Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-0501-9 OPEN

Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward

Samuele Lo Piano [1,2]✉

Decision-making on numerous aspects of our daily lives is being outsourced to machine-learning (ML) algorithms and artificial intelligence (AI), motivated by speed and efficiency in the decision process. ML approaches—one of the typologies of algorithms underpinning artificial intelligence—are typically developed as black boxes. The implication is that ML code scripts are rarely scrutinised; interpretability is usually sacrificed in favour of usability and effectiveness. Room for improvement in practices associated with programme development have also been flagged along other dimensions, including *inter alia* fairness, accuracy, accountability, and transparency. In this contribution, the production of guidelines and dedicated documents around these themes is discussed. The following applications of *AI-driven* decision-making are outlined: (a) risk assessment in the criminal justice system, and (b) autonomous vehicles, highlighting points of friction across ethical principles. Possible ways forward towards the implementation of governance on AI are finally examined.

[1] School of the Built Environment, University of Reading, Reading, UK. [2] Open Evidence, Universitat Oberta de Catalunya, Barcelona, Catalonia, Spain. ✉email: s.lopiano@reading.ac.uk

Introduction

Artificial intelligence (AI) is the branch of computer science that deals with the simulation of intelligent behaviour in computers as regards their capacity to mimic, and ideally *improve*, human behaviour. To achieve this, the simulation of human cognition and functions, including learning and problem-solving, is required (Russell, 2010). This simulation may limit itself to some simple predictable features, thus limiting human complexity (Cowls, 2019).

AI became a self-standing discipline in the year 1955 (McCarthy et al., 2006) with significant development over the last decades. AI resorts to ML to implement a predictive functioning based on data acquired from a given context. The strength of ML resides in its capacity to learn from data without need to be explicitly programmed (Samuel, 1959); ML algorithms are autonomous and self-sufficient when performing their learning function. This is the reason why they are ubiquitous in AI developments. Further to this, ML implementations in data science and other applied fields are conceptualised in the context of a final decision-making application, hence their prominence.

Applications in our daily lives encompass fields, such as (precision) agriculture (Sennaar, 2019), air combat and military training (Gallagher, 2016; Wong, 2020), education (Sears, 2018), finance (Bahrammirzaee, 2010), health care (Beam and Kohane, 2018), human resources and recruiting (Hmoud and Laszlo, 2019), music composition (Cheng, 2009/09), customer service (Kongthon et al., 2009), reliable engineering and maintenance (Dragicevic et al., 2019), autonomous vehicles and traffic management (Ye, 2018), social-media news-feed (Rader et al., 2018), work scheduling and optimisation (O'Neil, 2016), and several others.

In all these fields, an increasing amount of functions are being ceded to algorithms to the detriment of human control, raising concern for loss of fairness and equitability (Sareen et al., 2020). Furthermore, issues of garbage-in-garbage-out (Saltelli and Funtowicz, 2014) may be prone to emerge in contexts when external control is entirely removed. This issue may be further exacerbated by the offer of new services of auto-ML (Chin, 2019), where the entire algorithm development workflow is automatised and the residual human control practically removed.

In the following sections, we will (i) detail a series of research questions around the ethical principles in AI; (ii) take stock of the production of guidelines elaborated in the field; (iii) showcase their prominence in practical examples; and (iv) discuss actions towards the inclusion of these dimensions in the future of AI ethics.

Research questions on the ethical dimensions of artificial intelligence

Critical aspects in AI deployment have already gained traction in mainstreaming literature and media. For instance, according to O'Neil (2016), a main shortcoming of ML approaches is the fact these resort to proxies for driving trends, such as person's ZIP code or language in relation with the capacity of an individual to pay back a loan or handle a job, respectively. However, these correlations may be discriminatory, if not illegal.

Potential *black swans* (Taleb, 2007) in the code should also be considered. These have been documented, for instance, in the case of the *Amazon* website, for which errors, such as the quotation of plain items (often books) up to 10,000 dollars (Smith, 2018) have been reported. While mistakes about monetary values may be easy to spot, the situation may become more complex and less intelligible when incommensurable dimensions come to play. That is the reason why a number of guidelines on the topic of ethics in AI have been proliferating over the last few years.

While reflections around the ethical implications of machines and automation deployment were already put forth in the '50s and '60s (Samuel, 1959; Wiener, 1988), the increasing use of AI in many fields raises new important questions about its suitability (Yu et al., 2018). This stems from the complexity of the aspects undertaken and the plurality of views, stakes, and values at play. A fundamental aspect is how and to what extent the values and the perspectives of the involved stakeholders have been taken care of in the design of the decision-making algorithm (Saltelli, 2020). In addition to this ex-ante evaluation, an ex-post evaluation would need to be put in place so as to monitor the consequences of *AI-driven* decisions in making winners and losers.

To wrap up, it is fundamental to assess how and if ethical aspects have been included in the *AI-driven* decision-making implemented by asking questions such as:

- What are the most prominent ethical concerns raised by large-scale deployment of AI applications?
- How are these multiple dimensions interwoven?
- What are the actions the involved stakeholders are carrying out to address these concerns?
- What are possible ways forward to improve ML and AI development and use over their full life-cycle?

We will firstly examine the production of relevant guidelines in the fields along with academic secondary literature. These aspects will then be discussed in the context of two applied cases: (i) recidivism-risk assessment in the criminal justice system, and (ii) autonomous vehicles.

Guidelines and secondary literature on AI ethics, its dimensions and stakes

The production of dedicated documents has been skyrocketing from 2016 (Jobin et al., 2019). We here report on the most prominent international initiatives. A suggested reading on national and international AI strategies providing a comprehensive list of documents (Future of Earth Institute, 2020).

The *France's Digital Republic Act* gives the right to an explanation as regards decisions on an individual made through the use of administrative algorithms (Edwards and Veale, 2018). This law touches upon several aspects including:

- how and to what extent the algorithmic processing contributed to the decision-making;
- which data was processed and its source;
- how parameters were treated and weighted;
- which operations were carried out in the treatment.

Sensitive governmental areas, such as national security and defence, and the private sector (the largest user and producer of ML algorithms by far) are excluded from this document.

An international European initiative is the multi-stakeholder *European Union High-Level Expert Group on Artificial Intelligence*, which is composed by 52 experts from academia, civil society, and industry. The group produced a deliverable on the required criteria for AI trustworthiness (Daly, 2019). Even articles 21 and 22 of the recent European Union *General Data Protection Regulation* include passages functional to AI governance, although further action has been recently demanded from the European Parliament (De Sutter, 2019). In this context, China has also been allocating efforts on privacy and data protection (Roberts, 2019).

As regards secondary literature, Floridi and Cowls (2019) examined a list of statements/declarations elaborated since 2016 from multi-stakeholder organisations. A set of 47 principles has been identified, which mapped onto five overarching dimensions

(Floridi and Cowls, 2019): *beneficence, non-maleficence, autonomy, justice* and, *explicability*. The latter is a new dimension specifically acknowledged in the case of AI, while the others were already identified in the controversial domain of *bioethics*.

Jobin et al. (2019) reviewed 84 documents, which were produced by several actors of the field, almost half of which from private companies or governmental agencies. The classification proposed by Jobin et al. (2019) is around a slightly different set of values: *transparency, justice and fairness, non-maleficience, responsibility* and *privacy*. Other potentially relevant dimensions, such as accountability and responsibility, were rarely defined in the studies reviewed by these authors.

Seven of the most prominent value statements from the AI/ML fields were examined in Greene et al. (2019): *The Partnership on AI to Benefit People and Society*; *The Montreal Declaration for a Responsible Development of Artificial Intelligence*; *The Toronto Declaration Protecting the rights to equality and non-discrimination in machine-learning systems*; *OpenAI*; *The Centre for Humane Technology*; *Fairness, Accountability and Transparency in Machine Learning*; *Axon's AI Ethics Board for Public Safety*. Greene et al. (2019) found seven common core elements across these documents: (i) design's moral background (universal concerns, objectively measured); (ii) expert oversight; (iii) values-driven determinism; (iv) design as locus of ethical scrutiny; (v) better building; (vi) stakeholder-driven legitimacy; and, (vii) machine translation.

Mittelstadt (2019) critically analysed the current debate and actions in the field of AI ethics and noted that the dimensions addressed in AI ethics are converging towards those of medical ethics. However, this process appears problematic due to four main differences between medicine and the medical professionals on one side, and AI and its developers on the other. Firstly, the medical professional rests on common aims and fiduciary duties, which AI developers lack. Secondly, a formal profession with a set of clearly defined and governed good-behaviour practices exists in medicine. This is not the case for AI, which also lacks a full understanding of the consequences of the actions enacted by algorithms (Wallach and Allen, 2008). Thirdly, AI faces the difficulty of translating overarching principle into practices. Even its current setting of seeking maximum speed, efficiency and profit clashes with the resource and time requirements of an ethical assessment and/or counselling. Finally, the accountability of professionals or institutions is at this stage mainly theoretical, having the vast majority of these guidelines been defined on a merely voluntary basis and hence with the total lack of a sanctionary scheme for non-compliance.

Points of friction between ethical dimensions. Higher transparency is a common refrain when discussing ethics of algorithms, in relation to dimensions such as how an algorithmic decision is arrived at, based on what assumptions, and how this could be corrected to incorporate feedback from the involved parties. Rudin (2019) argued that the community of algorithm developers should go beyond explaining black-box models by developing interpretable models in the first place.

On a larger scale, the use of open-source software in the context of ML applications has already been advocated for over a decade (Thimbleby, 2003) with an indirect call for tools to execute more interpretable and reproducible programming such as *Jupyter Notebooks*, available from 2015 onwards. However, publishing scripts expose their developers to the public scrutiny of professional programmers, who may find shortcomings in the development of the code (Sonnenburg, 2007).

Ananny and Crawford (2018) comment that resorting to full algorithmic transparency may not be an adequate means to address their ethical dimensions; opening up the black-box would not suffice to disclose their *modus operandi*. Moreover, developers of algorithm may not be capable of explaining in plain language how a given tool works and what functional elements it is based on. A more social relevant understanding would encompass the human/non-human interface (i.e., looking *across* the system rather than merely *inside*). Algorithmic complexity and all its implications unravel at this level, in terms of relationships rather than as mere self-standing properties.

Other authors pointed to possible points of friction between transparency and other relevant ethical dimensions. de Laat (2018) argues that transparency and accountability may even be at odds in the case of algorithms. Hence, he argues against full transparency along four main lines of reasoning: (i) leaking of privacy sensitive data into the open; (ii) backfiring into an implicit invitation to game the system; (iii) harming of the company property rights with negative consequences on their competitiveness (and on the developers reputation as discussed above); (iv) inherent opacity of algorithms, whose interpretability may be even hard for experts (see the example below about the code adopted in some models of autonomous vehicles). All these arguments suggest limitations to full disclosure of algorithms, be it that the normative implications behind these objections should be carefully scrutinised.

Raji et al. (2020) suggest that a process of algorithmic auditing within the software-development company could help in tackling some of the ethical issues raised. Larger interpretability could be in principle achieved by using simpler algorithms, although this may come at the expenses of accuracy. To this end, Watson and Floridi (2019) defined a formal framework for interpretable ML, where explanatory accuracy can be assessed against algorithmic simplicity and relevance.

Loss in accuracy may be produced by the exclusion of politically critical features (such as gender, race, age, etc.) from the pool of training predictive variables. For instance, *Amazon* scrapped a gender-biased recruitment algorithm once it realised that despite excluding gender, the algorithm was resorting to surrogate gender variables to implement its decisions (Dastin, 2018). This aspect points again to possible political issues of a trade-off between fairness, demanded by society, and algorithmic accuracy, demanded by, e.g., a private actor.

Fairness may be further hampered by reinforcement effects. This is the case of algorithms attributing credit scores, that have a reinforcement effect proportional to people wealth that de facto rules out credit access for people in a more socially difficult condition (O'Neil, 2016).

According to Floridi and Cowls (2019) a prominent role is also played by the *autonomy* dimension; the possibility of refraining from ceding decision power to AI for overriding reasons (e.g., the gain of efficacy is not deemed fit to justify the loss of control over decision-making). In other words, machines autonomy could be reduced in favour of human autonomy according to this *meta-autonomy* dimension.

Contrasting dimensions in terms of the theoretical framing of the issue also emerged from the review of Jobin et al. (2019), as regards interpretation of ethical principles, reasons for their importance, ownership and responsibility of their implementation. This also applies to different ethical principles, resulting in the trade-offs previously discussed, difficulties in setting prioritisation strategies, operationalisation and actual compliance with the guidelines. For instance, while private actors demand and try to cultivate trust from their users, this runs counter to the need for society to scrutinise the operation of algorithms in order to maintain developer accountability (Cowls, 2019). Attributing responsibilities in complicated projects where many parties and developers may be involved, an issue known as *the problem of many hands* (Nissenbaum, 1996), may indeed be very difficult.

Conflicts may also emerge between the requirements to overcome potential algorithm deficits in accuracy associated with large data bases and the individual rights to privacy and autonomy of decision. Such conflicts may exacerbate tensions, further complicating agreeing on standards and practices.

In the following two sections, the issues and points of friction raised are examined in two practical case studies, criminal justice and autonomous vehicles. These examples have been selected due to their prominence in the public debate on the ethical aspects of AI and ML algorithms.

Machine-learning algorithms in the field of criminal justice

ML algorithms have been largely used to assist juridical deliberation in many states of the USA (Angwin and Larson, 2016). This country faces the issue of the world's highest incarcerated population, both in absolute and per-capita terms (Brief, 2020). The COMPAS algorithm, developed by the private company *Northpointe*, attributes a 2-year recidivism-risk score to arrested people. It also evaluates the risk of violent recidivism as a score.

The fairness of the algorithm has been questioned in an investigative report, that examined a pool of cases where a *recidivism score* was attributed to >18,000 criminal defendants in Broward County, Florida and flagged up a potential racial bias in the application of the algorithm (Angwin and Larson, 2016). According to the authors of the report, the recidivism-risk was systematically overestimated for black people: the decile distribution of white defendants was skewed towards the lower end. Conversely, the decile distribution of black defendants was only slightly decreasing towards the higher end. The risk of violent recidivism within 2 years followed a similar trend. This analysis was debunked by the company, which, however, refused to disclose the full details of its proprietary code. While the total number of variables amounts to about 140, only the core variables were disclosed (Northpointe, 2012). The race of the subject was not one of those.

Here, a crucial point is how this fairness is to be attained: whether it is more important a fair treatment across groups of individuals or within the same group. For instance, let us take the case of gender, where men are overrepresented in prison in comparison with women. As to account for this aspect, the algorithm may discount violent priors for men in order to reduce their recidivism-risk score. However, attaining this sort of algorithmic fairness would imply inequality of treatment across genders (Berk et al., 2018).

Fairness could be further hampered by the combined use of this algorithm with others driving decisions on neighbourhood police patrolling. The fact these algorithms may be prone to drive further patrolling in poor neighbourhoods may result from a training bias as crimes occurring in public tend to be more frequently reported (Karppi, 2018). One can easily understand how these algorithms may jointly produce a vicious cycle—more patrolling would lead to more arrests that would worsen the neighbourhood average *recidivism-risk score*, which would in turn trigger more patrolling. All this would result in exacerbated inequalities, likewise the case of credit scores previously discussed (O'Neil, 2016).

A potential point of friction may also emerge between the algorithm dimensions of fairness and accuracy. The latter may be theoretically defined as the classification error in terms of rate of false positive (individuals labelled at risk of recidivism, that did not re-offend within 2 years) and false negative (individuals labelled at low risk of recidivism, that did re-offend within the same timeframe) (Loi and Christen, 2019). Different classification accuracy (the fraction of observed outcomes in disagreement with the predictions) and forecasting accuracy (the fraction of predictions in disagreement with the observed outcomes) may exist across different classes of individuals (e.g., black or white defendants). Seeking equal rates of false positive and false negative across these two pools would imply a different forecasting error (and accuracy) given the different characteristics of the two different training pools available for the algorithm. Conversely, having the same forecasting accuracy would come at the expense of different classification errors between these two pools (Corbett-Davies et al., 2016). Hence, a trade-off exists between these two different shades of fairness, which derives from the very statistical properties of the data population distributions the algorithm has been trained on. However, the decision-making rests again on the assumptions the algorithm developers have adopted, e.g., on the relative importance of false positive and false negative (i.e., the weights attributed to the different typologies of errors, and the accuracy sought (Berk, 2019)). When it comes to this point, an algorithm developer may decide (or be instructed) to train his/her algorithm to attribute, e.g., a five/ten/twenty times higher weight for a false negative (re-offender, low recidivism-risk score) in comparison with a false positive (non re-offender, high recidivism-risk score).

As with all ML, an issue of transparency exists as no one knows what type of inference is drawn on the variables out of which the recidivism-risk score is estimated. Reverse-engineering exercises have been run so as to understand what are the key drivers on the observed scores. Rudin (2019) found that the algorithm seemed to behave differently from the intentions of their creators (Northpointe, 2012) with a non-linear dependence on age and a weak correlation with one's criminal history. These exercises (Rudin, 2019; Angelino et al., 2018) showed that it is possible to implement interpretable classification algorithms that lead to a similar accuracy as COMPAS. Dressel and Farid (2018) achieved this result by using a linear predictor-logistic regressor that made use of only two variables (age and total number of previous convictions of the subject).

Machine-learning algorithms in the field of autonomous vehicles

The case of autonomous vehicles, also known as self-driving vehicles, poses different challenges as a continuity of decisions is to be enacted while the vehicle is moving. It is not a one-off decision as in the case of the assessment of recidivism risk.

An exercise to appreciate the value-ladenness of these decisions is the moral-machine experiment (Massachussets Institute of Technology 2019)—a serious game where users are requested to fulfil the function of an autonomous-vehicle decision-making algorithm in a situation of danger. This experiment entails performing choices that would prioritise the safety of some categories of users over others. For instance, choosing over the death of car occupants, pedestrians, or occupants of other vehicles, et cetera. While such extreme situations may be a simplification of reality, one cannot exclude that the algorithms driving an autonomous-vehicle may find themselves in circumstances where their decisions may result in harming some of the involved parties (Bonnefon et al., 2019).

In practice, the issue would be framed by the algorithm in terms of a *statistical trolley dilemma* in the words of Bonnefon et al. (2019), whereby the risk of harm for some road users will be increased. This corresponds to a *risk management* situation by all means, with a number of nuances and inherent complexity (Goodall, 2016).

Hence, autonomous vehicles are not bound to play the role of silver bullets, solving once and forever the vexing issue of traffic fatalities (Smith, 2018). Furthermore, the way decisions enacted could backfire in complex contexts to which the algorithms had

no extrapolative power, is an unpredictable issue one has to deal with (Wallach and Allen, 2008; Yurtsever et al., 2020).

Coding algorithms that assure fairness in autonomous vehicles can be a very challenging issue. Contrasting and incommensurable dimensions are likely to emerge (Goodall, 2014) when designing an algorithm to reduce the harm of a given crash. For instance, in terms of material damage against human harm. Odds may emerge between the interest of the vehicle owner and passengers, on one side, and the collective interest of minimising the overall harm, on the other. Minimising the overall physical harm may be achieved by implementing an algorithm that, in the circumstance of an unavoidable collision, would target the vehicles with the highest safety standards. However, one may want to question the fairness of targeting those who have invested more in their own and others' safety. The algorithm may also face a dilemma between low probability of a serious harm and higher probability of a mild harm. Unavoidable normative rules will need to be included in the decision-making algorithms to tackle these types of situations.

Accuracy in the context of self-autonomous vehicles rests on their capacity to correctly simulate the course of the events. While this is based on physics and can be informed by the numerous sensors these vehicles are equipped with, unforeseen events can still play a prominent role, and profoundly affect the vehicles behaviour and reactions (Yurtsever et al., 2020). For instance, fatalities due to autonomous-vehicle malfunctioning were reported as caused by the following failures: (i) the incapability of perceiving a pedestrian as such (National Transport Safety Board 2018); (ii) the acceleration of the vehicle in a situation when braking was required due to contrasting instructions from different algorithms the vehicle was hinged upon (Smith, 2018). In this latter case, the complexity of autonomous-vehicle algorithms was witnessed by the millions lines of code composing their scripts, *a universe no one fully understands* in the words of *The Guardian* (Smith, 2018), so that the causality of the decisions made was practically impossible to scrutinise. Hence, no corrective action in the algorithm code may be possible at this stage, with no room for improvement in accuracy.

One should also not forget that these algorithms are learning by direct experience and they may still end up conflicting with the initial set of ethical rules around which they have been conceived. Learning may occur through algorithms interaction taking place at a higher hierarchical level than the one imagined in the first place (Smith, 2018). This aspect would represent a further open issue to be taken into account in their development (Markham et al., 2018). It also poses further tension between the accuracy a vehicle manufacturer seeks and the capability to keep up the agreed fairness standards upstream from the algorithm development process.

Discussion and conclusions

In this contribution, we have examined the ethical dimensions affected by the application of algorithm-driven decision-making. These are entailed both ex-ante, in terms of the assumptions underpinning the algorithm development, and ex-post as regards the consequences upon society and social actors on whom the elaborated decisions are to be enforced.

Decision-making-based algorithms rest inevitably on assumptions, even silent ones, such as the quality of data the algorithm is trained on (Saltelli and Funtowicz, 2014), or the actual modelling relations adopted (Hoerl, 2019), with all the implied consequences (Saltelli, 2019).

A decision-making algorithm will always be based on a formal system, which is a representation of a real system (Rosen, 2005). As such, it will always be based on a restricted set of relevant relations, causes, and effects. It does not matter how complicated the algorithm may be (how many relations may be factored in), it will always represent one-specific vision of the system being modelled (Laplace, 1902).

Eventually, the set of decision rules underpinning the AI algorithm derives from human-made assumptions, such as, where to define the boundary between action and no action, between different possible choices. This can only take place at the human/non-human interface: the response of the algorithm is driven by these human-made assumptions and selection rules. Even the data on which an algorithm is trained on are not an objective truth, they are dependent upon the context in which they have been produced (Neff et al., 2017).

Tools for technically scrutinising the potential behaviour of an algorithm and its uncertainty already exist and could be included in the workflow of algorithm development. For instance, global sensitivity analysis (Saltelli, 2008) may help in exploring how the uncertainty in the input parameters and modelling assumptions would affect the output. Additionally, a modelling of the modelling process would assist in the model transparency and in addressing questions such as: *Are the results from a particular model more sensitive to changes in the model and the methods used to estimate its parameters, or to changes in the data?* (Majone, 1989).

Tools of post-normal-science inspiration for knowledge and modelling quality assessment could be adapted to the analysis of algorithms, such as the NUSAP (Numeral Unit Spread Assessment Pedigree) notation system for the management and communication of uncertainty (Funtowicz and Ravetz, 1990; Van Der Sluijs, 2005) and sensitivity auditing (Saltelli and Funtowicz, 2014), respectively. Ultimately, developers should acknowledge the limits of AI, and what its ultimate function should be in the equivalent of an Hippocratic Oath for ML developers (O'Neil, 2016). An example comes from the field of financial modelling, with a manifesto elaborated in the aftermath of the 2008 financial crisis (Derman and Wilmott, 2009).

As to address these dimensions, value statements and guidelines have been elaborated by political and multi-stakeholder organisations. For instance, *The Alan Turing Institute* released a guide for responsible design and implementation of AI (Leslie, 2019) that covers the whole life-cycle of design, use, and monitoring. However, the field of AI ethics is just at its infancy and it is still to be conceptualised how AI developments that encompass ethical dimensions could be attained. Some authors are pessimistic, such as Supiot (2017) who speaks of *governance by numbers*, where quantification is replacing the traditional decision-making system and profoundly affecting the pillar of equality of judgement. Trying to revert the current state of affairs may expose the first movers in the AI field to a competitive disadvantage (Morley et al., 2019). One should also not forget that points of friction across ethical dimensions may emerge, e.g., between transparency and accountability, or accuracy and fairness as highlighted in the case studies. Hence, the development process of the algorithm cannot be perfect in this setting, one has to be open to negotiation and unavoidably work with imperfections and clumsiness (Ravetz, 1987).

The development of decision-making algorithms remains quite obscure in spite of the concerns raised and the intentions manifested to address them. Attempts to expose to public scrutiny the algorithms developed are yet scant. As are the attempt to make the process more inclusive, with a higher participation from all the stakeholders. Identifying a relevant pool of social actors may require an important effort in terms of stakeholders' mapping so as to assure a complete, but also effective, governance in terms of number of participants and simplicity of working procedures. The post-normal-science concept of extended peer communities could

assist also in this endeavour (Funtowicz and Ravetz, 1997). Example-based explanations (Molnar, 2020) may also contribute to an effective engagement of all the parties by helping in bridging technical divides across developers, experts in other fields, and lay-people.

An overarching meta-framework for the governance of AI in experimental technologies (i.e., robot use) has also been proposed (Rego de Almeida et al., 2020). This initiative stems from the attempt to include all the forms of governance put forth and would rest on an integrated set of feedback and interactions across dimensions and actors. An interesting proposal comes from Berk (2019), who asked for the intervention of *super partes* authorities to define standards of transparency, accuracy and fairness for algorithm developers in line with the role of the *Food and Drug administration* in the US and other regulation bodies. A shared regulation could help in tackling the potential competitive disadvantage a first mover may suffer. The development pace of new algorithms would be necessarily reduced so as to comply with the standards defined and the required clearance processes. In this setting, seeking algorithm transparency would not be harmful for their developers as scrutiny would be delegated to entrusted intermediate parties, to take place behind *closed doors* (de Laat, 2018).

As noted by a perceptive reviewer, ML systems that keep learning are dangerous and hard to understand because they can quickly change. Thus, could a ML system with real world consequences be "locked down" to increase transparency? If yes, the algorithm could become defective. If not, transparency today may not be helpful in understanding what the system does tomorrow. This issue could be tackled by hard-coding the set of rules on the behaviour of the algorithm, once these are agreed upon among the involved stakeholders. This would prevent the algorithm-learning process from conflicting with the standards agreed. Making mandatory to deposit these algorithms in a database owned and operated by this entrusted super-partes body could ease the development of this overall process.

Received: 29 January 2020; Accepted: 12 May 2020;
Published online: 17 June 2020

References

Ananny M, Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society 20:973–989

Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C (2018) Learning certifiably optimal rule lists for categorical data. http://arxiv.org/abs/1704.01701

Angwin J, Larson J (2016) Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bahrammirzaee A (2010) A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Comput Appl 19:1165–1195

Beam AL, Kohane IS (2018) Big data and machine learning in health care. JAMA 319:1317

Berk R (2019) Machine learning risk assessments in criminal justice settings. Springer International Publishing, Cham

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: the state of the art. Soc Methods Res 004912411878253

Board NTS (2018) Vehicle automation report. Tech. Rep. HWY18MH010, Office of Highway Safety, Washington, D.C.

Bonnefon J-F, Shariff A, Rahwan I (2019) The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. Proc IEEE 107:502–504

Brief WP (2020) World prison brief- an online database comprising information on prisons and the use of imprisonment around the world. https://www.prisonstudies.org/

Cheng J (2009) Virtual composer makes beautiful music and stirs controversy. https://arstechnica.com/science/news/2009/09/virtual-composer-makes-beautiful-musicand-stirs-controversy.ars

Chin J (2019) The death of data scientists. https://towardsdatascience.com/the-death-of-data-scientists-c243ae167701

Corbett-Davies S, Pierson E, Feller A, Goel S (2016) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. Washington Post. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Cowls J (2020) Deciding how to decide: six key questions for reducing AI's democratic deficit. In: Burr C, Milano S (eds) The 2019 Yearbook of the Digital Ethics Lab, Digital ethics lab yearbook. Springer International Publishing, Cham. pp. 101–116. https://doi.org/10.1007/978-3-030-29145-7_7

Daly A et al. (2019) Artificial intelligence, governance and ethics: global perspectives. SSRN Electron J. https://www.ssrn.com/abstract=3414805

Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

De Sutter P (2020) Automated decision-making processes: ensuring consumer protection, and free movement of goods and services. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2020/01-22/Draft_OQ_Automated_decision-making_EN.pdf

Derman E, Wilmott P (2009) The financial modelers' manifesto. SSRN Electron J. http://www.ssrn.com/abstract=1324878.

Dragičević T, Wheeler P, Blaabjerg F (2019) Artificial intelligence aided automated design for reliability of power electronic systems. IEEE Trans Power Electron 34:7161–7171

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4:eaao5580

Edwards L, Veale M (2018) Enslaving the algorithm: from A -right to an explanation- to A -right to better decisions-? IEEE Security, Priv 16:46–54

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Science Review. https://hdsr.mitpress.mit.edu/pub/l0jsh9d1

Funtowicz SO, Ravetz JR (1990) Uncertainty and quality in science for policy. Springer Science, Business Media, Berlin, Heidelberg

Funtowicz S, Ravetz J (1997) Environmental problems, post-normal science, and extended peer communities. Études et Recherches sur les Systémes Agraires et le Développement. INRA Editions. pp. 169–175

Future of Earth Institute (2020) National and International AI Strategies. https://futureoflife.org/national-international-ai-strategies/

Gallagher S (2016) AI bests Air Force combat tactics experts in simulated dogfights. https://arstechnica.com/information-technology/2016/06/ai-bests-air-force-combat-tactics-experts-in-simulated-dogfights/

Goodall NJ (2014) Ethical decision making during automated vehicle crashes. Transportation Res Rec: J Transportation Res Board 2424:58–65

Goodall NJ (2016) Away from trolley problems and toward risk management. Appl Artif Intell 30:810–821

Greene D, Hoffmann AL, Stark L (2019) Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences

Hmoud B, Laszlo V (2019) Will artificial intelligence take over human-resources recruitment and selection? Netw Intell Stud VII:21–30

Hoerl RW (2019) The integration of big data analytics into a more holistic approach-JMP. Tech. Rep., SAS Institute. https://www.jmp.com/en_us/whitepapers/jmp/integration-of-big-data-analytics-holistic-approach.html

Jobi A, Ienca M, Vayena E (2019) Artificial intelligence: the global landscape of ethics guidelines. Nat Mach Intell 1:389–399

Karppi T (2018) The computer said so-: on the ethics, effectiveness, and cultural techniques of predictive policing. Soc Media + Soc 4:205630511876829

Kongthon A, Sangkeettrakarn C, Kongyoung S, Haruechaiyasak C (2009) Implementing an online help desk system based on conversational agent. In: Proceedings of the International Conference on Management of Emergent Digital EcoSystems, MEDES '09, vol. 69. ACM, New York, NY, USA. pp. 450–69:451. Event-place: France. https://doi.org/10.1145/1643823.1643908

de Laat PB (2018) Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos Technol 31:525–541

Laplace PS (1902) A philosophical essay on probabilities. J. Wiley, New York; Chapman, Hall, London. http://archive.org/details/philosophicaless00lapliala

Leslie D (2019) Understanding artificial intelligence ethics and safety. http://arxiv.org/abs/1906.05684

Loi M, Christen M (2019) How to include ethics in machine learning research. https://ercim-news.ercim.eu/en116/r-s/how-to-include-ethics-in-machine-learning-research

Majone G (1989) Evidence, argument, and persuasion in the policy process. Yale University Press, Yale

Markham AN, Tiidenberg K, Herman A (2018) Ethics as methods: doing ethics in the era of big data research-introduction. Soc Media + Soc 4:205630511878450

Massachussets Institute of Technology (2019) Moral machine. Massachussets Institute of Technology. http://moralmachine.mit.edu

McCarthy J, Minsky ML, Rochester N, Shannon CE (2006) A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. AI Mag 27:12–12
Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507
Molnar C (2020) Interpretable machine learning (2020). https://christophm.github.io/interpretable-ml-book/
Morley J, Floridi L, Kinsey K, Elhalal A (2019) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Tech Rep. https://arxiv.org/abs/1905.06876
Neff G, Tanweer A, Fiore-Gartland B, Osburn L (2017) Critique and contribute: a practice-based framework for improving critical data studies and data science. Big Data 5:85–97
Nissenbaum H (1996) Accountability in a computerized society. Sci Eng Ethics 2:25–42
Northpointe (2012) Practitioner's guide to COMPAS. northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf
O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy, 1st edn. Crown, New York
Rader E, Cotter K, Cho J (2018) Explanations as mechanisms for supporting algorithmic transparency. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. ACM Press, Montreal QC, Canada. pp. 1–13. http://dl.acm.org/citation.cfm?doid=3173574.3173677
Raji ID et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency pp 33–44 (Association for Computing Machinery, 2020). https://doi.org/10.1145/3351095.3372873
Ravetz JR (1987) Usable knowledge, usable ignorance: incomplete science with policy implications. Knowledge 9:87–116
Rêgo de Almeida PG, Denner dos Santos C, Silva Farias J (2020) Artificial intelligence regulation: a meta-framework for formulation and governance. In: Proceedings of the 53rd Hawaii International Conference on System Sciences (2020). http://hdl.handle.net/10125/64389
Roberts H et al. (2019) The Chinese approach to artificial intelligence: an analysis of policy and regulation. SSRN Electron J. https://www.ssrn.com/abstract=3469784
Rosen R (2005) Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life. Columbia University Press, New York
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. http://arxiv.org/abs/1811.10154
Russell SJ (2010) Artificial intelligence : a modern approach. Prentice Hall, Upper Saddle River, NJ
Saltelli A et al. (2008) Global sensitivity analysis: the primer. Wiley, Hoboken, NJ
Saltelli A (2019) A short comment on statistical versus mathematical modelling. Nat Commun 10:3870
Saltelli A (2020) Ethics of quantification or quantification of ethics? Futures 116:102509
Saltelli A, Funtowicz S (2014) When all models are wrong. Issues Sci Technol 30:79–85
Samuel AL (1959) Some studies in machine learning using the game of checkers. IBM J Res Dev 3:210–229
Sareen S, Saltelli A, Rommetveit K (2020) Ethics of quantification: illumination, obfuscation and performative legitimation. Palgrave Commun 6:1–5
Sears (2018) The role of artificial intelligence in the classroom. https://elearningindustry.com/artificial-intelligence-in-the-classroom-role
Sennaar K (2019) AI in agriculture-present applications and impact. https://emerj.com/ai-sector-overviews/ai-agriculture-present-applications-impact/
Van Der Sluijs JP et al. (2005) Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: The NUSAP system. Risk Anal 25:481–492
Smith A (2018) Franken-algorithms: the deadly consequences of unpredictable code. The Guardian. https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger
Sonnenburg S et al. (2007) The need for open source software in machine learning. J Mach Learn Res 8:2443–2466
Supiot A (2017) Governance by numbers: the making of a legal model of allegiance. Hart Publishing, Oxford; Portland, Oregon
Taleb NN (2007) The Black Swan: the impact of the highly improbable. Random House Publishing Group, New York, NY
Thimbleby H (2003) Explaining code for publication. Softw: Pract Experience 33:975–1001
Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford, USA
Watson D, Floridi L (2019) The explanation game: A formal framework for interpretable machine learning. https://papers.ssrn.com/abstract=3509737
Wiener N (1988) The human use of human beings: cybernetics and society. Da Capo Press, New York, N.Y, new edition
Wong YH et al. (2020). Deterrence in the age of thinking machines: product page. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2797.html
Ye H et al. (2018) Machine learning for vehicular networks: recent advances and application examples. IEEE Vehicular Technol Mag 13:94–101
Yu H et al. (2018) Building ethics into artificial intelligence. http://arxiv.org/abs/1812.02953
Yurtsever E, Capito L, Redmill K, Ozguner U (2020) Integrating deep reinforcement learning with model-based path planners for automated driving. http://arxiv.org/abs/2002.00434

Acknowledgements

I would like to thank Kjetil Rommetveit, Andrea Saltelli and Siddarth Sareen for the organisation of the Workshop *Ethics of Quantification*, and the Centre for the Study of Sciences and the Humanities of the University of Bergen for the travel grant, at which a previous version of this paper was presented. Thomas Hodgson, Jill Walter Rettberg, Elizabeth Chatterjee, Ragnar Fjelland and Marta Kuc-Czarnecka for their useful comments in this venue. Finally, Stefn Thor Smith and Andrea Saltelli for their suggestions and constructive criticism on a draft version of the present manuscript.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.L.P.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.