POLICY FORUM

MACHINE LEARNING

“Explaining” machine learning reveals policy challenges

The need to make objectives explicit may expose policy trade-offs that had previously been implicit and obscured

By **Diane Coyle**[1] and **Adrian Weller**[2,3]

There is a growing demand to be able to “explain” machine learning (ML) systems’ decisions and actions to human users, particularly when used in contexts where decisions have substantial implications for those affected and where there is a requirement for political accountability or legal compliance (*1*). Explainability is often discussed as a technical challenge in designing ML systems and decision procedures, to improve understanding of what is typically a “black box” phenomenon. But some of the most difficult challenges are nontechnical and raise questions about the broader accountability of organizations using ML in their decision-making. One reason for this is that many decisions by ML systems may exhibit bias, as systemic biases in society lead to biases in data used by the systems (*2*). But there is another reason, less widely appreciated. Because the quantities that ML systems seek to optimize have to be specified by their users**,** explainable ML will force policy-makers to be more explicit about their objectives, and thus about their values and political choices, exposing policy trade-offs that may have previously only been implicit and obscured. As the use of ML in policy spreads, there may have to be public debate that makes explicit the value judgments or weights to be used. Merely technical approaches to “explaining” ML will often only be effective if the systems are deployed by trustworthy and accountable organizations.

The promise of ML is that it could lead to better decisions, yet concerns have been raised about its use in policy contexts such as criminal justice and policing. A fundamental element of the demand for explainability is for explanation of what the system is “trying to achieve.” Most policy decision-making makes extensive use of constructive ambiguity to pursue shared objectives with sufficient political consensus. There is thus a tension between political or policy decisions, which trade off multiple (often incommensurable) aims and interests, and ML, typically a utilitarian maximizer of what is ultimately a single quantity and which typically entails explicit weighting of decision criteria.

Will the demand for explainable ML systems for decisions in the justice system require more explanations from policy-makers, too?

PHOTO: THE WOLF LAW LIBRARY/FLICKR/CC BY-NC-ND

We focus on public policy decision-making using ML algorithms that learn the relationships between data inputs and decision outputs. As a first step, policy-makers need to decide among a number of possible meanings of explainability. These range from causal accounts and post hoc interpretations of decisions (*3*) to assurance that outcomes are reliable or fair in terms of the specified objectives for the system (*4*). For example, the explainability requirements for ML systems used by local authorities to determine benefit payments will differ greatly from those required for the enforcement of competition policy with respect to pricing by online merchants. Each of the specific meanings of explainability has different technical requirements, which will imply choices about where efficiency and cost might need to be sacrificed to deliver both explainability and the desired outcomes. Choosing which meaning is relevant is far from a technical question (though what can be provided depends on what is technically feasible). Thus, those seeking explainability will need to specify, in terms translatable to how ML systems operate, what exactly they mean, and what kind of evidence would satisfy their demand (*5*). It must also be possible to monitor whatever explanations are provided, and there must be practical methods to enforce compliance.

Furthermore, policy institutions starting to deploy algorithmic or ML-based decision systems, such as the police, courts, and government agencies, are operating in the context of declining trust in some aspects of public life. This context is important for understanding demands for explainability, as these may in part reflect broader legitimacy demands of the policy-making process. If an organization is not trusted, its automated decision procedures will likely also be distrusted. This implies a broader need for trustworthy processes and institutions, for “intelligent accountability” as the result of informed and independent scrutiny, communicated clearly to the public (*6*). Satisfying the demand for explainability implies testing the trustworthiness of the organizations using ML systems to make decisions affecting individuals. Evaluation requires comparing outcomes against a benchmark, which can be the baseline situation, or a specified desired outcome.

Taking the demand for explainability as a demand for accountability, the promise of ML is that it could lead to more legitimate and better decisions than humans can make, on some measure. Potential benefits are clearly demonstrable in some forms of medical diagnosis (*7*) or monitoring attempted financial fraud (*8*). In these domains, there is general agreement on a straightforward quantity to optimize, and the incentives of principals (citizens or customers) and agents (public or corporate decision-makers) are aligned. Public concern about the use of ML focuses on other domains, such as marketing or policing, where there may be less agreement about (or trust in) the aim of either the ML system or the organization using it.

[1] Bennett Institute for Public Policy, University of Cambridge, Cambridge, UK. [2] University of Cambridge, UK. [3] The Alan Turing Institute, London, UK. Email: dc700@cam.ac.uk; aw665@cam.ac.uk

These concerns highlight a key challenge posed by the use of ML in policy decisions, which is that ML processes are almost always set up to optimize an objective function; this optimization goal can be described in anthropomorphic terms as the “intention” of the system. Yet there is often little or no explicit discussion by policy-makers when considering using ML systems about what conflicting goals, benefits, and risks may trade off against each other as a result. One reason for this is that it is inherently challenging to specify a concrete objective function in sociopolitical domains (*9*). For example, like current ML systems, economists’ decisions are informed by estimates of statistical relationships between directly observable and unobservable variables, derived from data generated by a complex environment. Yet economic policies such as tax changes often fail to take into account all relevant factors in the decision environment, or likely behavior changes, in specifying the objective function (*10*). The use of ML systems in other policy contexts will expand the scope of such unintended consequences.

Given that the dominant paradigm of machine learning is based on optimization, the use of ML in policy decisions thus speaks to a fundamental debate about social welfare. From the perspective of ethical theories, ML is largely consequentialist: A machine system is configured on the basis of its ability to achieve a desired outcome. Conventional policy analysis is similarly typically based on consequentialist economic social welfare criteria. The well-known impossibility theorems in social choice theory (*11*) establish that when the goal is to aggregate individual choices under a set of reasonable social decision rules, it is impossible to satisfy a set of desirable criteria simultaneously, and thus impossible to achieve a set of desired outcomes by optimizing a single quantity. Critics of consequentialist economic policy analysis argue that people have multidimensional, probably incommensurable, and possibly contradictory objectives, so that imposing utilitarian decision-making procedures will conflict both with reality and with ethical intuitions (*12*).

Nevertheless, policy choices are made, so there has always been an unavoidable, albeit often implicit, trade-off or weighting of different objectives (*12*). For example, cost-benefit analysis can incorporate environmental and cultural, as well as financial, considerations, but converts all of these into monetary values. Any choice made when there are multiple interests or trade-offs will imply weights on the different components. As these trade-offs are codified into ML objective functions, the weights given to competing objectives comprise a first-line characterization of how conflicts will be resolved. Using ML systems in political contexts is extending the use of optimization; progress in making these ML systems more understandable to policy-makers will make the de facto choices between competing objectives more explicit than they have been previously (*13*). Greater explainability is therefore likely to have to lead to a more explicit political, not wholly technical, debate.

Distilling concrete, unambiguous objectives in this way may turn out to be extremely challenging, for ambiguity about objectives is often useful in policy-making precisely because it blurs uncomfortable conflicts of interest. In many domains, policies generally emerge as a pragmatic compromise between fundamentally conflicting aims. For example, people who disagree about whether the justice system should be retributive or rehabilitative may well be able to agree on specific sentencing policies. Such incompletely theorized agreements “Play an important function in any well-functioning democracy consisting of a heterogeneous population” (*14*, p. 1738). The omission of discussion of ultimate aims can make it easier to achieve consensus on difficult issues. As there is some (limited) scope to interpret means to achieve the objective with flexibility, the “weighting” of different fundamental aims remains implicit, and diverse political communities can make progress.

An optimistic conclusion would be that being forced by the use of ML systems to be more explicit about policy objectives could promote useful debate leading in the long run to more considered outcomes. ML systems can be used to explore choices and outcomes on different counterfactual high-level objectives, such as retribution or rehabilitation in justice, enabling considered human judgments. However, it may in practice be impossible to specify what we collectively truly want in rigid code. For example, many local governments do not seem to be engaging in public consultation when they adopt predictive ML systems, such as to flag “troubled” families that are likely to need interventions. Although steps such as explicitly adding uncertainty to the ML objective might address this challenge of imperfectly specified objectives in future, ML systems are unable at present to offer wisely moderated solutions to ambiguous objectives (*15*).

Human decision-makers can make use of common sense or tacit knowledge, and often override decisions indicated by an economic model or other formal policy analysis, and they will be able to do the same when assisted by ML. Yet, demanding that ML systems be explainable is likely to make the trade-offs between different objectives far more explicit than has been the norm previously. Ultimately, the use of explainable ML systems in the public sector will make a broader debate about social objectives and social justice newly salient. Providing explanations requires being transparent about the systems’ objectives — forcing clarity about choices and trade-offs previously often made implicitly — and how their predictions or decisions draw on patterns revealed by a fundamentally biased social and institutional system. Moreover, whereas democratic political systems often look to resolve conflicts through constructive ambiguity—or in other words, the failure to explain—ML systems may require ambiguous objectives to be resolved unequivocally. So, although the need for explainability certainly poses technical challenges, it poses political challenges too, which have not to date been widely acknowledged. Yet, the increasing scope of ML, and progress in delivering explainability, in politically salient areas of policy could shine a helpful spotlight on the conflicting aims and the implicit trade-offs in policy decisions, just as it already has on the biases in existing social and economic systems. ■

REFERENCES AND NOTES

1. B. Dattner, T. Chamorro-Premuzic, R. Buchband, L. Schittler, *The legal and ethical implications of using AI in hiring*, *Harv. Bus. Rev*. April, 25 (2019); https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring.
2. R. Richardson, J. Schultz, K. Crawford, *New York Univ. Law Rev.* **192**, 204 (2019).
3. Z. Lipton, The mythos of model interpretability (2017); https://arxiv.org/pdf/1606.03490.pdf .
4. T. Miller, Explanation in artificial intelligence: Insights from the social sciences (2018); https://arxiv.org/pdf/1706.07269.pdf.
5. P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence (2019); https://arxiv.org/pdf/1903.02409.pdf.

6 O. O’Neill, *Int. J. Philos. Stud.* **26**, 293 (2018).

7. J. De Fauw *et al.*, *Nat. Med*. **24**, 1342 (2018).
8. S. Aziz, M. Dowling, in *Disrupting Finance: FinTech and Strategy in the 21st Century*, T. Lynn, G. Mooney, P. Rosati, M. Cummins, Eds. (Palgrave, 2019), pp. 33–50.
9. P. Samuelson, *Foundations of Economic Analysis* (Harvard University Press, 1979), chap. 8, pp. 203–252.
10. J. Le Grand, *Br. J. Polit. Sci*. **21**, 423 (1991).
11. A. Sen, *Am. Econ. Rev*. **89**, 349 (1999).
12. E. Anderson, *Value in Ethics and Economics* (Harvard Univ. Press, 1993).
13. S. Grover, C. Pulice, G. I. Simari, V. S. Subrahmanian, *IEEE Trans. Comput. Soc. Syst.* **6**, 350 (2019).
14. C. R. Sunstein, *Harv. Law Rev*. **108**, 1733 (1995).
15. M. Hildebrandt, *Smart Technologies and the End(s) of Law* (Edward Elgar, 2016).

ACKNOWLEDGMENTS

We are grateful to M. Kenny and N. Rabinowitz for helpful comments. A.W. acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grants EP/N510129/1 and TU/B/000074, the Leverhulme Trust via CFI, and the Centre for Data Ethics and Innovation.

10.1126/science.aba9647

Science

"Explaining" machine learning reveals policy challenges

Diane Coyle and Adrian Weller

Science **368** (6498), 1433-1434.
DOI: 10.1126/science.aba9647

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.