers Find Racial Bias in Hospital Algorithm', *The Wall Street Journal*, 28th October, 2019, available: https://www.wsj.com/articles/researchers-find-racial-bias-in-hospital-algorithm-11571941096

Fleming S. et al. (2019), 'Brussels steps up pressure on US over global digital tax deal', *Financial Times*, 5th December, 2019, available: https://www.ft.com/content/db6148fc-1748-11ea-9ee4-11f260415385.

From Principles to Practice : an interdisciplinary framework to operationalise AI ethics (2020), VDE, Bertelsmann Stiftung.

Hao K. (2019), 'This is how AI bias really happens—and why it's so hard to fix', *MIT Technology Review*, 4th February, 2019, available: https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

Hofheinz P. (2019), *The Ethics of Artificial Intelligence: How AI can End Discrimination and Make the World a Smarter, Better Place*, Lisbon Council, Brussels, available, https://lisboncouncil.net/publication/publication/148-the-ethics-of-artificial-intelligence-how-ai-can-end-discrimination-and-make-the-world-a-smarter-better-place.html.

ICO (2019), Human bias and discrimination in AI systems, webpage, ICO, 25th June 2019, available: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-human-bias-and-discrimination-in-ai-systems/.

International Privacy Conference (2016), *Artificial Intelligence, Robotics, Privacy and Data Protection*, Marrakesh, October 2016, available at: https://edps.europa.eu/sites/edp/files/publication/16-10-19_marrakesh_ai_paper_en.pdf.

Jacobs J. B. and Henry J. S (1996), 'The Social Construction of a Hate Crime Epidemic', *Journal of Criminal Law & Criminology*, Vol 86 No 2. Pp.366-391.

Jacobs, J., and Potter, K. (1997). Hate Crimes: A Critical Perspective. *Crime and Justice*, 22, 1-50. Retrieved November 8, 2020. Available: http://www.jstor.org/stable/1147570.

Jobin A. I. et. al. (2019), 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, 2nd September, 2019: 8, 15.

Johnson D. J. et. al. (2019), 'Officer characteristics and racial disparities in fatal officer-involved shootings', *Proceedings of the National Academy of Sciences of the United States of America*, 22nd July, 2019. available: https://www.pnas.org/content/pnas/116/32/15877.full.pdf.

Kay M. et. al. (2015) 'Unequal Representation and Gender Stereotypes in Image Search Results for Occupations' in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM, 2015): 3819–3828, available: https://www.csee.umbc.edu/~cmat/Pubs/KayMatuszekMunsonCHI2015GenderImageSearch.pdf.

Kleinberg J. et. al. (2016), 'A Guide to Solving Social Problems with Machine Learning', *Harvard Business Review*, 8th December, 2016, available: https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning.

Kosof M. (2019), 'Alexandria Ocasio-Cortez Says Algorithms Can Be Racist. Here's Why She's Right', *Live Science*, 29th January, 2019, available: https://www.livescience.com/64621-how-algorithms-can-be-racist.html.

Larson J. et. al. (2016), 'How We Analyzed the COMPAS Recidivism Algorithm', *ProPublica*, 23rd May, 2016, available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Liu S.(2019), 'Forecast of Big Data market size, based on revenue, from 2011 to 2027', Statista (webpage), 9th August, 2019, available at: https://www.statista.com/statistics/254266/global-big-data-market-forecast/.

MacMillen A. (2019), 'Can an Algorithm Identify Repeat Offenders?', *Chicago Policy Review*, 12th March, 2019, available: https://chicagopolicyreview.org/2019/03/12/can-an-algorithm-identify-repeat-offenders/.

McCarthy J. (n.d.) 'What is AI? / Basic Questions', webpage. Project JMC. Available: http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html.

McQuinn A. (2015), ‘From Kodak to Google, How Privacy Panics Distort Policy’, *TechCrunch*, 1st October, 2015, available: https://techcrunch.com/2015/10/01/from-kodak-to-google-how-privacy-panics-distort-policy/.

Metz C. (2016), ‘How Google’s AI Viewed the Move No Human Could Understand’, *Wired*, 14th March, 2016, available: https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/.

Mind Matters News (2019), ‘A type of reasoning AI can’t replace’, *Mind Matters News*, 10th October, 2019, available: https://mindmatters.ai/2019/10/a-type-of-reasoning-ai-cant-replace/.

Montgomery J. (2019), ‘AI for society: creating AI that supports equality, transparency, and democracy’, webpage, The Royal Society, 21st February, 2019, available: https://blogs.royalsociety.org/in-verba/2019/02/21/ai-for-society-creating-ai-that-supports-equality-transparency-and-democracy/.

Nasa (1962), ‘John F. Kennedy Moon Speech - Rice Stadium’, Nasa, available: https://er.jsc.nasa.gov/seh/ricetalk.htm.

National Crime Victims’ Rights Week Resource Guide (2017), Homicide Fact Sheet, NCVRW, available, https://www.ncjrs.gov/ovc_archives/ncvrw/2017/images/en_artwork/Fact_Sheets/2017NCVRW_Homicide_508.pdf.

Nonnecke B. (2019), ‘Artificial intelligence can make our societies more equal. Here’s how’, World Economic Forum, 21st September, 2017, available: https://www.weforum.org/agenda/2017/09/applying-ai-to-enable-an-equitable-digital-economy-and-society/.

Northpointe, *Demonstrating Accuracy Equity and Predictive Parity*, Northpointe, 8th July, 2016, available: http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Obermeyer Z. et. al. (2019), ‘Dissecting racial bias in an algorithm used to manage the health of populations’, *Science*, 25th October, 2019, Vol 366, Issue 6464: 447.453.

Perrault R. et. al. (2019), *The AI Index 2019 Annual Report*, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, December 2019, available at: https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_report.pdf.

Reilly W. (2019), 'The Hate-Crime Epidemic That Never Was: A Seattle Case Study', *Quillette*, 7th July, 2019. Available: https://quillette.com/2019/07/07/the-hate-crime-epidemic-that-never-was-a-seattle-case-study/.

Royal Society (2018), *AI, society and social good*, Note of discussions at a Royal Society and American Academy workshop', Centre for Advanced Study in the Behavioural Sciences, Stanford University, 8th November, 2018, available: https://royalsociety.org/~/media/policy/Publications/2018/ai-and-society-workshop-notes.pdf?la=en-GB.

Royal Society (2019), 'The AI revolution in scientific research', available: https://royalsociety.org/-/media/policy/projects/ai-and-society/AI-revolution-in-science.pdf?la=en-GB&hash=5240F21B56364A00053538A0BC29FF5F.

Satariano A. and Pronczuk M. (2020) 'Europe, Overrun by Foreign Tech Giants, Wants to Grow Its Own', *New York Times*, 20th February, 2020, available: https://www.nytimes.com/2020/02/19/business/europe-digital-economy.html.

Schagrin M. L. (2019), 'Fallacy', *Encyclopedia Britannica*, Encyclopædia Britannica, inc., 7th August, 2019, available: https://www.britannica.com/topic/fallacy.

Searle J. R. (1980), 'Minds, brains, and programs', *The Behavioural and Brain Sciences* (1980).

Shen J. (2019), 'AI engineer the most popular job in China: report' *technode*, 7th March, 2019, available at: https://technode.com/2019/03/07/image-recognition-engineer-job.

Silberg J., and Manyika J. (2019), 'Tackling bias in artificial intelligence (and in humans)', McKinsey Global Institute, June 2019, available at: https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans.

Silver D. and Hassabis D. (2017), 'AlphaGo Zero: Starting from scratch', Blog, 18th October, 2017, available: https://deepmind.com/blog/article/alphago-zero-starting-scratch.

Steele S. (1990), White Guilt, *The American Scholar*, Vol. 59. No. 4 (Autumn 1990) pp.497-506., Available: https://www.jstor.org/stable/pdf/41211829.pdf?casa_token=DpMLr0ezpkEAAAAA:3mm2egM9iuKwg3lKyBOsPZQnz-7YULsVb8KgUVvyZTYfxF8c5xxxysJyqlm6n7eGKxOABlt5NRVcSeQ8mSuEOreRa0pSW4U-cMbgtCOpKI20eM1iAhU6.

Stevenson L. (2018), ‘Artificial intelligence: how a council seeks to predict support needs for children and families’, webpage, CommunityCare, Surrey County Council, 1st March, 2018, available: https://www.communitycare.co.uk/2018/03/01/artificial-intelligence-council-seeks-predict-support-needs-children-families/.

Stolton S., (2020) ‘LEAK: Commission considers facial recognition ban in AI ‘white paper’, *Euractiv*, 17th January 2020, available at: https://www.euractiv.com/wp-content/uploads/sites/2/2020/01/AI-white-paper-EURACTIV.pdf.

Swartz A. (2019), ‘Has Europe lost its tech giants?’, *Hackerroom*, 26th January, 2019, available: https://hackernoon.com/had-europe-lost-its-tech-giants-ab8df9669aeb.

Sweeney P. (2018),‘One problem to explain why AI works’, The Explainable Startup, available, https://www.explainablestartup.com/2018/05/one-problem-to-explain-why-ai-works.html.

Siegel E. (2018), ‘How to Fight Bias with Predictive Policing’, *Scientific American*, 19th February, 2018, available, https://blogs.scientificamerican.com/voices/how-to-fight-bias-with-predictive-policing/.

Simonite T. (2018), ‘When It Comes to Gorillas, Google Photos Remains Blind’, *Wired*, 11th January, 2018, available at: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Spielkamp M. (2017), ‘Inspecting Algorithms for Bias’, *MIT Technology Review*, 12th June, 2017, available: https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/.

Tolan et. al. (2019), 'Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia', International Conference on AI and Law, Montreal, QC, Canada, 17–21 June, 2019, available: https://chato.cl/papers/miron_tolan_gomez_castillo_2019_machine_learning_risk_assessment_savry.pdf.

Turing A. (1950), Computing Machinery and Intelligence. *Mind* 49: 433-460.

Ulam S. (1958), 'John Von Neumann 1903-1957.' *Bulletin of the American Mathematical Society*, 64, part 20.

V. H. ‘A.I. (2018), ’Bias' could create disastrous results, experts are working out how to fight it’, CNBC.Com, 14th December, 2018, available: https://www.cnbc.com/2018/12/14/ai-bias-how-to-fight-prejudice-in-artificial-intelligence.html.

Vinge V. (1993), 'The Coming Technological Singularity: How to Survive in the Post-Human Era', in *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, G. A. Landis, ed., NASA Publication CP-10129: 11–22.

von der Leyden U. (2019), *My agenda for Europe*, European Commission (2019).

What-If Tool, webpage, available: https://pair-code.github.io/what-if-tool/.

Whitfield S. (2014), 'Refusing Marcuse: 50 Years After *One-Dimensional Man*', *Dissent*, Fall 2014, available: https://www.dissentmagazine.org/article/refusing-marcuse-fifty-years-after-one-dimensional-man.

Weiner M. F. (2014),'The Ideologically Colonized Metropole: Dutch Racism and Racist Denial', *Sociology Compass*, 8/6 (2014): 731-744, available: https://onlinelibrary-wiley-com.uoelibrary.idm.oclc.org/doi/epdf/10.1111/soc4.12163.

Wynn A. (2019), 'Why Tech's Approach to Fixing Its Gender Inequality Isn't Working', *Harvard Business Review*, 15th October, 2019.

Young E. (2018), 'A popular algorithm is no better at predicting crimes than random people', *Atlantic*, 17th January, 2018, available: https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/.

Zunger Y. (2019), 'Can an Algorithm be Racist?', *Mind Matters News*, 3rd January, 2019, available:https://mindmatters.ai/2019/01/can-an-algorithm-be-racist/.

Books

Appiah K. A. (2018), *The Lies That Bind: Rethinking Identity*, New York, Liveright. (originally broadcast as a BBC Reith Lecture, Mistaken Identities: Colour, on November 12th, 2016).

Dennett D. C. (1996), *Darwin's Dangerous Idea: Evolution and the Meanings of Life*, London, Penguin Books.

Dennett, D. C. (2013), *Intuition Pumps and Other Tools for Thinking*, London, Allen Lane.

Dennett, D. C. (2017). *From Bacteria to Bach*. London, Penguin Books.

Delgado R. and Stefancic J. (2017), *Critical Race Theory: An Introduction*, Third Edition. New York, NYU Press.

Deutsch, D. (1998), *The Fabric of Reality*. London, Penguin Books.

Emord J. *Freedom, Technology, and the First Amendment*, Pacific Research Institute for Public Policy, 1991.

Friedman S. and Laurison D. (2020), *The Class Ceiling: Why it pays to be privileged*. Bristol, Bristol University Press.

Gramsci (1971), *Selections from the Prison Notebooks*, Lawrence & Wishart, London.

Hayek F. A. (1973), 'Rules and Order', in *Law, Legislation and Liberty*, London, Routledge 1982.

Kurzweil R. (2005), *The Singularity Is Near: When Humans Transcend Biology*. New York, Penguin,

Lakoff G., Johnson M. (2003), *Metaphors We Live By*, Chicago, University of Chicago Press.

Levin J. and McDevitt J. (1993), *The Rising Tide of Bigotry and Bloodshed*. New York, Plenum.

Marcuse H. (1964), *One Dimensional Man*, London, edn., Routledge, 1991.

Matsuda M. J. (1993), *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Westview Press, 1993.

Murray D. (2017), *The Strange Death of Europe*, London, Bloomsbury.

Popper K. R. (1963), *Conjectures and Refutations: the Growth of Scientific Knowledge*, London: Routledge.

Ryle G. (1949), *The Concept of Mind*, London, Penguin University Books (reissue, 1973).

Said E. W. (1978), *Orientalism*, New York, Pantheon Books.

Legislation

Council of the European Union (2001), Directive on General Product Safety, Directive 2001/95/EC, 3 December 2001, available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32001L0095.

Council of the European Union (2006),Directive on Machinery, Directive 2006/42/EC, 17th May 2006, and amending Directive Directive 95/16/EC (recast), available: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1579870568589&uri=CELEX:02006L0042-20190726.

Council of the European Union (2014),Directive on Radio Equipment, Directive 2014/53/EU, 14th April, 2014, available: https://eur-lex.europa.eu/search.html?qid=1579875189091&text=Radio%20equipment%20directive&scope=EURLEX&type=quick&lang=en.

Council of the European Union (1985), Directive on Product Liability, Directive 85/374/EEC, 25th July, 1985, available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31985L0374.

Council of the European Union (2016), Police Directive, Directive 2016/680, 27th April, 2016, available: http://data.europa.eu/eli/dir/2016/680/2016-05-04.

Council of the European Union (2019), Open Data Directive, Directive 2019/1024, 20th June, 2019, available: http://data.europa.eu/eli/dir/2019/1024/oj.

European Commission (2016), General Data Protection Regulation (GDPR), Official Journal of the European Union, 27th April, 2016, Brussels, available: https://gdpr-info.eu/.