Canadian Journal of Philosophy (2022), 52: 1, 1–5
doi:10.1017/can.2022.28

ARTICLE

The Political Philosophy of Data and AI

Annette Zimmermann[1,2], Kate Vredenburgh[3] and Seth Lazar[4]

[1]Department of Philosophy, University of York, York, UK, [2]Carr Center for Human Rights Policy, Harvard University, Cambridge, MA, USA, [3]Department of Philosophy, Logic, and Scientific Method, London School of Economics, London, UK, and [4]Department of Philosophy, Australian National University, Canberra, Australia
Corresponding author: Annette Zimmermann. Email: annette.zimmermann@york.ac.uk

We are increasingly subject to the power of technological systems relying on big data and AI. These systems are reshaping the welfare state and the administration of criminal justice. They are used to police tax evasion, track down child abusers, and model the spread of the pandemic. And they are used to weaponize vast surveillance networks through facial recognition technology. But algorithmic power extends far beyond the state: we spend ever more time working, socialising, and consuming within digital platforms. Our experiences are governed by algorithms that are constantly monitoring and shaping our behaviour and our attention, automatically selecting what we do and do not see. These online experiences have offline consequences, among them an unprecedented challenge to democratic processes worldwide.

There is a thriving literature in other disciplines on the legal and political implications of big data and AI, as well as a rapidly growing literature within philosophy concerning ethical problems surrounding AI. There is relatively little work to date, however, from the perspective of political philosophy. This special issue was borne out of a recognition that political philosophy has a crucial role to play in conversations about how AI ought to reshape our joint political, social, and economic life. The widespread deployment of AI calls attention to fundamental, long-standing problems in political philosophy with renewed urgency, and creates genuinely new philosophical problems of political significance. Existing philosophical problems resurface at an unprecedented scale, such as the question of whether some rule-based decision-making procedure—whether that procedure is implemented by bureaucrats and administrative officials or by algorithmic systems—is just and legitimate, or the question of whether and when making judgments based on statistical generalisations is morally permissible. Other examples include long-standing debates in moral and legal philosophy on why discrimination is wrong; classic debates in political philosophy on political equality in light of unequal political influence; debates in political philosophy and the philosophy of economics on work and alienation; and wider debates on idealization and abstraction that cut across political and moral philosophy as well as the philosophy of science.

New philosophical problems emerge, too: for example, we cannot simply draw on established theories of privacy originally designed for small-scale eavesdropping scenarios or for the physical environment of the panoptic prison, and drag-and-drop them into a world of vast, decentralised mutual surveillance—a world in which the most intimate secrets can be inferred from people's digital footprints in order to shape 'dark patterns' that nudge consumers and voters into making particular choices. We cannot always invoke existing concepts of legitimacy and political obligation that locate political authority firmly within the power monopolies held by idealised nation-states governed by democratically authorised public officials, while disregarding the fact that technological change has created significant power and wealth oligopolies controlled by unregulated and democratically unauthorised corporations. And we cannot adequately formulate a theory of the

appropriate allocation of collective attention by relying on long-standing free-speech arguments that presuppose a long-gone media and communications environment.

* * *

To understand how big data and AI are used to exercise power, how they might undermine or promote justice, and how they are reshaping social structures and the human agents within them, we must understand the underlying technologies. In particular, during a time in which new technologies are being met with both undue optimism and undue pessimism in public discourse, philosophers must be careful to resist fatalistic, evangelising, or otherwise misguided narratives about what AI can and cannot realistically do here and now.

To this end, political philosophers ought to engage closely with cutting-edge work in computer science in order to develop a technologically sound understanding of *why* particular problems of political and moral significance arise across various AI deployment domains, and of how contemporary technical work has attempted to solve those problems. In addition, a constructive engagement with interdisciplinary, empirically grounded scholarship on the interplay of technology and society from law, political science, science and technology studies, and other fields is crucial. To understand the politics of AI we must understand not only its technological foundations, but also the sociotechnical systems of which AI is part. This ensures that we remain focused on real problems raised by AI, and that our work can feed into practically implementable solutions to those problems.

Overall, political philosophers have much to contribute to, but also much to gain, from a close exchange with relevant work in a wide range of other disciplines focused on political and moral problems associated with AI and big data: for instance, when it comes to evaluating whether purely technical interventions are sufficient (or indeed necessary) for ameliorating existing social and institutional structures in a particular domain—or whether solutions that move beyond improving existing technological tools themselves might better protect the equal freedom of all those subject to such tools. Importantly, philosophical contributions to these questions push forward not only the academic debate in computer science and other disciplines, but also that within philosophy itself: innovative work on the philosophy of AI can motivate and productively address pressing philosophical debates on how existing conceptual and normative frameworks might be transformed, extended, and questioned in light of recent technological innovations. Furthermore, this type of work has the potential to articulate bold, novel solutions to (re-)emerging political problems in an age of automation, while building on a clear-sighted assessment of what is and is not technologically feasible.

By fostering constructive exchange with existing work in disciplines outside of philosophy while maintaining a clear focus on distinctly philosophical problems, the papers in this special issue aim to articulate at least three things: first, a clear diagnosis of what is wrong with our existing social structures and the technological tools increasingly deployed within them; second, a conceptual analysis of what it means for systems—both technological and social—to be just and democratically legitimate; and third, action-guiding normative arguments about how to intervene upon technological and social realities in order to better approximate central political and moral goals.

* * *

This special issue brings together philosophers whose work is not only technologically and empirically informed, but also motivated by the aim of making original philosophical contributions that address problems of political and moral urgency. Our primary focus is on moral questions raised by contemporary and near-future uses of AI and other data-driven technologies interacting with complex social, political, and economic structures.

The first group of papers take the recent explosion of technical discussions of fairness in algorithmic decision-making as a springboard to make first-order philosophical progress on questions of justice and fairness. Two of the papers in this special issue raise fundamental questions

https://doi.org/10.1017/can.2022.28 Published online by Cambridge University Press

about how to design fair distributional procedures under conditions of uncertainty and structural injustice. Two others teach us general lessons about the fair distribution of benefits and burdens through situated moral reasoning about how AI has changed decision-making about punishment and fundamental goods such as credit or employment. They do so by examining the problem of procedural injustice in algorithmic decision-making, the moral significance of large-scale algorithmic arbitrariness, the limitations of ideal theorising about machine learning systems, and the right to be treated as an individual in the context of algorithmic decisions.

The second group of papers explore the new power relations created by data and AI in the workplace, on the tech platforms that shape our online and offline lives, and by the state. They raise moral concerns about the exercise of power through the use of artificial intelligence to deliver targeted interventions that aim to shape individuals' behaviours. These exercises of power rely on a social and economic environment permeated by surveillance mechanisms that turn individual actions into behavioural data. The commercialization of these behavioural data power the AI-infused economy, especially social media; and the products developed in the private sector can be repurposed by the state to surveil those under its purview, threatening free and equal societies governed by coercive states. These papers help to start overdue debates in political philosophy on the noninstrumental value of explanations, the circumstances under which the extraction of personal data can be permissible given the practical infeasibility of informed consent in the age of big data, the consequences of algorithmic manipulation for the quality and integrity of democratic processes, and the extent to which large-scale systems of influence and observation enabled by the rise of personalised advertising constitute illegitimate exercises of power.

1.a Zimmermann and Lee-Stronach

Algorithmic decision-making often leads to *substantively* unjust outcomes. Much recent scholarship in philosophy and computer science has analysed this problem, while seeming to assume that substantively unjust algorithmic decision-making might be at least *procedurally* just. Zimmermann and Lee-Stronach critique the latter assumption, arguing that deference to algorithmic outputs is procedurally unjust in contexts involving background conditions of structural injustice. Under nonideal conditions, algorithmic systems, if left unchecked, cannot meet a necessary condition of procedural justice: they fail to deliver a sufficiently nuanced picture of which cases count as relevantly similar given that background structures shape similarity and difference in complex ways. Human agents relying uncritically and entirely on algorithmic outputs risk prematurely adopting beliefs about decision subjects, and thus committing *doxastic negligence*, the authors argue. This poses a distinct, underexplored philosophical problem that escapes a purely technological solution. Resolving this problem requires that human agents exercise their unique deliberative capacities cautiously—for instance, by suspending belief and gathering additional information—when assessing if an algorithmic system truly treats like cases alike in a structurally unjust world.

1.b Hellman and Creel

Algorithms can be valuable decision aids when they increase a decision-maker's ability to predict an outcome of interest or to classify individuals. However, this increase in accuracy is sometimes bought at the cost of more *arbitrary* decision-making. Many have the strong intuition that such arbitrary decision-making is morally defective. In their contribution to this volume, Creel and Hellman argue that arbitrary decision-making is not morally objectionable on the individual level. Instead, they argue, it is morally problematic when and because it results in *standardisation* that leads to social exclusion, as individuals are prevented from accessing important opportunities because they do not qualify according to a common set of decision criteria. Troublingly, standardisation is likely to result from algorithmic decision-making used for economic and technical reasons. There are economic incentives to use a small number of existing data sets; that plus the

https://doi.org/10.1017/can.2022.28 Published online by Cambridge University Press

technical problem of overfitting leads to models that exploit similar, arbitrary features and correlations for prediction and classification. Creel and Hellman conclude by arguing that reducing systematicity is a more promising strategy than reducing arbitrariness to address the tendency of AI-powered decision systems to exacerbate social exclusion.

1.c Fazelpour, Lipton, and Danks

Fazelpour, Lipton, and Danks's contribution to this volume poses a methodological critique of current approaches to research in so-called "fair machine learning," which aims to widen the range of values that guide algorithm design to fairness and justice. Current approaches in fair machine learning measure the fairness or justness of a system and set the just or fair target state using evaluation metrics that operate on static, local datasets. This approach, argue the authors, is flawed because it ignores interdependencies among actors and the social and institutional context, and fails to take uncertainty into account, both of which prevent the approach from designing successful interventions to realize the target state. However, these problems with static evaluation should not lead us to the view that one cannot evaluate decision procedures in terms of their (predicted) consequences. Instead, the authors argue, we can and should evaluate so-called "dynamic trajectories" in terms of good-making properties such as robustness and apt representational choices.

1.d Jorgensen

Jorgensen's contribution to the special issue examines moral questions raised by the use of actuarial inference in the criminal justice system, or inference about a person based on people that are similar to them. Risk-assessment tools, which have a long history in the US criminal justice system, aim to reduce the influence of human cognitive bias on pre- and posttrial decision-making through more accurate predictions and classifications. But this topic has gained new moral urgency with the introduction of a new generation of tools that aim to improve prediction performance by learning a model using existing datasets. Since these tools make actuarial inferences, however, they seem to violate the right to be treated as an individual. Jorgensen argues for a particular understanding of that moral concern—namely, that actuarial inference can violate the right to an individualized judgment. This right is grounded in agents' claims to a fair distribution of the burdens and benefits of the rule of law. Jorgensen concludes by drawing out the implications of the right for this new generation of risk-assessment tools: predictors must be transparent, the features that are the basis for prediction must be subject to agential control, and the burdens imposed by using a feature as a predictor must be outweighed by the benefits to those individuals.

2.a Vredenburgh

Opaque algorithms—algorithms whose outputs are not understood, and perhaps not capable of being understood, by affected parties—are increasingly used to structure the workplace. Their opacity raises well-charted instrumental concerns that opaque algorithmic decision-making may be unfair or cause unjustifiable harm. In this volume, Vredenburgh examines a less-discussed question: whether opaque algorithmic decision-making in the workplace is objectionable in and of itself. She argues that the opacity is bad in and of itself when and because it makes workers (subjectively) unfree. And it does so because it prevents them from developing a practical orientation, or a reflective attitude and way of relating to one's social world, that is rooted in a grasp of normative explanations of their workplace and the economy. So-called "technical opacity" is the least troubling mechanism that leads to the systematic unavailability of the required normative explanations. Instead, it is the greater control over the flow of information that algorithms grant managers, and the attendant isolation and loss of control of workers, that is the greater source of subjective unfreedom for workers.

https://doi.org/10.1017/can.2022.28 Published online by Cambridge University Press

2.b Voorhoeve and Wolmarans

Social networking services provide services in return for access to users' personal data, which is the commercial foundation of new products. Scholarship in law, philosophy, and media and communications studies has been vexed by the question of whether extensive data sharing is compatible with the rights of users and citizens. Consent is broadly recognized as incapable of providing a moral foundation of the permissibility of data sharing in the digital age, but compelling alternative frameworks are lacking. Wolmarans and Voorhoeve draw on the work of T. M. Scanlon to argue that sharing personal data with social media companies is permissible if individuals had sufficiently valuable opportunities to make choices and avoid harms. The value of those opportunities, they argue, must be assessed in light of users' different decision-making capabilities, and third parties. The authors conclude by contrasting the differing regulatory requirements suggested by consent-based accounts and the value of choice account.

2.c Christiano

Christiano's contribution tackles the question of whether advances in AI threaten democracy. He argues that AI-powered communication poses a pressing threat to political equality. The argument builds on the intuition that the power to uptake and disseminate information matters a great deal for advancing one's interests and one's conception of the common good. Artificial intelligence affords public and private actors greater scope for manipulating their audiences through mechanisms such as microtargeting, hypernudging, and hyperspecialization. Such manipulation threatens to undermine political equality in societies where some citizens are more vulnerable to manipulation. Christiano argues that modern society is such a society since a high cognitive dependence on one's social environment for information is combined with an inequality in the reliability of information received from one's network. Artificial intelligence thus poses a serious threat to political equality by creating conditions of unequal informational power.

2.d Benn and Lazar

Benn and Lazar offer a moral evaluation of Automated Influence—the use of AI to collect, integrate, and analyse people's data in order to deliver targeted interventions that shape their behaviour. Automated Influence has been challenged for violating individuals' privacy, exploiting them, and manipulating them, but Benn and Lazar show that in each case individualist versions of these objections carry relatively little weight. Instead of focusing on the claims and rights of individuals, we can best understand the moral qualms many of us have about Automated Influence by considering its structural, collective impacts—how even where it does not itself involve violating privacy, it creates structural incentives for those who gather behavioural data to do so. Even if we as individuals are not meaningfully exploited, the asymmetries of information enabled by Automated Influence mean that we are collectively exploited by those who hold the power of big data; and even if we as individuals are unlikely to be manipulated, AI enables stochastic manipulation of populations at large.

Funding statement. Seth Lazar's work on this special issue was supported by ARC grant numbers FT210100724 and DP170101394.

Cite this article: Zimmermann, A. Vredenburgh, K. and Lazar, S. 2022. The Political Philosophy of Data and AI. *Canadian Journal of Philosophy* 52: 1–5, doi:10.1017/can.2022.28