Lyria Bennett Moses

Is Your Algorithm Dangerous?

Much has been written about artificial intelligence, both from the perspective of possibilities and opportunities as well as from the perspective of risks and limitations. Here I make a simple point — evaluating the appropriateness of an algorithm requires understanding the domain space in which it will operate. While data science enables one to transcend expertise in a particular domain, it nevertheless requires a deep familiarity with the question it is required to answer. Focusing on the answer rather than the question presents significant dangers. These are not necessarily physical hazards, but rather dangers to things like social norms, rule of law values, and the experience of equality. Deploying algorithms that do not avoid these dangers risks injustice in individual cases as well as generating longer term threats to fundamental social and democratic values.

The machine not only needs to learn to answer a question, but to answer the right question.

Consider the context of criminal justice. Risk assessment tools are increasingly used, particularly in the United States, to make decisions about bail, parole, and sentencing. (For a recent example, see [1].) Such tools, based on learning from historic data sets and offender surveys, raise issues even within the discipline of data science. For example, a ProPublica investigation [2] found bias against African Americans, demonstrating that there was a higher rate of false positives among this subpopulation. Given the importance of non-discrimination as a social principle, differential impact should be a question for evaluation alongside accuracy and precision measures. This can be done entirely within the discipline of data science.

However, there is a deeper question about the use of data-driven decision-making in criminal justice that goes beyond metrics (however widely cast). This relates to the need for the machine not only to learn how to answer a question, but to answer the right question. This in turns requires an understanding of the *nature* and *purposes* of particular decisions. To take a simple example, would it be appropriate for a justice system to require someone to spend more time in jail (via bail, sentencing or parole decisions) because they have large feet, even on the assumption that people with large feet are historically more likely to commit a crime? This would seem to be the assumption employed in machine learning; quoting from an article comparing methods of forecasting recidivism to inform parole decisions: "For example, if other things equal, shoe size is a useful predictor of recidivism, then it can be included as a predictor. Why shoe size matters is immaterial" [3]. The authors are of course correct that pure prediction (complete with testing of accuracy and precision) does not require explanation. However, when I have explained how these kinds of tools work to judges, they are generally disturbed by the suggestion that this would be a reasonable basis for making a decision that affected an individual's liberty. Even where "dangerousness" is relevant, judges and data scientists have different ideas as to what kinds of factors are relevant to that assessment. Deploying a purely predictive algorithm changes rather than replicates the way that these decisions are made, with important consequences for justice, fairness, and due process.

Digital Object Identifier 10.1109/MTS.2018.2857618
Date of publication: 31 August 2018

Outside the context of crime, there are numerous contexts where it is important to understand the nature of the decision being made. For example, should universities admit students based on how demographic and other data correlates with the university's records of successful graduates and alumni? Not only would such a policy be regressive, it would resemble Gattaca's dystopia where innate characteristics rather than our actions and performance take precedence in determining our futures. This is problematic *even if* some of those innate characteristics are better predictors of future performance than past performance.

So — how does one ensure that tools such as machine learning do not displace important social values? For those writing the program, it is important to understand the task — is it purely about prediction or are there other ethical, social, or institutional factors that come into play? Conversely, it is important to communicate clearly up the chain — explaining inferences and assumptions to those relying on outputs in decision-making. Evaluation is also crucial — not only for accuracy but also for impact. Understanding impact requires broader reflection on the breadth of potential consequences and the likelihood and severity of potential harms. This has been a traditional function of technology assessment (as practiced by the now defunct Office of Technology Assessment in the U.S., or members of the European Parliamentary Technology Assessment network), but it needs to move beyond policy advice to practical ethics within organizations.

More broadly, policymakers and educators should be concerned about digital and algorithmic literacy across the population. Serious thought is needed as to what kinds of decisions can be delegated to what kinds of automated processes, not just within government, but also as a matter for public debate. Where accountability matters for human decision-making, this needs to be preserved in any move towards algorithms. These are not easy demands, but they are the best route for ensuring that algorithms are fit-for-purpose.

While data science is universalist allowing inferences to be drawn from any dataset, its use in decision-making requires an understanding of context. Predictive accuracy is not the only value that will be relevant, as can be seen from the examples of criminal justice and university admissions. A failure to acknowledge this will cause harms, at the level of both individuals subject to decisions and the broader social fabric. To avoid this, broad literacy in data science is needed to facilitate enhanced interdisciplinarity, appropriate deployment, and comprehensive evaluation.

How do we ensure that tools such as machine learning do not displace important social values?

Author Information

Lyria Bennett Moses is Associate Professor and Director of the Allens Hub for Technology, Law and Innovation at UNSW Law, Sydney, NSW 2052, Australia. Email: lyria@unsw.edu.au.

References

[1] *Wisconsin v Loomis*, 881 NW 2d 749 (Wis, 2016).

[2] J. Angwin *et al.*, "Machine bias," *ProPublica*, May 23, 2016; https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] R.A. Berk and J. Bleich, "Statistical procedures for forecasting criminal behavior: A comparative assessment," *American Society of Criminology*, vol. 12, no. 3, pp. 513-544, 2013.