Algorithmic Fairness, Intersectionality, and Uncertainty

Two Problems and Four Solutions

Johannes Himmelreich

Maxwell School of Citizenship and Public Affairs at Syracuse University

Overview

Structure

1. Statistical Uncertainty
2. Ontological Uncertainty
3. Solutions

Aims

- Reflect on theme accessibly
- Structure issues within theme
- Identify opportunities with practical and theoretical relevance

Based on joint work with Arbie Hsu, Ellen Veomett, and Kristian Lum [7]

Intersectional algorithmic fairness

Experiences of oppression depend on social identities that are constituted across several demographic categories simultaneously

How to audit algorithmic fairness intersectionally?

Statistical Uncertainty

Many Groups

Intersectionality Two channels [1]

1. **non-additive** effects of attributes
2. **switchy** (cond. prob.) effects

$\rightsquigarrow$ consider

- e.g., Maghrebi older women in France simultaneously
- instead of each ethnic origin, age, gender, location separately

Note: In Europe particularly relevant, but legally optional

Problem of many groups

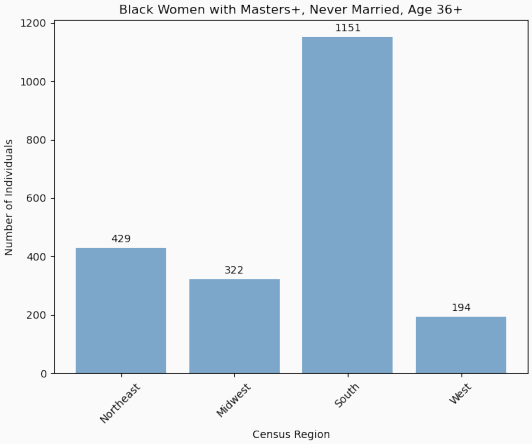
- Number of intersectional groups grows exponentially: $\prod k^n$ (for n k-valued attributes)
- 3 2-, 1 3-, 1 9-, and 1 12-valued attribute: 2,593 groups
- Policy upshot: Existing data not sufficiently disaggregated

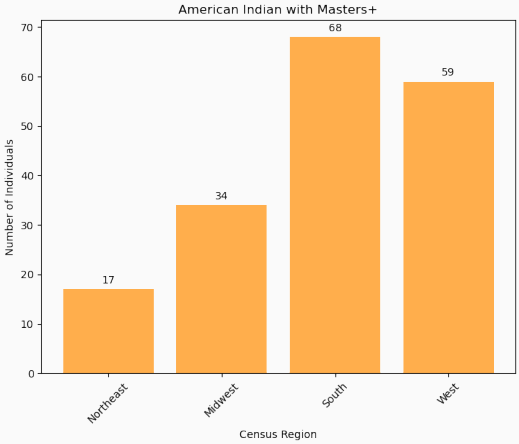
Small Groups: Dataset review

Dataset	Subgroup	*n*
compas	(Male, Native American, 25 - 45)	6
compas_violent	(Female, African-American, <25)	95
creditg	(male div/sep, male, <=25)	2
heart_disease	(female, >54)	103
meps20	(Non-White, 80's, female)	146
meps21	(Non-White, 80's, female)	142
nlsy	(Female, <18, GREEK)	2
ricci	W	68
student_math	(M, <18)	134
student_por	(M, >=18)	73
tae	1.0	29
titanic	(female, 60's)	10

Table 1: Subgroups with minimum accuracy on prediction task

Small Groups: ACS 2018, all states





What is the minimum # of observations we'd want? **1,538**

Back-of-the-Envelope Calculation

$$n \geq \frac{z_{1-\alpha/2}^2 \, p(1-p)}{m^2}$$

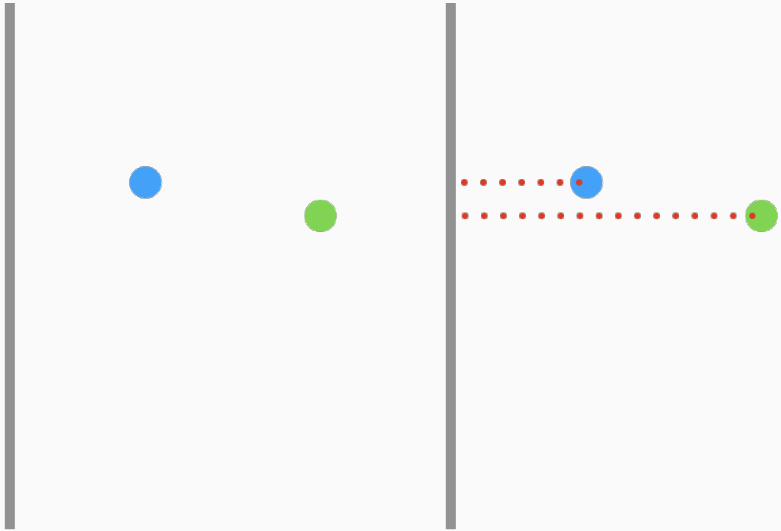
- $p = 0.20$ (anticipated error rate)
- $m = 0.02$ (tolerated half-width)
- $1 - \alpha = 0.95 \Rightarrow z_{1-\alpha/2} = 1.96$ (95% intervall)
- $p(1-p) = 0.20 \times 0.80 = 0.16$

$$n \geq \frac{1.96^2 \times 0.16}{0.02^2} = \frac{3.8416 \times 0.16}{0.0004} \approx 1538$$

Would need at least $\approx 1.6 \times 10^3$ observations per relevant group

Takeaway: Insufficient data for intersectional fairness, on typical approach

Problem of Statistical Uncertainty: Illustration



Problems for Typical Fairness Definitions

Fairness (typically)

$$|m(G) - m(\cdot)| < \epsilon$$

for some small $\epsilon, \forall\, G$

- $m(G)$ model performance for group g
- $m(\cdot)$ overall model performance
- i.e., accuracy, false-positive rate, ...

Problems of Statistical Uncertainty

- **Statistically biased** Over-estimates between-group disparity [9]
- **Point estimates $m(G)$ nonsensical** Fairness varies between samples/models on same population [3, 2]

Summary

Intersectionality $\rightsquigarrow$ many groups $\rightsquigarrow$ small groups $\rightsquigarrow$ nonsensical fairness estimates

Ontological Uncertainty

Social Ontology of Intersectionality

Question: Which groups $G \in \mathcal{G}$ warrant consideration in fairness audits?

Answer: *Some theory* that yields $\mathcal{G}$ — the set of groups to audit.

Problem: No such salient operational theory

$\rightsquigarrow$ **Fairness Gerrymandering**: Auditor can tweak $\mathcal{G}$ to skew audit results [8]

$\rightsquigarrow$ **Advocacy efficacy dampener**: What reforms of non-discrimination law to propose? What attribute disaggregation to prioritize?

Hypothesis: This social ontology problem is the place to incorporate hermeneutic accounts of identity production [13]

Competing Theories for $\mathcal{G}$

Schematically, space of theories: Relevant groups $\mathcal{G}$ are

(a) Intersections of *all* attributes — $\mathcal{G} = \prod_{A \in \mathcal{A}} A$

(b) Intersections of *"relevant"* attributes — $\mathcal{G} = \prod_{A \in \mathcal{A}} \pi(A)$ with "pruning" π

(c) A "curated" *list* of social groups — $\mathcal{G} = \{G_1, \ldots, G_k\}$

Attributes $A \in \mathcal{A}$, consisting of possible "values" of A

Hypothesis: Hermeneutic accounts fall into (c).

Challenge: Disagreements over right theory on both epistemic and normative grounds. E.g., Theory might rest on assumptions about historical injustice.

Solutions

Summary

Background concerns

- Pursue *algorithmic* fairness, despite its myopia
- Advance intersectionality — technically, legally, and theoretically

Two problems

1. **Statistical Uncertainty**: Intersectionality $\rightsquigarrow$ many groups $\rightsquigarrow$ small groups $\rightsquigarrow$ estimation problems
2. **Ontological Uncertainty**: Intersectionality $\rightarrow \mathcal{G}$, but how to get $\mathcal{G}$ is uncertainty, non-trivial (unorthodox?), essentially contested.

Solutions

1. Ontology $\mathcal{G}$ that is not vast
2. More data
3. Non-typical fairness "metrics"

} Problem of Statistical Uncertainty

Several proposals [5, 6, 4, 11, 10, 12]. Some might be problematic [7].

👍 Would solve *both* problems.

👍 Theoretically interesting and important (if hypothesis true)

Conclusion

Conclusion

Two problems, four solutions: Structured topic, identified opportunities within theme of intersectional algorithmic fairness

Finding $\mathcal{G}$ is worthwhile problem.

- Hermeneutical, not just analytical, intersectionality
- May solve problem of many groups (depending on size of $\mathcal{G}$)
- Policy-relevant: data disaggregation, non-discrimination law
- Broader importance than (myopic) algorithmic fairness

Questions?

References i

L. K. Bright, D. Malinsky, and M. Thompson.
Causally Interpreting Intersectionality Theory.
Philosophy of Science, 83(1):60–81, Jan. 2016.
Publisher: Cambridge University Press.

J. Choi, S. Karumbaiah, and J. Matayoshi.
Bias or Insufficient Sample Size? Improving Reliable Estimation of Algorithmic Bias for Minority Groups.
In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, pages 547–557, New York, NY, USA, Mar. 2025. Association for Computing Machinery.

References ii

- A. F. Cooper, K. Lee, M. Z. Choksi, S. Barocas, C. De Sa, J. Grimmelmann, J. Kleinberg, S. Sen, and B. Zhang.
  Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification.
  Proceedings of the AAAI Conference on Artificial Intelligence, 38(20):22004–22012, Mar. 2024.
- J. R. Foulds, R. Islam, K. N. Keya, and S. Pan.
  An intersectional definition of fairness.
  In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, 2020.

References iii

U. Gohar and L. Cheng.
A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges.
In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6619–6627, Aug. 2023.
arXiv:2305.06969 [cs].

C. Herlihy, K. Truong, A. Chouldechova, and M. Dudík.
A structured regression approach for evaluating model performance across intersectional subgroups.
In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 313–325, Rio de Janeiro Brazil, June 2024. ACM.

References iv


J. Himmelreich, A. Hsu, E. Veomett, and K. Lum.
The Intersectionality Problem for Algorithmic Fairness.
In *Proceedings of the Algorithmic Fairness Through the Lens of Metrics and Evaluation*, pages 68–95. PMLR, Apr. 2025.
ISSN: 2640-3498.


M. Kearns, S. Neel, A. Roth, and Z. Wu.
Preventing fairness gerrymandering: Auditing and learning for subgroup fairness.
In *The 35 th International Conference on Machine Learning*, 2018.

References v

- K. Lum, Y. Zhang, and A. Bower.
  De-biasing “bias” measurement.
  In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, pages 379–389, New York, NY, USA, June 2022. Association for Computing Machinery.
- M. Molina and P. Loiseau.
  Bounding and approximating intersectional fairness through marginal fairness.
  In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

References vi

- G. Morina, V. Oliinyk, J. Waton, I. Marusic, and K. Georgatzis.
  Auditing and achieving intersectional fairness in classification problems.
  CoRR, abs/1911.01468, 2019.
- L. M. Paes, A. T. Suresh, A. Beutel, F. P. Calmon, and A. Beirami.
  Multi-Group Fairness Evaluation via Conditional Value-at-Risk Testing.
  IEEE Journal on Selected Areas in Information Theory, 5:659–674, 2024.
  arXiv:2312.03867 [cs].
- E. Ruíz.
  Framing Intersectionality.
  In *The Routledge Companion to the Philosophy of Race*, pages 335–348. 2017.

Desiderata for Fairness Estimation

What is an adequate statistical setup — or: right "metric" — to analyze intersectional fairness?

Desiderata for Fairness Metrics

Minimal Justice

Incentive Compatibility

Minimal Justice

Any fairness metric...

Minimal Justice

should not encode a standard of fairness that is lower for certain groups

- "Don't disadvantage the disadvantaged"
- **Weakly prioritarian** Give these groups *at least the same* weight
- Example violation: Kearns et al. [8]

$$\alpha(G)|m(G) - m(\cdot)| < \epsilon \quad \forall G$$

where $\alpha(G) = Pr(G)$, proportion of group G in population
Violates Minimal Justice: importance of equality proportional to group size

Incentive Compatibility

Any fairness metric...

Incentive Compatibility

Should neither (a) discourage further data collection, nor (b) incentivize improving model performance with deliberate mistakes

- Often one *can* improve fairness by making deliberately inaccurate predictions (on group with high model performance)
- Example violation: Kearns et al. [8] can discourage minority group data collection

Desiderata: Overview

Any fairness metric...

Minimal Justice

should not encode a standard of fairness that is lower for certain groups

Incentive Compatibility

should neither (a) discourage further data collection, nor (b) incentivize improving model performance with deliberate mistakes