DOI:10.1145/3144172

Solon Barocas and danah boyd

Computing Ethics
Engaging the Ethics of Data Science in Practice

Seeking more common ground between data scientists and their critics.

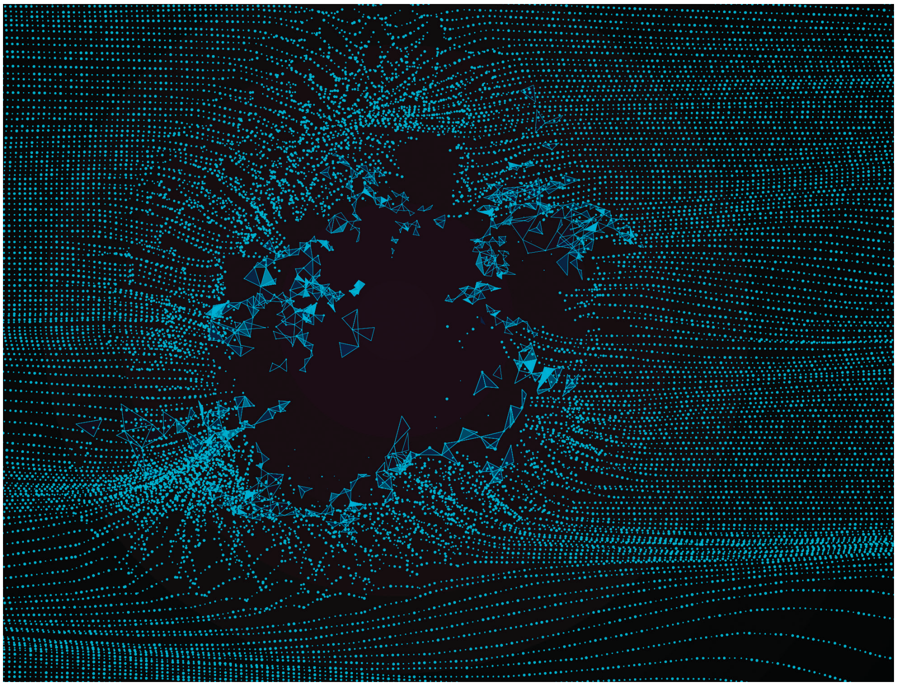

IMAGE BY LIU ZISHAN

CRITICAL COMMENTARY ON data science has converged on a worrisome idea: that data scientists do not recognize their power and, thus, wield it carelessly. These criticisms channel legitimate concerns about data science into doubts about the ethical awareness of its practitioners. For these critics, carelessness and indifference explains much of the problem—to which only they can offer a solution.

Such a critique is not new. In the 1990s, Science and Technology Studies (STS) scholars challenged efforts by AI researchers to replicate human behaviors and organizational functions in software (for example, Collins[3]). The scholarship from the time was damning: expert systems routinely failed, critical researchers argued, because developers had impoverished understandings of the social worlds into which they intended to introduce their tools.[6] At the end of the decade, however, Mark Ackerman reframed this as a social-technical gap between "what we know we *must* support socially and what we *can* support technically."[1] He argued that AI's deficiencies did not reflect a lack of care on the part of researchers, but a profound challenge of dealing with the full complexity of the social world. Yet here we are again.

Our interviews with data scientists give us reason to think we can avoid this repetition. While practitioners were quick to point out that common criticisms of data science tend to lack technical specificity or rest on faulty understandings of the relevant techniques, they also expressed frustration that critics failed to account for the careful thinking and critical reflection that data scientists already do as part of their everyday work. This was more than resentment at being subject to outside judgment by non-experts. Instead, these data scientists felt that easy criticisms overlooked the kinds of routine deliberative activities that outsiders seem to have in mind when they talk about ethics.

Ethics in Practice

Data scientists engage in countless acts of implicit ethical deliberation while trying to make machines learn something useful, valuable, and reliable. For example, dealing with dirty and incomplete data is as much a moral as a practical concern. It requires making a series of small decisions that are often fraught, forcing reflection at each step. How was this data collected? Does it capture the entire population and full range of behavior that is of interest? The same is true for validating a model and settling on an acceptable error rate. What must a data scientist do to prove to herself that a model will indeed perform well when deployed? How do data scientists decide that a reported error rate is tolerable—and defendable? Ethical considerations also emerge while making more fundamental decisions regarding the choice of learning algorithm, where practitioners frequently struggle to find an approach that maximizes the resulting models' performance while also providing some degree of interpretability. When is the ability to meaningfully interrogate a model sufficiently important to justify some cost in performance? What kinds of decisions—and real-world effects—drive data scientists to develop a model that they can explain, even if its decisions might be less accurate as a result?

These are difficult decisions, for which data scientists must employ carefully cultivated judgment. Yet, many data scientists do not use the language of ethics to talk about these practices. They may speak of trade-offs, but they primarily talk about what it takes to be *good* at what they do. Pressed about "ethics" directly, many data scientists say "this is not my area," even though they draw on a wide range of values to work through difficult tensions.

Broad critiques of data science practices cannot account for the diversity of practices, concerns, or efforts among data scientists. Instead, they often presume ignorance or corrupt intentions. All too often, the data scientists we have encountered are quite sympathetic to the sentiment behind the critiques they hear, but feel maligned and misunderstood, unacknowledged for their efforts, and frustrated by vague recommendations that are not actionable. Outsiders' use of the term "ethics" suggests that normative concerns must be dealt with independently or on top of technical practice—without noticing that ethical deliberation is embedded in the everyday work of data scientists.

Broad critiques of data science practices cannot account for the diversity of practices, concerns, or efforts among data scientists.

Even when attempting to address ethical issues more explicitly, practitioners face difficult trade-offs. One interviewee described a dilemma in choosing whether or not to "know" the gender of the individuals in his model—with that information, he could check whether his model might exhibit some kind of gender bias; without it, he could claim that this sensitive attribute did not figure into the model. Other researchers who are concerned about gender biases in data have attempted to build technical interventions to address them,[5] but such an approach requires trading off privacy in order to construct a viable fairness remedy, a decision that presents its own challenges.[4]

Where Ethics Is Not Enough

Critics are right to emphasize the seriousness of the implications of data science. And, as Cathy O'Neil has pointed out in *The Weapons of Math Destruction*,[8] data science is being deployed by powerful organizations to achieve goals that can magnify inequality and undermine democratic decision-making. She calls on data scientists to recognize how they are being used—and to push back against misuse of their skills.

Unfortunately, certain problems may stem from genuine value conflicts, not simply a lack of attention to the values at stake. Over the past year, a debate has unfolded over the use of data science in criminal justice, where courts rely on risk scores to make decisions about who should be released from prison while awaiting trial. The stakes are high: those given bail are more likely to keep their jobs, house, children, and spouse; those who are not are more likely to plead guilty, even when they are innocent.

A group of data scientists working with *ProPublica* established that black defendants in Broward County, FL, who did not reoffend were twice as likely to be mislabeled as posing a high risk of recidivism than white defendants.[2] They argued that the system exhibited a clear racial bias because errors imposed a far greater cost on black defendants, who were more likely to be wrongly incarcerated, while white defendants were more likely to be set free but nevertheless recidivate. Northpointe (now Equivant), the company behind the risk assessment, countered that its tool was equally accurate in predicting recidivism for black and white defendants. Since then, computer scientists and statisticians have debated the different qualities that an intuitive sense of fairness might imply: that a risk score is equally accurate in predicting the likelihood of recidivism for members of different racial groups; that members of different groups have the same chance of being wrongly predicted to recidivate; or that failure to predict recidivism happens at the same rate across groups. While each of these expectations of a fair score might seem like complementary requirements, recent work has established that satisfying all three at the same time would be impossible in most situations; meeting two will mean failing to comply with the third.[4,7] Even if Northpointe had been more sensitive to disparities in the false positive and false negative rates, the appropriate way to handle such a situation may not have been obvious. Favoring certain fairness properties over others could just as well have reflected a difference in values, rather than a failure to recognize the values at stake. One thing is for certain: this use of data science has prompted a vigorous debate, making clear that our normative commitments are not well articulated, that fuzzy values will be difficult to resolve computationally, and that existing ethical frameworks may not deliver clear answers to data science challenges.

Toward a Constructive Collaboration

The critical writing on data science has taken the paradoxical position of insisting that normative issues pervade all work with data while leaving unaddressed the issue of data scientists' ethical agency. Critics need to consider how data scientists learn to think about and handle these trade-offs, while practicing data scientists need to be more forthcoming about all of the small choices that shape their decisions and systems.

Technical actors are often far more sophisticated than critics at understanding the limits of their analysis. In many ways, the work of data scientists is a *qualitative* practice: they are called upon to parse an amorphous problem, wrangle a messy collection of data, and make it amenable to systematic analysis. To do this work well, they must constantly struggle to understand the contours and the limitations of both the data and their analysis. Practitioners want their analysis to be accurate and they are deeply troubled by the limits of tests of validity, the problems with reproducibility, and the shortcomings of their methods.

Technical actors are often far more sophisticated than critics at understanding the limits of their analysis.

Many data scientists are also deeply disturbed by those who are coming into the field without rigorous training and those who are playing into the hype by promising analyses that are not technically or socially responsible. In this way, they should serve as allies with critics. Both see a need for nuances within the field. Unfortunately, universalizing critiques may undermine critics' opportunities to work with data scientists to meaningfully address some of the most urgent problems.

Of course, even if data scientists take care in their work and seek to engage critics, they may not be well prepared to consider the full range of ethical issues that such work raises. In truth, few people are. Our research suggests the informal networks that data scientists rely on are fallible, incomplete, and insufficient, and that this is often frustrating for data scientists themselves.

In order to bridge the socio-technical gap that Ackerman warned about 20 years ago, data scientists and critics need to learn to appreciate each other's knowledge, practices, and limits. Unfortunately, there are few places in which such learning can occur. Many data scientists feel as though critics only talk *at* them. When we asked one informant why he did not try to talk back, he explained that social scientists and humanists were taught to debate and that he was not. Critics get rewarded for speaking out publicly, he said, garnering rewards for writing essays addressed to a general audience. This was not his skillset nor recognized as productive by his peers.

The gaps between data scientists and critics are wide, but critique divorced from practice only increases them. Data scientists, as the ones closest to the work, are often the best positioned to address ethical concerns, but they often need help from those who are willing to take time to understand what they are doing and the challenges of their practice. We must work collectively to make the deliberation that is already a crucial part of data science visible. Doing so will reveal far more common ground between data scientists and their critics and provide a meaningful foundation from which to articulate shared values. C

References

1. Ackerman, M.S. The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Human-Computer Interaction 15*, (2–3), 2000, 179–203.
2. Angwin, J. et al. Machine bias. *ProPublica.* (May 23, 2016); https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Collins, H.M. *Artificial Experts: Social Knowledge and Intelligent Machines.* MIT Press, 1993.
4. Corbett-Davies, S. et al. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post* (Oct. 17, 2016); https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/
5. Feldman, M. et al. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2015), 259–268.
6. Hess, D.J. Editor's Introduction, *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford University Press, 2001.
7. Kleinberg, J., Mullainathan, S. and Raghavan, S. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Arxiv.org.* 2016; https://arxiv.org/abs/1609.05807
8. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown, 2016.
9. Žliobaitė I. and Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law 24*, 2 (Feb. 2016), 183–201.

Solon Barocas (sbarocas@cornell.edu) is an Assistant Professor of Information Science at Cornell University.

danah boyd (danah@datasociety.net) is a Principal Researcher at Microsoft Research and the Founder/President of Data & Society.