What if algorithms could abide by ethical principles?

Algorithms are step-by-step procedures for solving a problem, usually expressed in computer code as a set of instructions for a computer to follow in order to complete a task. An algorithm can be hand-coded by a programmer or generated automatically from data, as in machine learning. The latter is considered a form of artificial intelligence. Day-to-day decisions around the world are based increasingly on data science techniques powered by machine-learning algorithms. For example, the intermediary platforms that propose accommodation (AirBnB) or transport services (Uber) use algorithms extensively. At the same time, algorithms, implicitly or explicitly, are not neutral, as their design is based on value-laden judgments that can potentially have race or sex biases, for example. This raises an important question: is it possible to ensure that algorithms are ethical?

Algorithms are widely employed to make decisions that have far-reaching impacts on individuals and society. They have the power, not least, to affect the distribution of social goods such as education, employment, police protection and medical care, as well as the protection of fundamental rights such as the right to life, the right to a fair trial and the presumption of innocence, the right to privacy, freedom of expression and workers' rights. Cases such as the Volkswagen algorithm, which enabled vehicles to pass emissions tests by reducing their nitrogen oxide emissions during those tests, and policing software for pinpointing repeat offenders, which frequently appears biased against black people, are instances of how algorithms' ethical shortcomings have led to 'algorithmic tragedies'.

© wladimir1804 / Fotolia.

Learning algorithms can invisibly reproduce and deepen various forms of prejudiced social classification, manifesting a new form of 'rational discrimination' harming people's life-chances. In fact, algorithms may amplify racial and gender biases as they contain the values and judgements of their human developers, who decide which data to include or exclude and how to weight each component. Biased or incomplete data can build flawed statistical models and reinforce societal biases if there are no impact assessment, audit or oversight procedures in place. Far from eradicating human biases, algorithms could magnify and entrench them, potentially leading to loss of human agency, especially if algorithms start creating new algorithms.

Thus, a 2016 ProPublica analysis of the use of automated decision-support software under the name COMPAS[1] uncovered evidence of racial bias within the US criminal justice system. Nevertheless, US judges are increasingly using this risk assessment algorithm to ground their decisions across a range of stages in the criminal justice process. In addition, recent evidence indicates discrimination against communities of colour resulting from the use of credit-scoring systems in the US. In the case of health data, there are three ways in which bias can have an impact: human bias; bias that is introduced by design; and bias in the ways systems use the data. In the same context, the development by UK local councils, amid mounting financial pressure, of 'predictive analytics' systems to algorithmically identify families for attention from child services may intrude into individual privacy and reinforce the stigmatisation of certain population groups.

[1] COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions

Author: Mihalis Kritikos, Scientific Foresight Unit (STOA)

Challenges to the design of ethical algorithms

The embedding of ethical principles in algorithmic decision-making would have its challenges, given the proprietary nature of algorithms and the need to safeguard privacy. Certain questions need to be asked before implementing an algorithmic decision-making system. Should the ethical assumptions in the algorithm be transparent and easy for users to identify? What about the development teams that create them – are they sufficiently diverse? Will people affected by these decisions have any influence over the system? Even if algorithms were made transparent, how could they be understood by multiple stakeholders of varying technical algorithmic literacy? Could ethical principles such as fairness and the right to privacy be encoded in the system and, if so, what should those principles be and who should decide upon their choice and weighting? Do our societies have universal, moral standards that can be codified?

In response to a mounting number of news articles about the ethics of algorithms, various market solutions that offer 'algorithmic accuracy, bias and fairness' certification are starting to emerge, including the AI fairness toolkit, Audit-AI by Pymetrics, Facebook's Fairness Flow and ORCAA. Recognition of the need to operationalise moral judgement for the development of autonomous vehicles and to integrate artificial moral agents that can manage complexity into new technologies has led to a series of algorithm design initiatives based on various ethical theories, such as one by the National Science Foundation. The emergence of artificial intelligence and advanced machine-learning may lead to self-driving cars being equipped with an ethical knob that could set key patterns of behaviour. IBM has meanwhile developed a new set of open-source software in order to help developers deal with black-box algorithms and understand how the artificial intelligence they use makes decisions.

What does developing ethical algorithms mean for European policy-making?

When it comes to ethical impacts of algorithmic decision-making systems, there are as yet no established certification models and procedures that could expressly address ethical considerations, including bias and transparency, in the domain of algorithms. This is partially due to a lack of existing standards on these issues to certify against. Developing ethical principles and codes for algorithms means identifying the decision-making principles and norms and the allocation of roles and responsibilities of the decision system.

The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems, the Toronto Declaration and Facebook's Fairness Flow indicate the need to set socially oriented goals and benchmarks for the development of algorithms. Given, however, that algorithms are unstable objects of ethical scrutiny, their ethics could still be investigated via the use of algorithmic impact assessments, and 'algorithmic audits' may need to become a legal requirement when implementing any systems of this kind. These audits would address ethical questions, such as the legitimacy of the use of an algorithmic decision-making system in certain contexts (e.g. evidence-based sentencing or lethal weapons), and be performed by ethical committees, accreditation bodies and certification agencies.

Their aim should be to evaluate the proposed uses of algorithmic decision-making in highly sensitive and/or safety-critical application domains and investigate suspected cases of rights violations in the frame of the same technological context. Moreover, these instruments could help system developers and decision-makers revisit some of their own assumptions of what an algorithm actually is, and explain decisions in areas such as credit, for instance.

Interestingly, the requirement for data controllers to provide data subjects with 'meaningful information about the logic involved' in an automated decision-making process – introduced by the General Data Protection Regulation (GDPR) – may pave the way for the development of practical algorithmic ethics that address virtues, consequences and norms. Shedding light on the assumptions built into the algorithm or disclosing the code of the system or information about its logic demands a careful examination of the relevant rules concerning intellectual property rights that may set limits on accessibility.

To conclude, EU policy-makers have a unique opportunity to lead the world in the ethical regulation of the digital revolution, by promoting the development of a general ethical framework governing the design, implementation and development of algorithms. These should remain under human oversight and control and be responsive to bias complaints and to the findings of reports on other undesired effects.