

# Measuring Justice in Machine Learning

Alan Lundgard  
Massachusetts Institute of Technology  
Cambridge, Massachusetts  
lundgard@mit.edu

## ABSTRACT

How can we build more just machine learning systems? To answer this question, we need to know both what justice is and how to tell whether one system is more or less just than another. That is, we need both a definition and a measure of justice. Theories of distributive justice hold that justice can be measured (in part) in terms of the fair distribution of benefits and burdens across people in society. Recently, the field known as fair machine learning has turned to John Rawls's theory of distributive justice for inspiration and operationalization. However, philosophers known as capability theorists have long argued that Rawls's theory uses the wrong measure of justice, thereby encoding biases against people with disabilities. If these theorists are right, is it possible to operationalize Rawls's theory in machine learning systems without also encoding its biases? In this paper, I draw on examples from fair machine learning to suggest that the answer to this question is no: the capability theorists' arguments against Rawls's theory carry over into machine learning systems. But capability theorists don't only argue that Rawls's theory uses the wrong measure, they also offer an alternative measure. Which measure of justice is right? And has fair machine learning been using the wrong one?

## CCS CONCEPTS

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

## KEYWORDS

Justice, machine learning, philosophy, operationalization, disability, distributive justice, capability, measure, fairness, bias, discrimination

### ACM Reference Format:

Alan Lundgard. 2020. Measuring Justice in Machine Learning. In *Conference on Fairness, Accountability, and Transparency (FAT\* '20)*, January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 1 page. <https://doi.org/10.1145/3351095.3372838>

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*FAT\* '20, January 27–30, 2020, Barcelona, Spain*

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6936-7/20/01...\$15.00

<https://doi.org/10.1145/3351095.3372838>

---

*Non-Archival Paper Submission.* *FAT\* '20* offers authors the choice of archival and non-archival paper submissions. Accepted non-archival papers only appear as abstracts in the proceedings. *FAT\* '20* offers a non-archival option to avoid precluding the future submission of these papers to area-specific journals. Note that all submissions have the same page length requirements and are judged by the same quality standards regardless of whether the authors choose the archival or non-archival option. <https://fatconference.org/2020/>

# Measuring Justice in Machine Learning

Alan Lundgard

Massachusetts Institute of Technology

ACM Conference on Fairness, Accountability, and Transparency

January 30, 2020

@AlanLundgard

1

Today I'll be talking about measuring justice in machine learning, focusing on one type of justice called distributive justice.

Measuring Justice in Machine Learning

Alan Lundgard

*ACM Conference on Fairness, Accountability, and Transparency*, 2020.

<https://dl.acm.org/doi/abs/10.1145/3351095.3372838>

# Theories of Distributive Justice

- John Rawls
  - Recently dubbed “AI’s favorite philosopher” [Procaccia 2019]
- Capability Theorists
  - Rawls’s measure encodes biases against disability and marginalized groups

The field known as fair machine learning has recently sought to operationalize philosophical theories of distributive justice within algorithmic systems. This especially true of John Rawls, who was dubbed "AI's favorite philosopher." But other philosophers known as capability theorists have long argued that Rawls uses the wrong measure of justice, thereby encoding biases against disabled people and other marginalized groups.

# What Is the Right Measure of Justice?

## ➤ Resources

- Defended by John Rawls, Erin Kelly, Thomas Pogge, et al.

## ➤ Capabilities

- Defended by Amartya Sen, Martha Nussbaum, Elizabeth Anderson, et al.

[Brighouse and Robeyns 2010]

@AlanLundgard

3

What is the right measure of justice in machine learning? Rawls defends what we may call a resource-oriented measure, whereas the capability theorists defend a capability measure. In this talk, I'll explain (at a high level) what makes these measures different, and how they have been (or could be) operationalized in machine learning, in potentially unjust ways.

# What Are Resources?

- Distributive Justice
  - Material possessions, income and wealth, opportunities, basic rights, etc.
- Machine Learning
  - Loans, college admissions, employment opportunities, predictive accuracy, etc.

First, resources. Machine learning often conceptualizes “unfairness” as a problem of “resource allocation,” or distributing resources across people according to some allocative principle. In philosophy, resources broadly include things that people can have in their possession, such as their income and wealth, opportunities for employment, loans, etc. In machine learning, resources include all those from philosophy, as well as computational artifacts such as the predictive accuracy of a machine learning model, its false positive and negative rates, and so on.

A distinguishing characteristic of resource measures is that they can be expressed as single-valued numerical quantities. This makes them amenable to operationalization in machine learning systems. However, capability theorists argue that this is (in part) where resource measures go wrong, and end up failing to do justice by marginalized groups of people. Consider two cases from fair machine learning.

## Case 1: Text Auto-Complete and Dialects

- Fairness Goal
  - Fairly distribute predictive accuracy across dialects
- Resource Measure
  - Predictive accuracy
- Measure Limitation
  - Insensitive to social, political context

[Hashimoto et al. 2018]

@AlanLundgard

5

Case 1. Researchers recently operationalized Rawls's theory when training a text-autocomplete model (that is, something like predictive typing on your phone's texting app). They chose the model's predictive accuracy as their "resource" and allocated more of it to speakers of the so-called African-American English dialect (AAE). Their well-intentioned goal was to make the model fairer (that is, more accurate) for speakers of this particular dialect. But is more predictive accuracy what AAE speakers actually want? For example, it's not difficult to imagine how auto-completing AAE words could be perceived as stereotyping, appropriating, and/or entrenching discourse inequality for actual speakers of AAE.

Further, the researchers did not engage with anyone from the Black community when designing and evaluating their model, instead relying on anonymous Mechanical Turk workers to perform \*as if\* they were speakers of AAE. Absent direct engagement with the affected community, a single-valued quantitative measure (that is, a resource measure) like predictive accuracy is not likely to encompass the social and political norms surrounding the use of certain words and dialects.

## Case 2: Algorithmic Hiring and Disability

- Fairness Goal
  - Fairly distribute employment opportunities across demographic groups
- Resource Measure
  - Employment opportunities (application tests)
- Measure Limitation
  - Insensitive to disability, heterogeneities

[Bogen and Rieke 2018]

Case 2. Algorithmic hiring on job platforms (like LinkedIn, Indeed, or Pymetrics) is currently beyond the scope of equal employment opportunity for disabled people. Yet, these platforms distribute job opportunities to candidates in the form of pre-employment application tests. Hypothetically speaking, even if these opportunities were fairly allocated across demographic groups (perhaps using a Rawlsian conception of fairness), there would still be no guarantee that the opportunities themselves (that is, the pre-employment application tests) will be fairly accessible to disabled people.

Indeed, it may be safer to assume that such tests will be unfair, given that online platforms are widely known to be inaccessible to blind people, or people who use screen readers, in various ways. Absent any equal employment opportunity protection (such as reasonable accommodation, itself often insufficient) a single-valued quantitative measure (like the number of job opportunities distributed to a jobseeker) is not likely to encompass the many ways of being and working online.

# Resource Measures

## Pros

Publicly legible and verifiable

Unambiguous comparisons

## Cons

Insensitive to social, political context

Insensitive to disability, heterogeneities

[Brighthouse and Robeyns 2010]

@AlanLundgard

7

In both of these cases, the problem with the resource measure is that it fails to be sensitive to the different ways people are able to use (or not use) the resources that are allocated to them. In the first case, higher predictive text accuracy may seem like an always-desirable design goal, but may actually perpetuate psychosocial harms, such as stereotype threat or discourse inequality. In the second case, employment opportunities “fairly” distributed via hiring platforms may achieve only a superficial conception of fairness, if those platforms are themselves inaccessible to disabled people who use screen readers.

In other words, resource measures are insensitive to what people are (or are not) capable of doing with those resources, given their social and political context and personal heterogeneities. In response to this shortcoming, capability theorists argue for an alternative measure of justice: one that is sensitive to context and personal heterogeneities. Namely, they argue for capabilities.

# What Are Capabilities?

- Distributive Justice
  - *Being able to...* attain employment, move without barriers, etc.
- Machine Learning
  - *Being able to...* complete online application tests, contest automated decisions, etc.

What are capabilities? (Aside: note that the “capability” terminology originated in the late 1970’s and so may sound somewhat outdated by today’s disability language norms. Also note that the capability theory is consistent with the social, rather than the medical, model of disability.)

In short, capabilities are what people are able to do (or not do) given their particular social and political context. Capabilities might include being able to attain employment (for anyone who wants it), move without barriers (regardless of disability), complete online application tests comfortably and effectively, or contest automated decisions. Whereas resource measures are single-valued numerical quantities, capability measures are heterogeneous, combining both qualitative and quantitative data, in direct collaboration with the affected communities.

# Operationalizing Capabilities

1. Choose relevant capabilities in collaboration with community members
2. Choose indicators for each capability (qualitative and quantitative)
3. Map each indicator to the unit interval  $[0, 1]$
4. Aggregate all indicators (average)
5. Iterate on system design until all community members achieve a baseline capability level

[Murphy and Gardoni 2012]

@AlanLundgard

9

For example, here's just one possible way to formulate a capability measure when designing a machine learning system.

1. First (prior to building anything), and in direct collaboration with the affected communities, choose the capabilities that matter to them, given their social and political context.
2. Then, select indicators for those capabilities (for example, quantitative data like predictive accuracy of the model, qualitative data like written testimony, likert scale usability/accessibility rankings, and so on).
3. Map each indicator to the unit interval and combine them (perhaps by averaging). (Note: This step is optional if it is preferred that qualitative data remain qualitative.)
4. Finally, iterate on the system design until all community members reach a baseline capability level.

In such a capability measure, a single-valued numerical quantity (such as predictive accuracy) could still be used as one of multiple indicators. But, unlike resource measures, it is essential that this indicator is not the primary measure of the system's justness or fairness.

# Capability Measures

## Pros

Sensitive to social, political context  
Sensitive to disability, heterogeneities

## Cons

Practicability concerns  
Ambiguous comparisons

[Brighthouse and Robeyns 2010]

@AlanLundgard

10

Because capability measures are designed (at the outset) in collaboration with people from the affected communities, they are sensitive to that community's social and political context and its members' personal heterogeneities, including disabilities. A capability approach to system design is theoretically analogous to value-sensitive, participatory, or inclusive design practices in the field of human-computer interaction.

However, in comparison with resources, capability measures have some shortcomings: they often require collecting a large amount of data, they're likely to be more expensive, and they're potentially ambiguous. But these shortcomings, I argue, are counterbalanced by their sensitivity to social and political injustices.

“Capability measures are...

...sensitive to structural and psychosocial injustices that interfere with individuals’ functioning as equals, although they are neither constituted nor remediated by distributions of resources.”

[Anderson 2010]

@AlanLundgard

11

According to Anderson: “Capability measures are sensitive to structural and psychosocial injustices that interfere with individuals’ functioning as equals, although [these injustices] are neither constituted nor remediated by distributions of resources.”

In other words, remediating injustice doesn’t only depend on (re)allocating resources more fairly, but also on how people are capable (or not) of using the resources in their possession. If fair machine learning aspires to address these injustices, then it will need to look beyond single-valued, quantitative resource measures. Capability measures provide one possible alternative.

# Measuring Justice in Machine Learning

Alan Lundgard

Massachusetts Institute of Technology

Conference on Fairness, Accountability, and Transparency

January 30, 2020

@AlanLundgard

12

Thank you.

## References

1. E. S. Anderson, “Justifying the Capabilities Approach to Justice,” in *Measuring Justice: Primary Goods and Capabilities*, H. Brighouse and I. Robeyns, Eds. Cambridge University Press, 2010, pp. 81–100.
2. H. Brighouse and I. Robeyns, *Measuring Justice : Primary Goods and Capabilities*. Cambridge; New York: Cambridge University Press, 2010.
3. T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, “Fairness Without Demographics in Repeated Loss Minimization,” in *International Conference on Machine Learning*, 2018, pp. 1929–1938.
4. C. Murphy and P. Gardoni, “Design, Risk and Capabilities,” in *The Capability Approach, Technology and Design*, I. Oosterlaken and J. van den Hoven, Eds. Springer Netherlands, 2012.
5. A. Procaccia, “AI Researchers Are Pushing Bias Out of Algorithms,” *Bloomberg*, 07-Mar-2019.
6. M. Bogen and A. Rieke, “Help Wanted: An Exploration of Hiring Algorithms, Equity, and Bias.” *Upturn*, 2018.

## References

1. E. S. Anderson, “Justifying the Capabilities Approach to Justice,” in *Measuring Justice: Primary Goods and Capabilities*, H. Brighouse and I. Robeyns, Eds. Cambridge University Press, 2010, pp. 81–100.
2. H. Brighouse and I. Robeyns, *Measuring Justice : Primary Goods and Capabilities*. Cambridge; New York: Cambridge University Press, 2010.
3. T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, “Fairness Without Demographics in Repeated Loss Minimization,” in *International Conference on Machine Learning*, 2018, pp. 1929–1938.
4. C. Murphy and P. Gardoni, “Design, Risk and Capabilities,” in *The Capability Approach, Technology and Design*, I. Oosterlaken and J. van den Hoven, Eds. Springer Netherlands, 2012.
5. A. Procaccia, “AI Researchers Are Pushing Bias Out of Algorithms,” *Bloomberg*, 07-Mar-2019.
6. M. Bogen and A. Rieke, “Help Wanted: An Exploration of Hiring Algorithms, Equity, and Bias.” *Upturn*, 2018.