9

Conclusions and Caveats

It is fitting that the last example we introduced in the book was about the Internet Research Agency's (IRA) use of social media, analytics, and recommendation systems to wage disinformation campaigns and sow anger and social discord on the ground. At first glance, it seems odd to think of that as primarily an issue of technology. Disinformation campaigns are ancient, after all; the IRA's tactics are old wine in new boxes. That, however, is the point. What matters most is not particular features of technologies. Rather, it is how a range of technologies affect things of value in overlapping ways. The core thesis of our book is that understanding the moral salience of algorithmic decision systems requires understanding how such systems relate to an important value, viz., persons' autonomy. Hence, the primary through line of the book is the value itself, and we have organized it to emphasize distinct facets of autonomy and used algorithmic systems as case studies.

To review, we have argued that three broad facets of autonomy are affected by algorithmic systems. First, algorithmic systems are relevant to what we owe each other as autonomous agents. That is the focus of Chapters 3 and 4. In Chapter 3 we addressed the material conditions that we owe others and argued that respecting people as autonomous demands that any algorithmic system they are subjected to must be one that they can reasonably endorse. It does not require that they value particular outcomes or that they not be made worse off by such systems. Rather, systems must either comport with agents' own ends or be consistent with fair terms of social cooperation. We argued that persons being able to reasonably endorse a system turns on the system's reliability, responsibility, stakes, and relative burden. Chapter 4 turned to the issues of what information we owe others. There we argued that people are owed information as a function of their practical agency (i.e., their ability to act and carry out plans in accord with their values) and as a function of their cognitive agency (i.e., their ability to exercise evaluative control over mental states, including beliefs, desires, and reactive responses). We offered several principles for information access grounded in agency.

The second connection between algorithmic systems and autonomy is ensuring the conditions under which people are autonomous. In Chapter 5 we considered the relationship between algorithmic systems and freedom. We explained that algorithms

https://doi.org/10.1017/9781108895057.009 Published online by Cambridge University Press

bear upon negative, positive, and republican freedom and offered a general account of freedom as ecological non-domination. Key to understanding that ecology is recognizing three key challenges to freedom: affective challenges, deliberative challenges, and social challenges. In Chapter 6 we offered some suggestions for addressing some facets of those challenges. Specifically, we argue that a kind of epistemic paternalism is both permissible and (under some conditions) obligatory.

Chapters 7 and 8 shift focus to the responsibilities *of* agents in light of the fact that they are autonomous. In Chapter 7 we argue that algorithmic systems allow agents deploying such systems to undermine a key component of responsibility, viz., providing an account for actions for which they are responsible. Specifically, we argue that complex systems create an opportunity for "agency laundering," which involves a failure to meet one's moral responsibility for an outcome by attributing causal responsibility to another person, group, process, or technology. Chapter 8 addresses a different facet of responsibility. Citizens within democratic states have a responsibility to exercise their autonomy in order to legitimate political authority. That is, they have a responsibility to help ensure that governments, laws, policies, and practices are justifiable. However, some kinds of algorithmic systems hinder citizens' abilities to do that. They can do so by undermining the epistemic conditions necessary to underwrite the "normative authority" path to legitimacy or by undermining the exercise of autonomy necessary to underwrite the "democratic will" path to legitimacy.

9.1 FURTHER WORK

In one sense, that is a lot of terrain to have covered. And yet even within the scope of autonomy and algorithmic systems, there is much more work to do. Throughout the book, we pause to point out how various topics bear upon one another. There are, however, connections across the chapters that warrant more attention.

In Chapter 4 we address informational components to autonomy, and we argue that people have claims to information about algorithmic systems based on practical agency, cognitive agency, and democratic agency. There is a question, though, about whether such information is a condition for people to be able to reasonably endorse systems. That is, the precise relationship between what we owe people materially (per Chapter 3 and the Reasonable Endorsement Test) and what we owe people informationally (per the principles of informed practical and cognitive agency) is worth examining. Similar concerns arise in understanding the relationship between practical, cognitive, and democratic agency and political legitimacy. We note in Chapters 4 and 8 that the ability to exercise democratic agency is a component of the legitimating process. We explain how that relationship functions within the normative authority and democratic will "arms" of legitimacy. But a number of questions remain. Just what kinds of processes are subject to legitimation at all? Certainly, direct actions of government agents can be legitimate or not, but what about private actions? Or what about private actors whose influence on

state affairs is enormous? Moreover, what is the extent of information necessary for citizens to fulfill their legitimating responsibilities?

There are further connections to be drawn between Chapter 5's discussion of quality of agency and other facets of autonomy. To the extent that the challenges to freedom limit people's quality of agency (and hence, positive freedom), are they also limitations on people's ability to reasonably endorse states of affairs? It also seems plausible that such challenges are an impediment to exercising practical, cognitive, and democratic agency. It is therefore worth exploring whether even greater epistemically paternalistic actions are justifiable (or even obligatory) than those we outline in Chapter 6.

We should also point out that the relevance of agency laundering may be even broader than we outline in Chapter 7. Laundering may be applicable in other cases we discuss throughout the book. For example, it would be worth considering it in background checks (Chapter 4) and predictive policing (Chapter 8). When presenting on the topic of agency laundering to academic audiences, we have often received questions about whether it could be applied to political actions (e.g., Brexit). While we cannot do justice to that question here, we can say that the use of sophisticated profiling and influence operations is a plausible mechanism for laundering. Hence, examining influence campaigns as potential sites of laundering is worthwhile.

And moving beyond the topics we've covered, the range of open questions is vast. Driverless vehicles, for example, raise numerous issues with respect to responsibility, human control, and worker displacement. Robots, including those that interact with humans, provide care and companionship, and displace labor are a topic of growing philosophical and moral concern. Weapons of war raise numerous issues relevant to human control and responsibility.

9.2 CAVEATS: BASELINE ISSUES

An important question that one might raise about this project concerns baseline comparisons. So, while COMPAS, EVAAS, TVAAS, PredPol, and other systems may have important problems with respect to autonomy, one might argue that despite those issues, the technologies are better than the relevant alternatives. What matters is not that algorithmic systems have flaws compared to some ideal, but whether they are meaningfully better than relevant alternatives. Having a system like COMPAS that assesses risk may be better than humans, who have well-known biases, who act arbitrarily, and who are harder to audit.

That's a reasonable criticism. However, it does not undercut the project for a number of reasons. First, even if it is the case that algorithmic systems are better than systems that precede them, it does not follow that they are justifiable. So, even if it is the case that using COMPAS is better than judges at determining likelihood of reoffense, it does not follow that use of COMPAS is itself justifiable. The space of reasonable alternatives need not be some antecedent system and some novel

algorithmic system. There could be better algorithms; lower-stakes algorithms; algorithms that do not impose disparate relative burdens; that respect practical, cognitive, and democratic agency; that preserve quality of agency; that do not serve to launder agency; that allow citizens to fulfill their responsibilities of legitimation; and so forth.

Second, even where technologies are better than some alternatives, they may reveal underlying moral concerns. Consider a study from October 2019.[1] Obermeyer et al. studied a machine learning system that examined health records in order to predict which patients were in "high risk" categories and either defaulted them into a care program or referred them for screening into a care program. The study determined that Black patients identified as high risk were significantly less healthy than White patients so-identified; that is, Black patients had higher numbers of chronic conditions than similarly categorized White patients. This entailed Black patients were less likely to receive appropriate care than White patients. The reason for the difference, according to the researchers, is that the machine learning algorithm reflected health-care expenditures. That is, risk levels were correlated with the amount of care and treatment that patients received in prior years. As the authors put it, "Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities. As a result, accurate prediction of costs necessarily means being racially biased on health."[2] That disparity may arise for a couple of reasons. One is that poor patients (who are disproportionately Black) face substantial barriers to receiving healthcare, even if they are insured. For example, transportation time to care facilities may be greater and they may face greater difficulty in getting time off of work. There are also social barriers, including worse treatment of Black patients by care providers and distrust of health-care providers.

Notice, though, that the algorithm at work in the Obermeyer study is likely better than a system that did not use sophisticated technologies to identify high-risk patients and nudge them toward care programs. That is, it was likely better for Black patients as well as White patients and worse for no one. It was just *more* advantageous for White patients. But there is nonetheless a moral problem with the algorithm. Hence, "better than a prior baseline" should not end our inquiries. In the health system case, the study's authors developed a different analysis that sorted patients into high-risk groups based on underlying health conditions, and it performed similarly well for Black and for White patients. Our conclusions in this book, we hope, can provide grounds for analyzing what better algorithmic systems would look like.

9.3 BIGGER PICTURES

A further objection to many of our arguments in the book will have to do with background conditions and structures and whether we have properly identified the

[1] Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations."
[2] Obermeyer et al., 450.

https://doi.org/10.1017/9781108895057.009 Published online by Cambridge University Press

most morally salient features of cases. In the case of COMPAS and *Loomis*, one might argue that the US criminal justice system has so many infirmities that focusing on use of a single tool is beside the point. In *Loomis*, one could argue that plea deals generally are unjust, in virtue of the fact that they take place in a system where alternatives to plea deals are trials with either expensive counsel or overstretched public defenders, high conviction rates for cases reaching jury verdicts, and long sentences for guilty verdicts. Plea deals to long sentences are common, in other words, because of other kinds of injustices and statutory penalties that outpace what is necessary for either deterrence or desert. Likewise, one might argue that the appropriate focus in analyzing use of VAMs to evaluate K-12 teachers is on issues such as school funding, the vast differences in resources available in wealthy and poor school districts, how those differences track race and ethnicity, the need to ensure excellent teaching, and administrative pressures on teachers generally. One might argue that background checks are a symptom of broader problems of wealth and income inequality, the austere state of healthcare for many Americans, and the fact that landlords have much greater legal power than tenants.

It is certainly the case that one can focus on different facets of problems, be they in criminal justice, education, rental markets, social media, disinformation campaigns, or anything else. But note that how criminal justice systems assess risk, allocate supervisory resources, and sentence people convicted of crimes are constitutive parts of the criminal justice system, not discrete actions. And different, constitutive parts of the criminal justice system may warrant different analyses, and it is not clear that overall systems are best analyzed in a univocal way. And in any case, our work here should be seen as sitting alongside other work on criminal justice, education, social media, disinformation, and so on. Systemic arguments and narrower arguments may be complementary rather than conflicting.

Finally, perhaps the biggest limitation to the project is that the sand is shifting beneath our feet. New ways of using technology continue apace, and some of the systems we discuss over the course of the book will soon have their problems resolved, change, or be replaced by new ones. But those newer algorithmic systems could also be used in ways that fail to respect autonomy, inhibit practical and cognitive agency, limit freedom, launder agency, and create drags on legitimation. That brings us full circle. What matters is not the particular technologies or the specific ways that those technologies are used. Rather, the underlying moral questions are a better anchor. Autonomy is certainly not the only moral value, and the nature, scope, and value of autonomy are contested. Moreover, the ways autonomy interacts with other values require continual reassessment. But, as we stated in Chapter 1, a rock-bottom assumption of this book is that autonomy matters – hopefully, considering autonomy helps us sort through a number of questions about technologies on the ground.