FATEN: A framework for governance in the era of data-driven decision-making algorithms[1]

Nuria Oliver, PhD
nuria@alum.mit.edu

Massive streams of human behavioural data, combined with increased technical and analytical capabilities (in particular, data-driven machine-learning methods), are enabling today's companies, governments and other public sector actors to use data-driven machine learning-based algorithms to tackle complex policy problems (Willson 2016). Decisions with both individual and collective impact that were previously taken by humans – often experts – are nowadays taken by data-driven artificial intelligence systems (i.e. algorithms), including decisions regarding the hiring of people, the granting of credits and loans, judicial judgements, policing, resource allocation, medical diagnoses and treatments, and the purchase/sale of shares in the stock market. Data-driven algorithms have the potential to improve our decision making. History has shown that human decisions are not perfect – they are subject to conflicts of interest, corruption, selfishness/greed and cognitive biases, which result in unfair and/or inefficient processes and outcomes (Fiske 1998). The interest in the use of algorithms can therefore be seen as the result of a demand for greater objectivity in decision making and for a better understanding of our individual and collective behaviours and needs.

Data-driven algorithmic decision making may indeed enhance overall government efficiency and public service delivery by optimising bureaucratic processes, providing real-time feedback and predicting outcomes (Sunstein 2012). In his recent book, *Technocracy in America*, Parag Khanna argues that a data-driven direct technocracy would be a superior alternative to today's representative democracy, because it could dynamically capture the specific needs of the people while avoiding the distortions of elected representatives and corrupt middlemen (Khanna 2017).

The potential for data-driven algorithmic decision making to make a positive impact in the world is massive and certainly motivates my work on this area (e.g. Froelich et al. 2009, Bogomolov et al. 2014, Torres Fernández et al. 2014, Lepri et al. 2017). Numerous efforts worldwide are also exploring this potential, including the New Deal on Data led by Alex Pentland at the World Economic Forum, which is focused on consensus policies and initiatives to give citizens control over the possession, use and distribution of their personal data; NGOs such as the Partnership on AI, Data-Pop Alliance (where I am Chief Data Scientist) and Flowminder, which are focused on leveraging large-scale data and machine-learning techniques for social good in areas such as financial inclusion, public health and climate change/natural disasters; United Nations initiatives including the World Data Forum, the Global Partnership for Sustainable Development Data and Global Pulse; OPAL, a project led by Data-Pop Alliance with the goal of taking advantage of big data and artificial intelligence for social good while preserving the privacy of people, in a sustainable, scalable, stable and commercially viable way; private sector initiatives in telecommunication companies or banks; the GSMA Big Mobile Data for Social Good initiative, led by the GSMA and the United Nations Foundation, in which 20 mobile operators participate to contribute through the analysis of aggregate and anonymous mobile data to solve problems in the areas of public health and climate change /natural disasters; and the AI for Good Global Summit of the ITU, an international summit of United Nations for the dialogue on artificial intelligence, aimed at identifying the practical applications of AI for the improvement of the sustainability of the planet. The latter is managed by the International Telecommunications Union (ITU) as a specialised agency of United Nations for information and communication technologies.

[1] Cite as "Governance in the Era of Data-driven Decision-making Algorithms", published in "Women Shaping Global Economic Governance", CEPR Press, pages 171-180, July 2019

However, data-driven decision making is not without limitations. Plato's words of some 2,400 years ago are surprisingly relevant today: *"A good decision is based on knowledge, not on numbers".*

Algorithmic decision making for public policymaking may generate inefficiencies and negative consequences (Easterly 2014). Turning to the use, governance and deployment of algorithmic and data-driven approaches in the public sector, we can draw several parallels with the 'tyranny of data' or the 'tyranny of algorithms' (Lepri et al. 2017).

At the heart of these issues is the fact that technology outpaces policy in most cases; mechanisms for the governance of algorithms have not kept pace with technological development.

When algorithmic decisions affect thousands or millions of people, important ethical dilemmas arise. For example, does this mean that automatic decisions are beyond our control? What level of security do these systems have to protect themselves from cyberattacks or malicious use? How can we guarantee that the decisions and/or actions do not have negative consequences for people? Who is responsible for these decisions? What will happen when an algorithm knows each one of us better than we know ourselves and can take advantage of that knowledge to manipulate our behaviour subliminally?

Beyond preserving human rights, the existing literature has proposed a set of ethical principles and working dimensions that will need to be addressed to ensure that data-driven decision making has a positive impact on society. I summarise these principles using the acronym 'FATEN' (Oliver 2018):

2. F is for Fairness

Fairness and non-discrimination should be central elements in the development of automatic decision-making (and action) systems based on artificial intelligence. Decisions based on algorithms can discriminate (Barocas and Selbst 2016) because the data used to train the algorithms might have biases that can give rise to discriminatory decisions, because of the properties of an algorithm itself, or through the misuse of certain models in different contexts. Algorithmic decisions can reproduce and magnify patterns of discrimination due to decision makers' prejudices or reflect biases already present in society (Pager and Shepherd 2008). A recent study by ProPublica of the COMPAS Recidivism Algorithm (used to inform criminal sentencing decisions by predicting recidivism) found that the algorithm was significantly more likely to label black defendants than white defendants, despite similar overall rates of prediction accuracy between the two groups (Angwin 2016). Along these lines, in her book, *Weapons of Math Destruction*, Cathy O'Neil details several case studies on harm and risks to public accountability associated with data-driven algorithmic decision making, particularly in the areas of criminal justice and education (O'Neil 2016). In addition, data-driven algorithmic decision-making processes may result in opportunities being denied to people not due to their own actions, but to the actions of others with whom they share certain characteristics. For example, some credit card companies have reduced the credit limit of their customers as a result of analysing the behaviour of other customers with a history of poor payments who made purchases in the same establishments as the customers concerned. Although various solutions have been proposed in the literature to deal with algorithmic discrimination and to maximise justice, I would like to underline the urgency for experts from across various disciplines (including law, economics, ethics, computer science, philosophy and political science) to create, evaluate and validate different metrics of justice for different tasks in the real world. In addition to this empirical research, a framework of theoretical modelling is needed – supported by empirical evidence – that helps the users of these algorithms ensure that decisions are made as fairly as possible.

To the principle of fairness I would also like to add a principle of *cooperation***.** Due to the transversal nature of data-driven algorithms and their potential application to all areas, a constructive exchange of resources and knowledge between the private, public and social sectors should be encouraged and developed to achieve their maximum potential of application and competitiveness. This need for cooperation not only between different sectors but also between nations – given today's globalisation – has been emphasised by the well-known Israeli historian and thinker, Yuval Noah Harari (Harari 2018).

2. A is for Autonomy, Accountability and intelligence Augmentation

Autonomy is a central value in Western ethics according to which each person should have the ability to decide their own thoughts and actions, thus ensuring free choice and freedom of thought and action. Today, however, we can build computational models of our desires, needs, personalities and behaviour with the ability to influence our decisions and behaviour subliminally. Therefore, we should ensure that autonomous intelligent systems always preserve human autonomy and dignity. For this, the systems need to behave in accordance with accepted ethical principles of the society in which they are used. There are numerous institutes and research centres created for this purpose, such as the AI Now Institute at New York University and the Digital Ethics Lab at the University of Oxford. This is an active area of research, however, and there is no single recognised method for embedding ethical principles into data-driven algorithmic decision processes. It is also important to highlight that all developers and professionals working on the development of artificial intelligence systems that affect or interact with people (algorithms for decision making, recommendation and personalisation systems, chatbots, and so on) should behave in accordance with a clear code of conduct and ethics defined by the organisations where they work. As Roy E. Disney wisely said, "*It is not difficult to make decisions when you know your values."*

We also need to be clear about the *attribution of responsibility* for the consequences of the actions or decisions taken by autonomous systems, in the same way as happens with the rest of the products used in society. Transparency is often considered a fundamental factor in contributing to accountability, but transparency and audits are not enough to guarantee clear accountability. Computational methods can help provide clarity regarding the attribution of responsibility, as shown by Kroll (2015), even when some information is hidden.

Finally, it is constructive to have a synergistic vision between humans and data-driven decision-making systems. This is often referred to as 'intelligence *augmentation'*, as such systems are used to increase or complement human intelligence, not to replace it. For example, an internet search engine can be considered a system to increase our intelligence, since it expands our knowledge with the capacity to process billions of documents and find the most relevant ones; an algorithm to automatically detect tumours in X-rays augments the intelligence of oncologists and radiologists by providing better-than-human detection capabilities which humans can use to make more informed decisions regarding their diagnoses and prescribed treatments.

3. T is for Trust and Transparency

Trust is a basic pillar in human relationships with other humans or with institutions. Technology needs the trust of its users who delegate their lives to digital services. However, the technology sector is experiencing a loss of trust due to recent scandals, such as the Facebook / Cambridge Analytica or the Huawei scandals. In order to develop a trusting relationship, three conditions need to be met: (1) first of all, the *competence* regarding the specific task that the trust will be deposited onto; (2) secondly, *reliability*, that is, sustained competence over time; and (3) finally, *honesty* and *transparency*. Thus, the T is also for transparency.

Transparency here refers to the ability to understand a computational model, and it can be a mechanism that contributes to the attribution of responsibility for the consequences of the use of said model. A model is transparent if a person can observe and understand it easily, and this is not necessarily the case in algorithmic decision making (Zarsky 2016, Pasquale 2015).

Burrell (2016) proposes three different types of opacity (i.e. lack of transparency) in algorithmic decisions:

(1) **Intentional opacity**, the objective of which is the protection of the intellectual property of the inventors of the algorithms. This type of opacity could be mitigated with legislation that would force the use of open software systems. The new European General Data Protection Regulation (GDPR) with its right to an explanation is an example of such legislation. However, powerful commercial and governmental interests can make it difficult to eliminate this type of opacity.

(2) **Illiterate opacity**, which arises because the vast majority of people lack the technical skills to understand how algorithms and data-driven computational models work. This type of opacity would be attenuated with educational programmes in digital competences – as I have explained previously – and by allowing independent experts to advise those affected by data-driven algorithmic decision-making processes.

(3) **Intrinsic opacity**, which arises from the nature of certain machine-learning methods (for example, deep learning models; see LeCun et al. 2015). This opacity is well known in the machine-learning research community and is also referred to as the problem of interpretability.

It is essential that artificial intelligence systems are transparent not only in relation to what data they capture and analyse on human behaviour and for what purposes – which is contemplated in the GDPR at the European level – but also in relation to situations in which humans are interacting with artificial systems (e.g. chatbots) as opposed to with other humans.

4. E for Education, bEneficence and Equality

We need to invest in *education* at all levels, starting with compulsory education curricula by adding Computational Thinking as a core subject from primary school, coupled with an emphasis on nurturing our creativity and the abilities of our social and emotional intelligence. Education is also needed for our policymakers and politicians, for professionals – particularly those whose jobs will be affected by the development of artificial intelligence – and finally for citizens so they can make informed decisions. Computational thinking includes five core areas of knowledge: algorithms, programming, data, networks and hardware.

Moreover, the use of data-driven algorithmic decisions should always focus on their *beneficence*, that is, on maximising its positive impact on society with sustainability, veracity and diversity. We cannot forget that not every technological development implies progress. What we should strive for and what we should focus on is progress. Of course, we need then to define progress. From my perspective, progress entails an improvement in the quality of life of people –of all people, not just a few--, of the rest of living beings on our planet, and of our planet itself.

Sustainability

Technological progress in general, and artificial intelligence systems in particular, consumes significant amounts of energy, with a negative impact on the environment (Andrae 2017). Today's pervasive deep-learning techniques require high computing capabilities with prohibitive energy costs, especially if we consider the deployment of such systems on a large scale. It is increasingly important that technological development is aligned with the human responsibility to guarantee the basic conditions for life on our planet and to preserve the environment for future generations. At the same time, data-driven machine-learning algorithms will be key to enabling us to address some of the most important challenges in the context of the environment (climate change, the scarcity of resources, etc.) as well as allowing us to develop sustainable transportation (autonomous electric cars, for example) and more efficient and sustainable energy models (smart grids, for example).

Veracity

Today, it is possible for data-driven machine-learning algorithms to create synthetic content (text, photos, videos) that is indistinguishable from 'real' content. This has led to the emergence of 'fake news', which can define public opinion on important issues – such as who should be the next president of a country or whether or not a country should remain a member of the EU – to favour the interests of a minority that generates and disseminates such content. The *veracity* of both the data used to train machine-learning algorithms and the content we consume is therefore of utmost importance.

Diversity

Given the variety of use cases in which data-driven machine-learning algorithms can be applied, it is important to reflect on the frequent lack of diversity in the teams that create such systems, which tend to be composed of homogeneous groups of computer scientists. In future, we should ensure that teams are diverse both in terms of their areas of knowledge and their demographics (in particular gender, given that women occupy fewer than 20% of technical positions in most technology companies).

Likewise, personalisation and recommendation algorithms often suffer from lack of diversity in their results and tend to pigeonhole their users based on certain patterns of tastes, which gives rise a 'filter bubble' (Pariser 2012). This lack of diversity in personalisation/recommendation is undesirable as it limits the opportunities of

technology to help us discover 'content' – be it movies, books, music, news or even friends – that differs from our own tastes and that therefore would help us understand other points of view and encourage open-mindedness.

Finally, we need to reflect about *equality*. The spirit of equality and solidarity perhaps is disappearing in the 4th Industrial Revolution. The development and growth of the Internet and the World Wide Web have been undoubtedly instrumental to enable the democratization in the access to information. However, the principles of universal access to knowledge and technology are questioned today partly due to the dominance of the technology giants in the US (Alphabet/Google, Amazon, Apple, Facebook, Microsoft) and in China (Tencent, Alibaba, Baidu) which have been coined as a "winner takes all" phenomenon. Together, these technology companies have a market value of more than 5 trillion USD and market shares in the US of more than 90% in Internet search (Google), more than 70% in social networking (Facebook) and about 50% in e-commerce (Amazon).

In fact, the XXI century is characterized by a polarization in the accumulation of wealth. According to a study by Credit Suisse, the 1% richest in the planet owns half of the world's wealth and the 100 richest people own more than the poorest 4 billion people. This situation of wealth concentration in the hands of very few has been attributed, at least partially, to the 4th Industrial Revolution and technological development that has led to it.

We see an evolution in the sources of wealth over time. Starting with the Agrarian Revolution in the Neolithic and during thousands of years, the ownership of the land entailed wealth. The First Industrial Revolution in the XVIII and XIX centuries in Europe and the US changed the meaning of wealth, which became associated with the ownership of factories and machines. Today, we could argue that the data –and more importantly, the ability to do something useful with such data—is the asset that generates the most wealth, leading to what is known as the data economy.

We should not forget that 3 out of the 5 most populated countries in the world (Facebook, WhatsApp, China, India and Instagram) are owned by Facebook. Digital countries, with global reach and less than 15 years of existence, which are governed by a non-democratically elected president. As a consequence of this phenomenon, a large percentage of today's data –and specially human behavioural data, that is, data about each of us—is private data that has been captured and is analized and leveraged by these technology giants which know not only our tastes, needs, interests or social relationships, but also our sexual or political orientation, our happiness or educational levels and even the state of our mental health.

Thus, if we could like to maximize the positive impact on society of the technologies that we are developing in the 4th Industrial Revolution –and particularly of Artificial Intelligence—we should think about new models of ownership, management, sharing, exploitation and regulation of data. Europe's GDPR is an example in this direction. However, the complexity of its practical application makes it evident the difficulties associated with defining the concept of property when we are talking about an intangible, distributed, varied, dynamic asset which is replicable infinite times at practically zero cost.

And finally, the N in FATEN is for non-maleficence.

5. N is for Non-maleficence

The principle of non-maleficence refers to minimising the negative impact that the development of data-driven decision-making algorithms might have on society. In the context of data-driven algorithmic decisions, I would like to highlight six components of the non-maleficence principle: reliability, security, reproducibility, robustness, prudence and privacy.

Reliability and security

The vast majority, if not all, of the systems, products and goods we use (food, household appliances, vehicles, clothing, toys, medicines, medical devices, industrial machinery, etc.) are subject to strict quality, safety and reliability controls to minimise the potential negative impact that they may have on society. Data-driven

algorithmic decision-making systems are also expected to be subject to similar processes. Beyond the theoretical processes of security, verification and reliability, it might make sense to create a European-level authority that would certify the quality, security and reliability of AI-based systems before they are commercialised or implemented within society. Also, autonomous systems should ensure the safety and integrity of the people who use them or are affected by their actions, and their own security against manipulation and cyberattacks.

Reproducibility and robustness

To generate confidence, systems should have consistency in their operation so that their behaviour is not only understandable by a human but is also reproducible, that is, it is replicable when subjected to the same input data or the same situation/context. In addition, there should be certain guarantees of the robustness of the data-driven algorithmic decision making systems that we might use. We know that Artificial Intelligence algorithms –like most software—are not fool-proof. In fact, there is an entire research area called Adversarial Machine Learning whose objective is the development of algorithms that would fool existing AI systems.

Prudence

The development of data-driven machine learning-based algorithms requires professionals to meet strict requirements, such as ensuring the availability of sufficient (high-quality) data, the analysis of working hypotheses from different perspectives and the availability of experts and resources to analyse and interpret the models and their results. The principle of prudence emphasises the importance of considering different options in the initial phases of the design of any system to maximise its positive impact and minimise the potential risks and negative consequences derived from its application.

Data protection and privacy

In a world in which data are generated and consumed in a ubiquitous and massive way, the rights to personal data protection and respect for privacy are constantly questioned and pushed to their limits. Numerous studies have focused on the misuse of the personal data provided by users of services and the aggregation of data from different sources by entities such as data brokers, with direct implications for people's privacy. An element that is often ignored, however, is that advances in machine-learning algorithms, combined with the availability of new sources of data on human behaviour (social media data, for example), allow the inference of private information – such as sexual orientation, political inclination, or levels of education and emotional stability – that has never been explicitly revealed. In a recent research project we showed that from non-person data, it is possible to infer attributes as personal as some dimensions of personality, level of education or interests (Park et al. 2018). This element is essential to understanding the implications of the use of algorithms to model, or even influence, human behaviour at the individual level, as was made clear in the recent Facebook/Cambridge Analytica scandal. Certain attributes and characteristics (sexual orientation, religion, etc.) should remain in the private sphere and should not be or inferred or used by AI systems unless the person expressly decides otherwise. Europe has assumed some leadership in this with the recent implementation of GDPR, which adds rights such as the right to establish and develop relationships with other human beings, the right to technological disconnection and the right to be free of vigilance. Other rights that could, or should, be added include the right to meaningful human contact – for example, in services operated exclusively by chatbots – and the right to not be measured, analysed, profiled, oriented or subliminally influenced via algorithms.

Finally, humans should always be placed at the core. The potential of algorithmic decision making will only be realised when policymakers are able to analyse the data, to study human behaviours and to test policies in the real world. A possible way forward is to build living laboratories – communities of volunteers willing to try new ways of doing things in a natural setting (Centellegher et al. 2016). These could provide a test-bed for designing and evaluating algorithmic policymaking approaches that encode societal values.

I believe that it is only when we respect these principles that we will be able to move forward and achieve a *model of democratic governance based on data and artificial intelligence, by and for the people*. The path forward must place humans and their societal values at the centre of discussions, as humans are ultimately both the actors and the subjects of the decisions made by algorithmic and human means. By involving people and ensuring that their values are upheld, we should be able to realise the immense positive potential of data-

driven algorithmic decision making while minimising the risks and the possible negative unintended consequences.

References

Andrae, A (2017), "Total Consumer Power Consumption Forecast", *Nordic Digital Business Summit*.

Angwin, J L (2016), *Machine bias*, ProPublica.

Barocas, S and A D Selbst (2016), Big data's disparate impact", *California Law Review* 104(3): 671-732.

Bogomolov, A, B Lepri, J Staiano, N Oliver, F Pianesi and A Pentland (2014), "One upon a crime: towards crime prediction from demographics and mobile data", *Proceedings of the 16th International Conference on Multimodal Interaction*, Istanbul.

Burrell, J (2016), "How the machine 'thinks': Understanding opacity in machine learning algorithms", *Big Data and Society* 3(1).

Centellegher, S, M De Nadai, M Caraviello, C Leonardi, M Vescovi, Y Ramadian, N Oliver, F Pianesi, A Pentland, F Antonelli and B Lepri (2016), "The Mobile Territorial Lab: a multilayered and dynamic view on parents' daily lives", *EPJ Data Science* 5(3).

Easterly, W (2014), *The tyranny of experts: Economists, dictators, and the forgotten rights of the poor*, Basic Books.

Fiske, S (1998), "Stereotyping, prejudice, and discrimination", in S T Fiske, D T Gilbert and G Lindzey (eds), *Handbook of Social Psychology*, McGraw-Hill, pp. 357-411.

Froelich, J, J Neumann and N Oliver (2009), "Sensing and Predicting the Pulse of the City through Shared Bicycling", *Proceedings of Twenty-First International Joint Conference on Artificial Intelligence*, Pasadena, CA.

Harari, Y N (2018), *21 lessons for the 21st century*, Penguin Random House.

Khanna, P (2017), *Technocracy in America: Rise of the info-state*, CreateSpace.

Kroll, J (2015), "Accountable Algorithms", PhD Dissertation, Princeton University.

LeCun, Y, Y Bengio and G Hinton (2015), "Deep Learning", *Nature* 521: 436-444.

Lepri, B, N Oliver, E Letouzé, A Pentland and P Vinck (2017), "Fair, Transparent, and Accountable Algorithmic Decision-making Processes", *Philosophy & Technology* 31(4): 611-627.

Oliver, N (2018), *Artificial Intelligence: fiction, reality and...dreams*, Spanish Royal Academy of Engineering.

Oliver, N, B Rosario and S Pentland (2000), "A Bayesian computer vision system for modeling human interactions", *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(8): 831-843.

O'Neil, C (2016), *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown.

Pager, D and H Shepherd (2008), "The sociology of discrimination: Racial discrimination in employment, housing, credit and consumer market", *Annual Review of Sociology* 34: 181-209.

Pariser, E (2012), *The filter bubble: how the personalized web is changing what we read and how we think*, Penguin Books.

Park, S, A Matic, K Garg and N Oliver (2018), "When Simpler Data Does Not Imply Less Information: A Study of User Profiling Scenarios With Constrained View of Mobile HTTP (S) Traffic", *ACM Transactions on the Web (TWEB)* 12(2).

Pasquale, F (2015), *The Black Blox Society: The secret algorithms that control money and information*, Harvard University Press.

Staiano, J, N Oliver, B Lepri, R de Oliveira, M Caraviello and N Sebe (2014), "Money walks: a human-centric study on the economics of personal mobile data", *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle, WA.

Sunstein, C (2012), "*Regulation in an uncertain world"*, National Academy of Sciences, 20 June (https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/speeches/regulation-in-an-uncertain-world-06202012.pdf).

Torres Fernández, Y, D Pastor Escuredo, A Morales Guzmán et al. (2014), "Flooding through the lens of mobile phone activity", *IEEE Global Humanitarian Technology Conference*, San Jose, CA.

Vieira, M, V Frias-Martinez, N Oliver and E Frias-Martinez (2010), "Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics", *IEEE Second International Conference on Social Computing*, Minneapolis, MN.

Willson, M (2016), "Algorithms (and the) everyday", *Information, Communication & Society* 20(1): 137-150.

Wilson, T D (2014), "Just think: The challenges of the disengaged mind", *Science* 345: 75-77.

Zarsky, T (2016), "The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making", *Science, Technology, and Human Values* 41(1): 118-132.