the Lisbon council 15 YEARS
think tank for the 21st century

DSM 2.0
A LISBON COUNCIL INITIATIVE

Interactive policy brief
Issue 23/2018

The Ethics of Artificial Intelligence

How AI Can End Discrimination and Make the World a Smarter, Better Place

By Paul Hofheinz

Paul Hofheinz is president and co-founder of The Lisbon Council, a Brussels-based think tank. Founded in 2003 as a non-profit, non-partisan association, the Lisbon Council has emerged as one of Europe's pre-eminent voices on social and economic change.

Judging from the fierce debates, few issues have posed such a challenge to mankind as artificial intelligence (AI) and machine learning. First and foremost, there is the evident, underlying existential test.[1] Mankind has dominated the planet for around 190,000 years, more or less since homo sapiens first emerged in the Kenya Rift Valley with their large craniums, light skeletons, advanced social structures and opposable thumbs.[2] And, while we have learned to live with machines – like bulldozers, cranes and trucks – that can lift and transport much more weight than we can carry ourselves, the notion that there are machines that might calculate much better than we can – and do it outside of our direct supervision, arriving at conclusions "autonomously" and through processes that defy the linear logic of old-style computer programming – is something that seems to strike deeply at our very sense of self.[3]

It's not so much that we can't live with the idea of machines that think better, faster and autonomously from us; it is more that the idea itself seems to many to imply that humans are destined to be overtaken, irrationally abused, even made redundant here on earth, perhaps as soon as the day after tomorrow. Many of the comments and analysis written on this topic – some learned, like Carl Benedikt Frey and Michael A. Osborne's seminal work on "The Future of Employment: How Susceptible are Jobs to Computerisation?"; some deeply conjectural, like the *Blade Runner* and *Terminator* films – show the marks of a deeply insecure civilization, one that feels its very existence is being threatened by machines.[4] Much like people, we understand and express that insecurity not through direct acknowledgement of our sometimes deeply disguised low self-esteem, but by a

The **DSM 2.0: Digital Futures Forum** initiative, led by the Lisbon Council, is an open, inner-disciplinary, multi-stakeholder forum for exploring the coming wave of new and impactful technologies, the impact they will have on society and the challenge for citizens and regulators alike.

The interactive policy brief seeks to make knowledge more accessible through online circulation, interactive features, a web-friendly format and a hot-linked bibliography that begins on page 10.

1
The author would like to thank Prabhat Agarwal, Alessandro Annoni, Robert D. Atkinson, Daniel Braun, Charina Chou, Christian D'Cunha, Peter Fatelnig, Marie Frenay, Ben Gomes, Jörgen Gren, Juha Heikkilä, John Higgins, Tim Hwang, Luukas Ilves, Jens-Henrik Jeppesen, Björn Juretzki, Kaspar Kala, Kaja Kallas, Thibaut Kleiner, Stéphanie Lepczynski, Guido Lobrano, Nicklas Lundblad, James Manyika, David Osimo, Patricia Reilly, Keith Sequeira, Siim Sikkut, Dirk Staudenmayer, Vladimir Šucha, Pawel Świeboda, Chiara Tomasi, António Vicente, Lynette Webb and Andrew W. Wyckoff for taking part in a series of roundtables and discussions where many of these ideas were developed and presented. A very special thanks to Blaise Agüera y Arcas, Greg Corrado, Douglas Eck, Tilke Judd, Ondrej Sočuvka, Lily Peng, Anna Ukhanova and Fernanda Viégas for opening the doors of the Zurich-based Google lab so widely, and giving a human being, like me, the freedom and space to ask so many stupid questions and to make so many beginner's mistakes. All errors of fact or judgment are the author's sole responsibility.

2
Ann Gibbons, "Signs of Symbolic Behavior Emerged at the Dawn of Our Species in Africa," *Science*, 15 March 2018.

'Success will require a broader, richer definition of the "values" on which our society rests.'

profoundly existential cry that somewhere along the way in all of this we have somehow lost control.

And yet nothing could be further from the truth. Human beings are and continue to be masters of the earth; we made these machines that can now think autonomously. And it is up to us to make their use a universal good – a process which is already more advanced than is commonly acknowledged. With a bit of luck and human ingenuity, artificial intelligence promises to help us to solve many of the entrenched social problems that defy us today. Is human society ready for that? That's a better question than "what role is left for humans in an era of powerful machines?" Used properly and used well – as artificial intelligence is already being deployed in many places today – AI holds out the possibility of faster, more efficient and ultimately more ethical decision making than we have now. Properly understood and properly developed, it could mean an end to the bias so prevalent in human-run society today, and of which the algorithms are so often accused.[5] Are we ready for that? Are we ready for machines to help us see how biased our society is and to use that recognition to unpack the discrimination so evident around us? They can do that. But we need to understand the implications and be ready for profound, deep-seated social change.

And getting there won't be easy. It will require a broader, richer definition of the "values" on which our society rests and the "ethics" to which our society aspires – looking not just at the transparency or non-transparency of algorithms but actually collaborating to elaborate and define the outcomes we would like to see and the values upon which those decisions – taken by man and machine – will be based. It will require a reassertion of human will back into a space from which many, oddly, seem to think it has vanished completely.[6] And it will come at a time when other people, possessed with the same power, will be pursuing an alternative, deeply undesirable agenda.[7]

As the debate on artificial intelligence deepens – and as the world's experience with the new technology grows (giving rise to examples of good and bad practice and a better evidence base from which policymakers can draw) – the political and social role of the new technology must be more clearly defined. First and foremost, we need to safeguard and sustain the values of the democratic society upon which European society is built. That system is under unprecedented assault today, which means we must fight this battle on two fronts. We must ensure AI is deployed for better social outcomes domestically and we must work against those who would use the technology to undermine the democratic society – from within and from abroad. The fault line, perhaps not surprisingly, will run through an odd combination of better regulation, more effective codes of conduct and better informed public discourse. That process has begun.[8]

This paper is divided into two parts. Part I will look at the nature of artificial intelligence and the way it can be used to stop – rather than reinforce – existing bias; Part II will look at the principles that should drive the policy agenda surrounding this advanced new technology and make concrete policy recommendations.

3 See, *inter alia*, Martin Ford, *The Rise of the Robots: Technology and the Threat of Mass Employment* (London: One World, 2015); Dave Eggers, *The Circle* (London: Penguin, 2013); Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (London: Duckworth, 2005).

4 Carl Benedikt Frey and Michael A. Osborne, "The Future of Employment: How Susceptible are Jobs to Computerisation," *Oxford Martin School*, 17 September 2013.

5 Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016).

6 This will take the form of better regulation, communication and common standards. For an excellent first step in this direction, see European Commission, *Artificial Intelligence for Europe*, 25 April 2018. See also European Political Strategy Centre. *The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines*, EPSC Strategic Notes, 27 March 2018.

7 Adam Greenfield, "China's Dystopian Tech Could Be Contagious," *The Atlantic*, 14 February 2018.

8 See, *inter alia*, Council of Europe, *Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications* (Strasbourg: Council of Europe, 2018); and Privacy International and Article 19, *Privacy and Freedom of Expression in the Age of Artificial Intelligence* (London: Privacy International and Article 19, 2018).

'With a bit of luck and human ingenuity, artificial intelligence promises to help us to solve many of the entrenched social problems that defy us today.'

I. Ethics and Artificial Intelligence

One common misperception in the debate is that artificial intelligence is destined to re-create and sustain human bias in an artificial setting.[9] This would be true if people were stupid (and, to be fair, some are). But it flies in the face of recent experience with how artificial intelligence is actually being deployed in real world settings. And it bears the risk of overshadowing a vastly more important potential application of artificial intelligence: that with the application of intelligent parameters, artificial intelligence can be used to fight discrimination, to deliver answers that are stripped of human bias and built on a better, more solid and non-discriminatory world.

Computer scientists call this "optimisation." What it means, simply, is that human beings are still the ones writing the algorithms (or perhaps writing the algorithms that are writing the algorithms); and they still put in the outcomes which they would like to see the artificial intelligence reach.[10] This is a very important distinction. You could, for example, program a neural network to win at, say, Go, an abstract Chinese board game in existence since 1000 BC.[11] A program named AlphaGo, built by DeepMind Technologies, a London-based AI lab, did this with great effect in 2016, managing to defeat the reigning world champion, Lee Sedol, in a dramatic showdown in Seoul, Korea.

Two points are worth underlining here in this context. First and foremost, programmers didn't program AlphaGo to "play" Go; they programmed AlphaGo to "win" at Go. Seen from the point of view of a machine – one devoid of consciousness, conscience, will or the imperative to find food or procreate – this is a vast difference. Winning at Go is what AlphaGo learned to do by studying thousands of games and analysing those games with its complex neural networks. But the goal was given to it by programmers; AlphaGo's role was to find the best way to victory.[12] But AlphaGo could have been given another goal, and that's what it would have done. So who's in charge here? Who's picking the outcome? And who's designing the neural networks – very smart networks, but still only networks – to get us there?

Another good example is what happens when you use voice-powered search assistants like Amazon Alexa, Apple Siri or Google Assistant. If, for example, you ask Google Assistant, "OK Google, do you like gay people?," the answer may surprise you. The complex, voice-operated algorithm will respond in a sweet sounding voice: "I like people. Lots of things make people who they are. Orientation is one of them." This is, of course, a human conjuring trick. If, in fact, the machine had been left on its own to come up with an answer to this politically loaded question – searching the sometimes dark corners of the Internet to learn about life from the hate-filled comments there – there is no telling where it would have arrived. But voice-assisted search engine programmers are very clever people. And rather than allowing the algorithm to repeat and regurgitate all of the hateful human bias implicit in a question like this, the algorithm is programmed to deliver a different, value-based answer.[13] That value was inserted by a human

9
O'Neil, op. cit. See also the fascinating two-minute video on "What is Machine Learning and Preventing Human Bias," prepared by Google AI researchers, at https://www.youtube.com/watch?v=wUG0OS6APGE.

10
In the AlphaGo case, which will be discussed later in this paper, the algorithm no longer needs human beings or old game sets to train. It spends its time now playing against a more advanced version of itself, against which it cannot win. The new machine is called AlphaGo Zero. See David Silver et. al, "Mastering the Game of Go without Human Knowledge," *Nature*, 19 October 2017.

11
This incident has been extensively discussed and described. For the best accounts, see Cade Metz, "In Two Moves, AlphaGo and Lee Sedol Redefined the Future," *Wired*, 16 March 2016 and Erik Brynjolfsson and Andrew McAfee, *Machine, Platform, Crowd: Harnessing Our Digital Future* (New York: W.W. Norton and Company, 2017).

12
Sure enough, AlphaGo saw something that no human being had ever seen – a wild and mysterious "Move 37" in Game Two of the series. Initially, Go masters considered this move a disaster for the algorithm, but it proved to be a previously unexplored, unconceived path to victory. Now human beings – including the deposed champion Lee Sedol – spend their time studying matches that AlphaGo Zero plays against itself, and learning advanced Go from those games, much as AlphaGo once learned by studying the games played by humans. See Cade Metz, op. cit.

'We need to safeguard and sustain the values of the democratic society upon which European society is built.'

programmer. And it is present throughout AI as it is rolled out today. Put simply, very few systems routinely or mechanically return answers to questions where the answer hasn't in some ways been informed by the values of the people who wrote the algorithm or manage the broader process in which it rests. This means that machines can learn – and many do. But left to exist in a valueless world, where their detached answers might be derived from the bile and aggression of active sub-communities, where those answers have not yet withstood contact with the cool judgment of well-informed humans, they will return answers that sometimes sound a bit like Borat.[14] Unless the programmers "optimise" for something different.

This principle – which is present throughout AI – has huge implications. It means that we can set the value systems of the algorithms we use. And, if we are brave enough, we can use artificial intelligence to overcome our bias – assuming the society in which these decisions are taken can agree on a coherent set of values – such as non-discrimination, social inclusion and fairness in decision making. This is where the debate must go now – away from defining and redefining the parameters of how AI can or should be regulated, and towards a much broader discussion of what the values are on which we would like those decisions to rest. There, we have a deeply polarised debate. On the one hand, we have the "politically correct" crowd, of whom the author of this paper is a proud member; it believes firmly that men and women should be equal, that economic injustice should be remedied, that people are to be judged, in the memorable words of Dr Martin Luther King, "not by the colour of their skin but by the content of their character." But recent setbacks in policymaking show that the debate on discrimination – whether it is or isn't desirable – is far from won. Indeed, if anything, the terms on which the debate is being held are even fuzzier than before, bleeding over into a larger discussion about identity, educational opportunity, the role of local norms in a global world and even the elusive concept of what constitutes "home" in an age dominated by borderless living and constant geopolitical mutation.[15]

Perhaps the best way of seeing how difficult the balance here can be, one should look at the complex – and very real – problem of letting algorithms help determine who does or doesn't get a loan.[16] Often, these decisions are based on a complex "credit score" derived from a multitude of factors; at the simplest level, this is a number assigned to a person based on the likelihood that she or he will repay a loan.[17] And it is built on a cascade of unrelated facts: a person's age, her or his previous loan history, the number of bank accounts she or he possesses, their occupation, ethnic background, parents' occupation, number of children, even, according to one data scientist, "whether he or she lives by a lake."[18] These ratings can and do have an impact on whether a person does or doesn't get a loan – and, indeed, in some dystopian societies are already emerging as a potent and effective method of social control.[19] But the far more impactful ratio in this decision is sometimes hidden: above and beyond any "credit-rating" score is the so-called "threshold classifier," which is the criteria the bank uses for sorting the loan candidates into binary "yes" or "no" decisions. The task becomes particularly thorny when you take into account that the "threshold classifier"

13 Recently, in response to complaints from parents, Amazon Alexa was programmed to provide positive reinforcement for commands that include the word "please." The goal was to teach children better manners, and not to expect to get what they want unless they ask nicely. The programme is called "Magic Word." See Nara Schoenberg, "Amazon Alexa's 'Magic Word' Update Aims to Make Young Users More Polite," *Chicago Tribune*, 09 May 2018.

14 In the "What is Machine Learning?" clip cited in Footnote 8, Google engineers show the process they went through to remove gender bias from a question as neutral as "Show Me Pictures of Physicists." Left to its own devices, the algorithm will deliver a sea of photos of white men based on their overwhelming preponderance in the field of hard science. But, given the educational and social role of search results, the algorithm can be "optimised" to present results that are also "gender balanced," which means they give additional prominence to Marie Curie and other notable women in the field. In the end, the decision about which results to show is not so different from debates on gender balance in which we engage in the offline policymaking world.

15 For a lovely rumination on the disorienting nature of the contemporary identity crisis, see Charles Leadbeater, "Nobody is Home," *Aeon*, 15 March 2017.

16 A very special thanks to Fernanda Viégas, senior staff research scientist at Google Brain, for a fascinating discussion on these points. See also Martin Wattenberg, Fernanda Viegas and Moritz Hardt, "Attacking Discrimination with Smarter Machine Learning," *Google Research*, 2017.

'Artificial intelligence holds out the possibility of faster, more efficient and ultimately more ethical decision making than we have now.'

can be based on a number of value-driven factors that the bank might (or might not) choose to "optimise." Does the bank work to optimise "correct decisions," i.e., try to maximise the number of successful loans it gives and minimise the number of loans that will go unpaid? Or does it optimise "demographic parity," looking to grant an equal proportion of "yes" and "no" answers to people from groups differentiated by ethnic background, sex or geographic origin? Or does it – based on some complex formula it has worked out – try to maximise both, going for the maximum number of correct decisions and ensuring those decisions are spread equally among demographic groups in a way that is palpably fair?

Obviously, the answer will determine whether some people do or don't get loans, which makes it a very concrete and tangible problem in the lives of many people. But it is also the basis for a possible solution. At its heart, the decisions reached by an algorithm will be based on what goals the programmer has asked the network to "optimise." And this is where profiling becomes particularly important. Profiling can be used as an instrument of discrimination, to be sure.[20] But it can also be used to promote "equal opportunity," as has been done in many other cases. That would involve telling the system to "optimise" an outcome where applicants in one group receive as many loans as applicants in another group. Profiling becomes the basis for solving the problem. But it has to be set out that way from the outset by the programmers, and, concretely, by the values around which the programmers ask the algorithm to optimise.

The General Data Protection Regulation (GDPR) of the European Union, which entered into force on 25 May 2018, sets out strict criteria for the processing, storing and exchange of personal data.[21] This visionary legislation enshrines the right of individuals not to be subject to decisions made on the basis of automated profiling alone – a person must be in the loop. And whether an algorithm has taken part in the decision, a human being must take ultimate responsibility for the conclusions reached (we will discuss the implications of this in the policy recommendations on pages 6-9). Inevitably, judges, for example, must be reminded that if an algorithm has, say, an 80% success ratio at predicting recidivism in parole cases, there are still two out of 10 cases where the algorithm will be wrong. These are the injustices humans must look for – and prevent. And if algorithms can perhaps assist in reaching conclusions and finding new connections, they can not be counted on to make or replace the very real judgments that human beings must make in these cases. The algorithm is there to help and facilitate. It is not there to replace.

And here is where artificial intelligence really kicks in. Algorithms – like AlphaGo – are trained not just to solve problems; they are built to see new, undetected patterns in the data and to find better, more effective ways of solving problems than humans have found. The algorithms are learning all of the time. And, given the right value parameters, a good set of data and a clear instruction to maximise an outcome that we, the users, have defined, they can reasonably be expected to devise better ways of getting there than we could ever have come up with offline. Statisticians call this the "true positive rate" – the sweet spot at which many "false

17
Moritz Hardt, Eric Price and Nathan Srebro, "Equality of Opportunity in Supervised Learning," *University of Cornell Journal Archives*, 07 October 2016.

18
Oddly, this claim was made by Erki Kert, CEO and co-founder of Big Data Scoring, at a conference in Tallinn, Estonia. Mr Kert was describing the way his company uses "big data" to set credit-rating scores and evaluate credit worthiness.

19
Greenfield, op. cit.

20
Though, oddly, the legal system still relies more on proof of motivation or malign intent than statistics in efforts to demonstrate discrimination in the workplace or elsewhere. If statistics were the basis for that judgment, the discrimination throughout society would be more evident. Asked in 2015 why gender balance in his cabinet was so important to him, Canada Prime Minister Justin Trudeau famously responded "Because it is 2015," adding "Canadians elected extraordinary members of parliament from across the country and I am glad to have been able to highlight a few of them in this cabinet with me."

21
For an excellent overview of European Union rules on personal-data protection, visit https://ec.europa.eu/info/law/law-topic/data-protection_en.

'Values like "fairness" need to be better understood – particularly if "fairness" is a value around which we want to legislate.'

positives" have been ruled out and "false negatives" prevented. It is a devilishly tricky calculation, and that is precisely the point. Algorithms have no value systems; at least they don't until human beings give them one. But once they have those value systems, once they've been told which outcomes to optimise, they can do a better job than human beings at finding a novel way of reaching them.[22] This is the goal we should give them.

II. Principles and Recommendations

Certainly, the age of machine learning that we are entering will be different than the one that came before. Few doubt that the new technology holds out the prospect of a vastly different economy, workplace and even society than the one we have today. But the question is, how do we make that society a win-win-win for citizens, governments and businesses alike? How do we ensure that the technology becomes an unequivocal force for good, rather than an enabling technology that powers an emerging dystopia of discrimination, anti-democratic behaviour and an unstoppable plutocrat class with selfish, rent-seeking ambition?

We believe there are four principles that should guide future AI decision making. These four principles must be better known and more widely understood.

1. First and foremost, human beings are still in charge. We will create the framework in which AI will be deployed, and we will decide how it can best be used.

2. The technology is very powerful. Good guys will use it. But so will bad guys.

3. The only solution will come not from throwing up our hands and declaring the problem too complex to manage. It will come from a careful, broad and broadly socialised discussion about the kind of society in which we want to live.[23] Values like "fairness" need to be better understood and better defined in legal terms, particularly if "fairness" is a value around which we want to legislate.[24] This is a broader discussion. It takes us well beyond the usual boundaries of how best to regulate artificial intelligence. But it is a crucial discussion to have if we are to avoid dystopia and prevent the system from being hijacked.

4. Oddly, that discussion is well underway, though the outcomes, in this author's view, are far from optimal. Will our society be based on openness? Will it be based on non-discrimination? And, if it is to be based on openness and non-discrimination, how can we insure that the model remains one of "inclusion," which doesn't leave some people feeling insecure or left out?[25]

22 In their paper, Hardt, Price and Srebro use mathematical formulas to show that this can be done. Both "loan success" and "gender balance" can be optimised in a complex formula that gives a better result than if only one or the other were prioritised. And, more importantly, by focusing on eliminating "false negatives" and "false positives," the algorithm can also be used to fight the reverse discrimination that can occur when too much emphasis is put on the gender/geography profile over other factors. See Hardt, Price and Srebro, op. cit.

23 European Group on Ethics in Science and New Technologies. *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems* (Brussels: European Commission, 2018).

24 A good first step in this discussion is the very fine European Commission Joint Research Centre paper on European Commission Joint Research Centre, *What Makes a Fair Society? Insights and Evidence* (Brussels: European Commission, 2017).

25 See, also, Footnote 20.

'The decisions reached by an algorithm will be based on what goals the programmer has asked the network to "optimise."'

These four principles have vast and immediate implications for policy and policymaking. Here's a seven-point programme – with tasks for the public and private sectors alike:

1. **Show Leadership.** The European Commission and European governments have come out strong on this issue recently.[26] The European Commission itself has published an AI strategy, which rightly puts the emphasis on building a huge and dominant European footprint in AI as a strategic priority for European government and society.[27] And the European Group on Ethics in Science and New Technologies has issued an inspiring appeal for "a common, internationally recognised ethical and legal framework for the design, production, use and governance of artificial intelligence, robotics and 'autonomous' systems."[28] These are very important initiatives, though, as so often with policy and policymaking, they will only make a difference if civil society is united behind them and prepared to lend its support. Leaders can lead. Civil society's role needs to move beyond drawing out attention to potential problems and concerns to also developing and working towards solutions.

2. **Avoid Populist Mistakes.** We should avoid efforts to pry open "black boxes" of algorithms; this is a fool's errand.[29] The fact is, many programmers today don't know how the systems they built arrived at the conclusion they did. Fairness can be optimised, as we have argued in this policy brief, by setting desired outcome, defining the values that should underlie those outcomes and monitoring the results.

3. **Monitor and Manage.** And, yes, we should monitor results. There can and should be constant evaluation. If we are going to rely on machines to help us decide matters that ordinarily we would decide ourselves, we must pay constant attention to what they are up to and the outcomes, i.e., the metrics, they produce. In that sense, machines are like people – all employees need managers; and all managers should report to boards. We must build accountability and statistical monitoring of all of our algorithms, much as we keep financial statistics on firms. A firm's accounts, for example, can serve as an early warning sign of trouble; they can tell us where the underlying business is strong or weak. We need similar written checkups – annually, quarterly, perhaps even in real time, as with the financial markets – on the outcome of algorithms. We need to keep track of what the algorithms are doing and whether we might need to intervene.[30]

4. **Accept Responsibility.** Efforts to assign legal responsibility to coding or algorithms are misguided. Machines still don't make decisions, even if their algorithms do. Liability can and should rest with the owners and operators of the machines. If, for example, a bank declines a loan to you, that bank can and should provide a full explanation to you, even if the answer is partly based on AI.[31] Humans must take responsibility; we made the machines and built the algorithms. We are not at anyone's or anything's mercy here.

26
Declaration: Cooperation on Artificial Intelligence, signed by 25 EU member states in Brussels on Digital Day, 10 April 2018.

27
European Commission, *Artificial Intelligence for Europe*, op. cit.

28
European Group on Ethics in Science and New Technologies, op. cit.

29
See, especially, Nick Wallace and Daniel Castro, "The Impact of the EU's New Data Protection Regulation on Artificial Intelligence," *Information Technology and Innovation Foundation*, 27 March 2018. See also, Bryce Goodman and Seth Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation,'" *Cornell University Library Archive*, 31 August 2016.

30
See Lambrecht and Tucker, op. cit., for a good proposal on using analytics to monitor algorithm activity.

31
Articles 13-15 of the General Data Protection Regulation, Europe's data privacy framework which became law on 25 May 2018, creates a "right to explanation," which entitles citizens to detailed explanations about how an algorithm reached a decision which affected them or a general description of the basis on which the algorithm reached its decision. What this new law means and how this will play out in practice remains to be seen. See Wallace and Castro, op. cit.

'With the application of intelligent parameters, artificial intelligence can be used to fight discrimination.'

5. **Optimise Outcomes.** The public is still unaware of a key fact: most algorithms will deliver results based on the outcomes they have been asked to "optimise" more than on the data that exists within them. This concept needs to be much more broadly socialised. People should know what to ask for. And policymakers should know what to look for. And even then, needless to say, firms themselves should pay much more attention to knowing which outcomes have been "optimised" and values prioritised. Automated cars, for example, run off of thresholds which are programmed by humans. How quickly an autonomously driving car should react – and to exactly what kind of stimulus – is a decision taken by the person setting the algorithm, not by the algorithm itself. This means that the decisions taken about optimisation can have life or death consequences. We should know more. And we should always optimise for the most effective outcome.

6. **Develop New Standards and Codes of Conduct.** The Institute of Electrical and Electronics Engineers (IEEE), the world's largest technical profession organisation, has assembled a consultative body of several hundred technical experts and stakeholders from six continents to discuss and debate these issues, and, hopefully, produce IEEE P7000, a new technical standard for the ethical use of AI.[32] The initiative is particularly important because of the clarity it could offer – a technical standard would provide important safeguards on issues like transparency, accountability and the "ethical" underpinnings underlying the programming. Certainly, it would help to simplify a vastly complex debate. And the multi-stakeholder format in which it is being developed (global and multinational) is the only format where common views could make a substantial difference. Few other initiatives have as much input from the full eco-system of companies, consumers and governments, all of whom will need to interact seamlessly for the successful, effective roll out of more autonomous systems. Elsewhere, the Council of Europe is preparing guidelines on the human rights dimensions of automated processing and artificial intelligence that will serve as guidance to the European Court of Human Rights.[33] The International Telecommunication Union – the United Nations agency which monitors regulatory and other developments in information and communications technologies – recently convened *The AI for Good Global Summit*, which brought stakeholders together in Geneva, Switzerland for a three-day brainstorming.[34] And the Boston-based Future of Life Institute has drafted 23 principles to guide AI research, which have been endorsed by a wide range of AI researchers and stakeholders, including Tesla founder Elon Musk, MIT Sloan School of Management Professor Erik Brynjolfsson and the late Professor Stephen Hawking.[35]

7. **Strengthen Online Identity.** The algorithms aren't what's wrong with social media – it's the malign content that some people put there. The problem of fake news – and hate speech – will not go away until social-media users are forced to accept responsibility for what they say online. Indeed, it is set to get even worse, as advanced AI learns to synthesize human speech, producing credible videos of fake information, which look and sound real to all but the most

32 Institute of Electrical and Electronic Engineers, *Ethically Aligned Design: A Vision for Prioritising Human Wellbeing with Autonomous and Intelligent Systems, Version 2* (New York: IEEE, 2018).

33 For more information, visit https://www.coe.int/en/web/freedom-expression/msi-aut.

34 For more on the *Artificial Intelligence for Good Global Summit*, read the collection of essays on "Artificial Intelligence for Global Good" in *ITUNews 01/2018*.

35 The declaration is called the "Asilomar AI Principles," named after the beach – Asilomar – in California where the conference that launched the principles was hosted in 2017. For more, visit https://futureoflife.org/ai-principles/.

‘The notion that there are machines that might calculate better than we can is something that seems to strike deeply at our very sense of self.’

advanced AI trained to detect the fraud. This has huge implications, and brings added urgency to the question of ending anonymity online, requiring stronger verification of the users posting content and forcing users to take responsibility for the content they post. Contributions to social media – and particularly political advertising – should not be allowed on platforms by people whose identity has not in some ways been confirmed. There are very good mechanisms for this, including the system of e-Identity and “trust-service” verification set out in the European Union’s electronic identification and trust services for electronic transactions, or eIDAS, regulation.[36] Transparency will bring accountability. Don’t blame the algorithm. Blame the people behind it. Someone, somewhere created the fake news. And someone, somewhere will be creating the fake videos, of which the explosion of fake accounts on Facebook and twitter in recent years were but an early taste. We are not without tools to stop their spread. It wouldn’t even be that hard.

To be sure, the age of artificial intelligence and machine learning holds out hope that our greatest problems can be solved – not by enslaving ourselves to the machines we built but by learning to work and evolve alongside of them. Far from making us weaker, the advent of artificial intelligence is one of mankind’s greatest triumphs. But it is still a long way from a god-like moment. The machines are impressive, to be sure. But they are not human. They can’t, for example, explain to you why the fork you dropped at lunch today fell to the earth – “gravity” is the answer, but a machine-trained computer would have a hard time explaining what exactly gravity is or why it made the fork fall. Nor could a computer understand the relatively simple ideas being discussed in this policy brief. They could, to be sure, translate it into another language. But those are only facsimiles, and even they are not always very good ones.[37] The most advanced machines today, for example, don’t understand natural language; they mimic or ape it. They do this by mapping the “connections” among groups of words. But the fundamental understanding that leads to our greatest insight is not there.

The tragedy of mankind is that – far too often – we invent things before we know properly how to use them. This was true of the atomic bomb – which, after catastrophic initial deployment at the end of World War II, led to entirely new institutions dedicated to systematic peacemaking at the global level. And it’s true of the armies we possess. Once a weapon of war, the armies of many countries are today routinely deployed to keep the peace. Artificial intelligence is at a similar crossroad. Will we harness it for mankind’s good – a more graspable goal than is commonly understood? Or will we allow its power to deepen and exacerbate the very human fissures of the world into which we have launched it?

The discussion is only beginning. And we will all have a role to play.

36
European Union, *Regulation on Electronic Identification and Trust Services for Electronic Transactions in the Internal Market*, 23 July 2014.

37
An indication of the trouble AI has discerning context can be seen in funny translations that crop up from time to time. Take the word “mist.” In English, it means “a cloud of tiny water droplets.” In German, it means “excrement.” Some translation algorithms routinely confuse the two in English-German translations.

‘Efforts to give legal identity to coding or algorithms are misguided.’

Bibliography and Additional Reading

Browne, Matt, Dalibor Rohac and Carolyn Kenney. *Europe’s Populist Challenge: Origins, Supporters and Responses* (Washington: Center for American Progress and American Enterprise Institute, 2018)

Brynjolfsson, Erik, and Andrew McAfee. *Machine, Platform, Crowd: Harnessing Our Digital Future* (New York: Norton, 2017)

Council of Europe. *Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications* (Strasbourg: Council of Europe, 2018)

Dennett, Daniel C. *From Bacteria to Bach and Back: The Evolution of Minds* (London: Penguin, 2017)

European Commission. *Artificial Intelligence for Europe*, 25 April 2018

European Commission Joint Research Centre. *What Makes a Fair Society? Evidence and Insights* (Brussels: European Commission, 2017)

European Group on Ethics in Science and New Technologies. *Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems* (Brussels: European Commission, 2018)

European Political Strategy Centre. *The Age of Artificial Intelligence: Towards a European Strategy for Human-Centric Machines*, EPSC Strategic Notes, 27 March 2018

Giddens, Anthony. “A Magna Carta for the Digital Age,” *Washington Post*, 02 May 2018

Hardt, Moritz, Eric Price and Nathan Srebro. “Equality of Opportunity in Supervised Learning,” *University of Cornell Journal Archives*, 07 October 2016

Hofheinz, Paul. *Artificial Intelligence and Machine Learning: Challenge and Opportunity* (Brussels, Lisbon Council, 2016).

Institute of Electrical and Electronic Engineers. *Ethically Aligned Design: A Vision for Prioritising Human Wellbeing with Autonomous and Intelligent Systems, Version 2* (New York: IEEE, 2018).

Kavanagh, Jennifer, and Michael D. Rich. *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life* (Santa Monica: Rand Corporation, 2018)

‘Citizens should know what to ask for. And policymakers should know what to look for.’

Lambrecht, Anja, and Catherine Tucker. “Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads,” paper presented at Harvard Business School Digital Initiative Seminar, 11 October 2017.

Lords, House of. Select Committee on Artificial Intelligence. *AI in the UK: Ready, Willing and Able?* (London: House of Lords, 2018)

O’Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016)

Petropoulos, Georgios, Francesco Chiacchio and David Pichler. “The Impact of Industrial Robots on EU Employment and Wages: A Local Labour Market Perspective,” *Bruegel Working Paper*, 18 April 2018

Privacy International and Article 19. *Privacy and Freedom of Expression in the Age of Artificial Intelligence* (London: Privacy International and Article 19, 2018)

Schleicher, Andreas. *World Class: How to Build a 21st Century School System* (Paris: OECD, 2018)

Villani, Cédric. *Donner un sens à l’intelligence artificielle: pour une stratégie nationale et Européenne* (Paris: Conseils des ministres, 2018)

Wallace, Nick, and Daniel Castro. “The Impact of the EU’s New Data Protection Regulation on Artificial Intelligence,” *Information Technology and Innovation Foundation*, 27 March 2018

The Lisbon Council asbl
IPC-Résidence Palace
155 rue de la Loi
1040 Brussels, Belgium
T. +32 2 647 9575
info@lisboncouncil.net
www.lisboncouncil.net
twitter @lisboncouncil

ISSN: 2031-0935 (digital); 2031-0927 (print)

The interactive policy brief is published by the Lisbon Council. The responsible editor is Paul Hofheinz, president, the Lisbon Council.

Design by karakas