The technical perspective on ethics: An overview and critique

Momin M. Malik

Senior Data Scientist – AI Ethics, Mayo Clinic

Fellow, Institute in Critical Quantitative, Computational, & Mixed Methodologies

Instructor, University of Pennsylvania School of Social Policy & Practice

Tuesday, March 29, 2022

The Center for Digital Ethics & Policy Annual International Symposium ‘22

Manzel Bowman,
Station No.99 (2018)

Goals and outline

- With what lens do “technical” people approach ethics?
- What does this lens involve?
- Where does this lens come from?
- Where does it break down and how?

Caveats

- We should not take any statements at face value as evidence of what the authors actually think: I myself frequently engage in strategic framing
  - Instead, we should take it as evidence of *what sort of framings are deemed acceptable* (and note that these phrasings are what passed peer review)
  - (One example I use, Corbett-Davies & Goel, does this explicitly, taking a turn halfway through the paper from math towards the limits of abstraction)
- Some of the framing I identify are already out of vogue; certainly, I raise issues when I am a reviewer
- My own perspective: hybrid, but primarily *technical*

The technical perspective

The background

- Zemel et al., 2013: “Information systems are becoming increasingly reliant on statistical inference and learning to render all sorts of decisions, including the setting of insurance rates, the allocation of police, the targeting of advertising, the issuing of bank loans, the provision of health care, and the admission of students.”
- Feldman et al., 2015: “Today, algorithms are being used to make decisions both large and small in almost all aspects of our lives, whether they involve mundane tasks like recommendations for buying goods, predictions of credit rating prior to approving a housing loan, or even life-altering decisions like sentencing guidelines after conviction.”
- Corbett-Davies & Goel, 2018: “In banking, criminal justice, medicine, and beyond, consequential decisions are often informed by statistical risk assessments that quantify the likely consequences of potential courses of action.”

The problem

- 2013: “This growing use of automated decision-making has sparked heated debate among philosophers, policy-makers, and lawyers. Critics have voiced concerns with bias and discrimination in decision systems that rely on statistical inference and learning.” [No citations]
- 2015: “How do we know if these algorithms are biased, involve illegal discrimination, or are unfair? These concerns have generated calls, by governments and NGOs alike, for research into these issues [17, 23].”
- 2018: “As the influence and scope of these risk assessments increase, academics, policymakers, and journalists have raised concerns that the statistical models from which they are derived might inadvertently encode human biases (Angwin et al., 2016; O’Neil, 2016).”

(Current language; 2021 FAccT)

- Singh et al.: “Deployment of machine learning algorithms to aid consequential decisions, such as in medicine, criminal justice, and employment, require revisiting the dominant paradigms of training and testing such algorithms.”
- Ron et al.: “Algorithmic decision making plays a fundamental role in many facets of our lives; criminal justice [10, 11, 29], banking [3, 18, 32, 40], online-advertisement [28, 30], hiring [1, 2, 4, 7] , and college admission [5, 26, 36] are just a few examples. With the abundance of applications in which algorithms operate, concerns about their ethics, fairness, and privacy have emerged.”
- Black & Frederickson: “Deep networks are becoming the go-to choice for challenging classification tasks due to their remarkable performance on many high-profile problems: they are used everywhere from recommendation systems [15] to medical research [8, 21], and increasingly in even more sensitive contexts, such as hiring [46], loan decisions [5, 51], and criminal justice [25]. Their continued rise in adoption has led to growing concerns about the tendency of these models to discriminate against certain individuals [4, 10, 13, 44], or otherwise produce outcomes that are seen as unfair.”

(Current language; 2021 FAccT)

- Nanda et al.: “Automated decision-making systems that are driven by data are being used in a variety of different real-world applications. In many cases, these systems make decisions on data points that represent humans (e.g., targeted ads [44, 53], personalized recommendations [3, 50], hiring [47, 48], credit scoring [31], or recidivism prediction [9]). In such scenarios, there is often concern regarding the fairness of outcomes of the systems [2, 18].”
- Taskeen et al.: “Nowadays, machine learning algorithms can uncover complex patterns in the data to produce an exceptional performance that can match, or even surpass, that of humans… Algorithms are conceived and function following strict rules of logic and algebra; it is hence natural to expect that machine learning algorithms deliver objective predictions and recommendations. Unfortunately, in-depth investigations reveal the excruciating reality that state-of-the-art algorithmic assistance is far from being free of biases.”

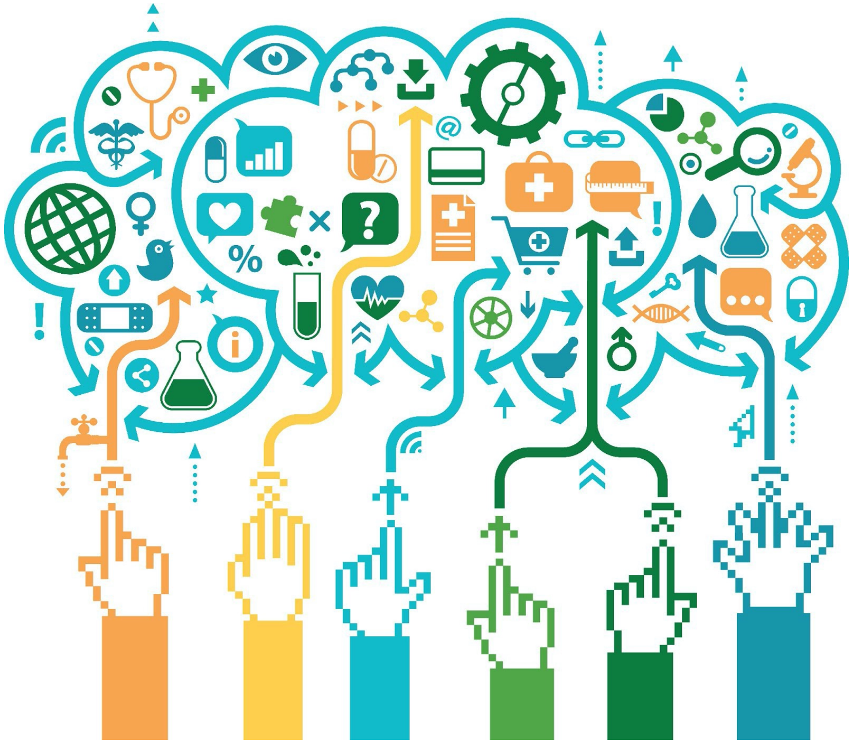
Comments

- Some version of a view: Machine learning has so much promise! But this promise comes with a flip side of unintended harm and consequences, that no one could have imagined, so we need to address it with the same tools we use to develop machine learning
- Even if not citing successes, *these take the application of machine learning as a given, or inevitable*
- None acknowledge (for example) the possibility of refusal, or that sometimes this might be a better way forward

Vision of the future (Morozov, 2013)

“If Silicon Valley had a designated futurist, her bright vision of the near future… would go something like this: Humanity, equipped with powerful self-tracking devices, finally conquers obesity, insomnia, and global warming as everyone eats less, sleeps better, and emits more appropriately. The fallibility of human memory is conquered too, as the very same tracking devices record and store everything we do. Car keys, faces, factoids: we will never forget them again…”

Vision of the future (Morozov, 2013)



"Politics, finally under the constant and far-reaching gaze of the electorate, is freed from all the sleazy corruption, backroom deals, and inefficient horse trading. Parties are disaggregated and replaced by Groupon-like political campaigns, where users come together—once—to weigh in on issues of direct and immediate relevance to their lives, only to disband shortly afterward. Now that every word—nay, sound—ever uttered by politicians is recorded and stored for posterity, hypocrisy has become obsolete as well. Lobbyists of all stripes have gone extinct as the wealth of data about politicians—their schedules, lunch menus, travel expenses— are posted online for everyone to review…"

Vision of the future (Morozov, 2013)

"Crime is a distant memory, while courts are overstaffed and underworked. Both physical and virtual environments—walls, pavements, doors, log-in screens—have become 'smart.' That is, they have integrated the plethora of data generated by the self-tracking devices and social-networking services so that now they can predict and prevent criminal behavior simply by analyzing their users. And as users don't even have the chance to commit crimes, prisons are no longer needed either. A triumph of humanism, courtesy of Silicon Valley."

The approach

- Eliminate correlations with protected attributes (Zemel et al.):

$$\frac{1}{|X_0^+|}\sum_{n\in X_0^+} M_{n,k} = \frac{1}{|X_0^-|}\sum_{n\in X_0^-} M_{n,k} \Rightarrow$$

$$\frac{1}{|X_0^+|}\sum_{n\in X_0^+} M_n\mathbf{w} = \frac{1}{|X_0^-|}\sum_{n\in X_0^-} M_n\mathbf{w} \Rightarrow$$

$$\frac{1}{|X_0^+|}\sum_{n\in X_0^+} y_n^+ = \frac{1}{|X_0^-|}\sum_{n\in X_0^-} y_n^- .$$

- Define a metric that has a provable relationship to the “80% rule” (Feldman et al.):

$$\text{BER}(f(Y),X) = \frac{\Pr[f(Y)=0|X=1]+\Pr[f(Y)=1|X=0]}{2}$$

- Express anti-classification, classification parity, and calibration (Corbett-Davies & Goel):

$$d(x) = d(x') \text{ for all } x, x' \text{ such that } x_u = x'_u$$

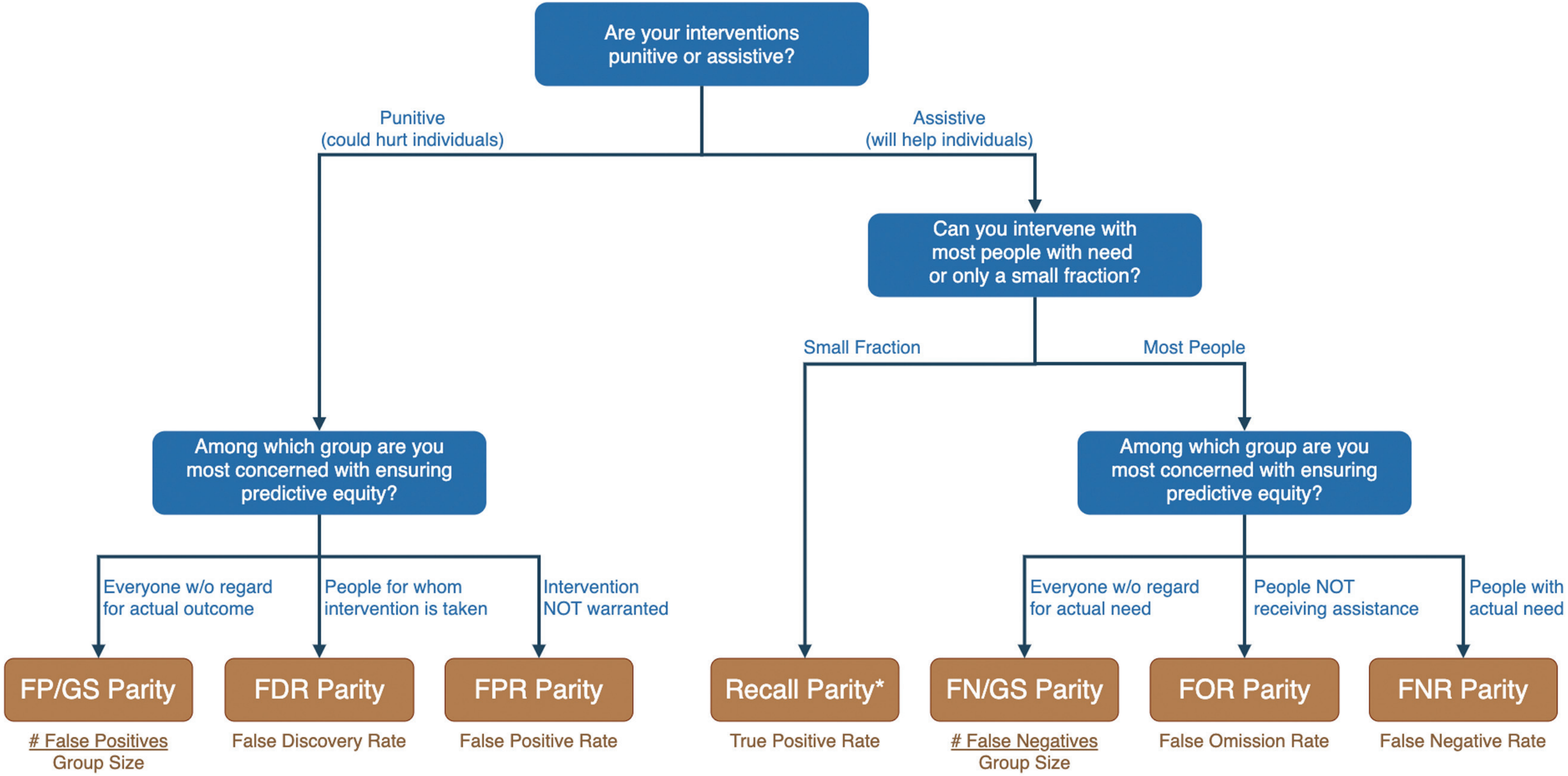
$$\Pr(d(X)=1 \mid Y=0,\, X_p) = \Pr(d(X)=1 \mid Y=0)$$

$$\Pr(Y=1 \mid s(X),\, X_p) = \Pr(Y=1 \mid s(X))$$

The wall of the technical perspective

- Alexandra Chouldechova (2017) showed that we cannot simultaneously satisfy three specific metrics: accuracy equality (equal accuracy across groups), equal opportunity (equal false negative rate across groups), and predictive parity (equal precision [positive predictive value] across groups)
  - (Partially what the COMPAS debate is about)
- So now, ML moves to: rely on domain experts to determine what fairness metric we should use

Landscape for fairness (Rodolfo et al.)



Are your interventions punitive or assistive?

Punitive (could hurt individuals)

Assistive (will help individuals)

Can you intervene with most people with need or only a small fraction?

Small Fraction

Most People

Among which group are you most concerned with ensuring predictive equity?

Everyone w/o regard for actual outcome

People for whom intervention is taken

Intervention NOT warranted

Among which group are you most concerned with ensuring predictive equity?

Everyone w/o regard for actual need

People NOT receiving assistance

People with actual need

FP/GS Parity

$\frac{\text{\# False Positives}}{\text{Group Size}}$

FDR Parity

False Discovery Rate

FPR Parity

False Positive Rate

Recall Parity*

True Positive Rate

FN/GS Parity

$\frac{\text{\# False Negatives}}{\text{Group Size}}$

FOR Parity

False Omission Rate

FNR Parity

False Negative Rate

Where this fits in

- This diagram is supremely useful, and can and should be a basis for auditing/formal analysis when we choose to use machine learning (or when we analyze an existing system)
- From a technical perspective, this is maybe as far as we can go
- But that doesn’t mean that there’s not a lot further *to* go

The idea of limits to abstraction is novel

- Reads as a fairly straightforward STS primer for outsiders
- But for some CS insiders, it was earth-shattering to consider the limits to abstraction
- Still, even for many of those people, it represented an endpoint; having pointed out the limits of abstraction, we are done, and there's nothing more to do (other than get back to working on those abstractions).
- I.e., could exist within the same assumptions of inevitability of using abstractions/ building systems

Fairness and Abstraction in Sociotechnical Systems

Andrew D. Selbst
Data & Society Research Institute
New York, NY
andrew@datasociety.net

danah boyd
Microsoft Research and
Data & Society Research Institute
New York, NY
danah@datasociety.net

Sorelle A. Friedler
Haverford College
Haverford, PA
sorelle@cs.haverford.edu

Suresh Venkatasubramanian
University of Utah
Salt Lake City, UT
suresh@cs.utah.edu

Janet Vertesi
Princeton University
Princeton, NJ
jvertesi@princeton.edu

ABSTRACT

A key goal of the fair-ML community is to develop machine-learning based systems that, once introduced into a social context, can achieve social and legal outcomes such as fairness, justice, and due process. Bedrock concepts in computer science—such as abstraction and modular design—are used to define notions of fairness and discrimination, to produce fairness-aware learning algorithms, and to intervene at different stages of a decision-making pipeline to produce "fair" outcomes. In this paper, however, we contend that these concepts render technical interventions ineffective, inaccurate, and sometimes dangerously misguided when they enter the societal context that surrounds decision-making systems. We outline this mismatch with five "traps" that fair-ML work can fall into even as it attempts to be more context-aware in comparison to traditional data science. We draw on studies of sociotechnical systems in Science and Technology Studies to explain why such traps occur and how to avoid them. Finally, we suggest ways in which technical designers can mitigate the traps through a refocusing of design in terms of process rather than solutions, and by drawing abstraction boundaries to include social actors rather than purely technical ones.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Fairness-aware Machine Learning, Sociotechnical Systems, Interdisciplinary

ACM Reference Format:
Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287598

1 INTRODUCTION

On the typical first day of an introductory computer science course, the notion of abstraction is explained. Students learn that systems can be described as black boxes, defined precisely by their inputs, outputs, and the relationship between them. Desirable properties of a system can then be described in terms of inputs and outputs alone: the internals of the system and the provenance of the inputs and outputs have been abstracted away.

Machine learning systems are designed and built to achieve specific goals and performance metrics (e.g., AUC, precision, recall). Thus far, the field of fairness-aware machine learning (fair-ML) has been focused on trying to engineer fairer and more just machine learning algorithms and models by using fairness itself as a property of the (black box) system. Many papers have been written proposing definitions of fairness, and then based on those, generating best approximations or fairness guarantees based on hard constraints or fairness metrics [24, 32, 39, 40, 72]. Almost all of these papers bound the system of interest narrowly. They consider the machine learning model, the inputs, and the outputs, and abstract away any context that surrounds this system.

We contend that by abstracting away the social context in which these systems will be deployed, fair-ML researchers miss the broader context, including information necessary to create fairer outcomes, or even to understand fairness as a concept. Ultimately, this is because while performance metrics are properties of systems in total, technical systems are subsystems. Fairness and justice are properties of social and legal systems like employment and criminal justice, not properties of the technical tools within. To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore to make a category error, or as we posit here, an abstraction error.

In this paper, we identify five failure modes of this abstraction error. We call these the Framing Trap, Portability Trap, Formalism Trap, Ripple Effect Trap, and Solutionism Trap. Each of these traps arises from failing to consider how social context is interlaced with technology in different forms, and thus the remedies also require a deeper understanding of "the social" to resolve problems [1]. After explaining each of these traps and their consequences, we draw on

FAT '19, January 29–31, 2019, Atlanta, GA, USA*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287598

59

Where does this lens come from, and how do people break out?

Phil Agre [ey-gree]

- PhD in 1989 from MIT (EECS)
- Influential works:
  - "Surveillance and Capture: Two Models of Privacy" (1994)
  - "The Soul Gained and Lost: Artificial Intelligence as a Philosophical Project" (1995)
  - *Computation and Human Experience* (1997)
  - Red Rock Eater News Service (1996-2002)
- Former associate professor at UCLA
  - Sister filed missing persons report in October 2009, after not seeing him since Spring 2008 and learning he abandoned his job and apartment
  - Found by LA County Sheriff's Department in January 2010
- Won't focus on him personally, but instead on his 1997 piece "Towards a critical technical practice: Lessons learned trying to reform AI"

From AI to social sciences

“My ability to move intellectually from AI to the social sciences – that is, to **stop thinking the way that AI people think, and to start thinking the way that social scientists think** – had a remarkably large and diverse set of historical conditions. AI has never had much of a reflexive critical practice, any more than any other technical field. Criticisms of the field, no matter how sophisticated and scholarly they might be, are certain to be met with the assertion that the author simply fails to understand a basic point. And so, **even though I was convinced that the field was misguided and stuck, it took tremendous effort and good fortune to understand how and why.”**

Autobiographical account of a crisis

“My college did not require me to take many humanities courses, or learn to write in a professional register, and so **I arrived in graduate school at MIT with little genuine knowledge beyond math and computers. This realization hit me with great force halfway through my first year of graduate school…**

“fifteen years ago, I had absolutely no critical tools with which to defamiliarize those ideas – to see their contingency or imagine alternatives to them. Even worse, I was unable to turn to other, nontechnical fields for inspiration. As an AI practitioner already well immersed in the literature, I had incorporated the field's taste for technical formalization so thoroughly into my own cognitive style that I literally could not read the literatures of nontechnical fields at anything beyond a popular level. **The problem was not exactly that I could not understand the vocabulary, but that I insisted on trying to read everything as a narration of the workings of a mechanism.”**

Some other perspectives

- Malazita & Resetarb, 2019, “Infrastructures of abstraction: how computer science education produces anti-political subjects”
- Hanna Wallach, 2018: “Spoiler alert: The punchline is simple. Despite all the hype, machine learning is not a be-all and end-all solution. We still need social scientists if we are going to use machine learning to study social phenomena in a responsible and ethical manner.”

Critical “awakening”

“At first I found [critical] texts impenetrable, not only because of their irreducible difficulty but also because **I was still tacitly attempting to read everything as a specification for a technical mechanism**… My first intellectual breakthrough came when, for reasons I do not recall, it finally occurred to me to stop translating these strange disciplinary languages into technical schemata, and instead simply to learn them on their own terms…”

Critical “awakening”

“I still remember the vertigo I felt during this period; I was speaking these strange disciplinary languages, in a wobbly fashion at first, without knowing what they meant — without knowing what *sort* of meaning they had…

“In retrospect, this was the period during which **I began to ‘wake up’, breaking out of a technical cognitive style that I now regard as extremely constricting**.”

Theorizing this process

- This bears remarkable resemblances to Paulo Freire’s idea of *critical consciousness*: become aware of our place in society to work for its betterment
- Follow-up work in education (specifically, Mezirow on “perspective transformation”; 1978) theorizes this process

Perspective transformation vs. Agre

✓ 1. A disorienting dilemma
✓ 2. Self-examination with feelings of guilt or shame
✓ 3. A critical assessment of assumptions
✗ 4. Recognition that one's discontent and process of transformation are shared and that others have negotiated a similar change
✗ 5. Exploration of options for new roles, relationships, and actions
? 6. Planning of a course of action
? 7. Acquisition of knowledge and skills for implementing one's plans
? 8. Provisionally trying out new roles
? 9. Building of competence and self-confidence in new roles and relationships
? 10. A reintegration into one's life on the basis of conditions dictated by one's new perspective.

Who experiences, how and why?

- Mezirow doesn't get at *who* experiences a perspective transformation
  - Empirical evidence/experience seems to be insufficient
  - Having a "disorienting dilemma", but then *reflecting* about it
- Work after Mezirow (Taylor & Snyder, 2012): went beyond the "rationalist" framing, recognized that self-actualization is not the only goal, recognized key role of interpersonal relationships

Ethics and interventions

- I contend: *Connecting to critical consciousness gives us a roadmap for "ethics" more important than ethical frameworks, or formal ethical reasoning: or at least <u>necessary</u>, if not sufficient*
- Interventions: build community with others who have negotiated a similar change; form coalitions with others; leverage our privilege, e.g., to oppose gatekeeping and bring in others, support the right of refusal; mentor others; give feedback to invest spontaneous actions with biographical significance
- Insofar as we maintain civilization on the current scale, abstraction is necessary: just because some things aren't current formalized doesn't mean they can't be. Even developing critical consciousness maybe could be included in formal education (Trbušić, 2014)
  - As a minimum of where, beyond a technical perspective, we can try to get technical people: allowing for the right of refusal, and for the option of opposing adoptions of ML in any given case

Assumptions in social research

Issue	**Positivism**	**Postpositivism**	**Critical theory et al.**	**Constructivism**	**Participatory**
Ontology	Reality independent of, prior to human conception of it and apprehensible.	Reality is “real” but only imperfectly and approximately apprehensible	There is a reality but it is secret/hidden	Relativism	Participative: multiple co-created realities
Epistemology	Singular, perspective-independent, neutral, atemporal, universally true findings	Findings are provisionally true, affected/distorted by society; multiple descriptions possible but equivalent	Truth is mediated by value; how we come to know something matters for what how meaningful it is	Transactional/subjectivist; co-created findings	Come to know things through involving other people
Methodology	Experimental/ manipulative; verification of hypotheses	Falsification of hypotheses; some qual, but only in service of quant	Dialogic/dialectical	Hermeneutical/dialectical	Collaborative, action-oriented; flatten hierarchies, jointly decide to engage in action
Axiology	Quant knowledge, people who have, have ultimate valuable	Quant knowledge most valuable, but qual can serve it	Marginalization is important, people who have it have unique insights	Value is relative; for us, understanding process of construction is valuable	Everyone is valuable; Reflexivity, co-created knowledge, non- western ways of knowing to combat erasure and dehumanization

Assumptions of social research paradigms (Malik & Malik, 2021). Based on Guba and Lincoln’s (2005) “Basic beliefs (metaphysics) of alternative inquiry paradigms.”

Summary and conclusion

- Technical perspective engenders a view where abstraction is the only legitimate way to engage with the world
- It fails to inculcate awareness of or appreciation of the limits of abstraction, or the possibility of sometimes rejecting abstraction
- Breaking out of this view is both difficult, requiring additional biographical inputs, and disorienting
- But this is necessary to get people engaged in ethical reasoning
- Ideally, this will go beyond what can pass as a "sociotechnical" perspective, to a fully constructivist, critical, and even participatory perspective

Thank you!

References

Agre, P. E. (1997). Towards a critical technical practice: Lessons learned trying to reform AI. In G. Bowker, S. L. Star, W. Turner, & L. Gasser (Eds.), *Social science technical systems and cooperative work: Beyond the great divide* (pp. 131–157). Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9781315805849-14

Black, E., & Fredrikson, M. (2021). Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 285–295). https://doi.org/10.1145/3442188.3445894

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153–163. Https://doi.org/10.1089/big.2016.0047

Corbett-Davies, D., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. https://arxiv.org/abs/1808.00023

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)* (259–268). https://doi.org/10.1145/2783258.2783311

Malazita, J. W., & Resetar, K. (2019). Infrastructures of abstraction: How computer science education produces anti-political subjects. *Digital Creativity, 30*(4), 300-312. https://doi.org/10.1080/14626268.2019.1682616

Malik, M. M., & Malik, M. (2021). Critical technical awakenings. *Journal of Social Computing, 2*(4), 365–384. https://doi.org/10.23919/JSC.2021.0035

Mezirow, J. (1978). Perspective transformation. *Adult Education Quarterly, 28*(2), 100-110. https://doi.org/10.1177/074171367802800202

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. Public Affairs.

Nanda, V., Dooley, S., Singla, S., Feizi, S., & Dickerson, J. P. (2021). Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 466–477). https://doi.org/10.1145/3442188.3445910

Rodolfa, K. T., Saleiro, P., & Ghani, R. (2021). Bias and fairness. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane, *Big data and social science: Data science methods and tools for research and practice* (pp. 281–312). CRC Press. https://doi.org/10.1201/9780429324383-11

Ron, T., Ben-Porat, O., & Shalit, U. (2021). Corporate social responsibility via multi-armed bandits. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 26–40). https://doi.org/10.1145/3442188.3445868

Selbst, A. D., boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (pp. 59–68). https://doi.org/10.1145/3287560.3287598

Singh, H., Singh, R., Mhasawade, V., & Chunara, R. (2021). Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 3–13). https://doi.org/10.1145/3442188.3445865

Taskesen, B., Blanchet, J., Kuhn, D., & Nguyen, V. A. (2021). A statistical test for probabilistic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 648–665). https://doi.org/10.1145/3442188.3445927

Taylor, E. W., & Snyder, M. J. (2012). A critical review of research on transformative learning theory, 2006–2010. In E. W. Taylor & P. Cranton (Eds.), *The handbook of transformative learning: Theory, research, and practice* (pp. 37–55). San Francisco: Jossey-Bass.

Trbušić, H. (2014). Engineering in the community: Critical consciousness and engineering education. *Interdisciplinary Description of Complex Systems, 12*(2), 108–118. https://doi.org/10.7906/indecs.12.2.1

Wallach, H. (2018). Computational social science ≠ computer science + social data. *Communications of the ACM, 61*(3), 42–44. https://doi.org/10.1145/3132698

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning, 28*(3), 325-333. https://proceedings.mlr.press/v28/zemel13.html