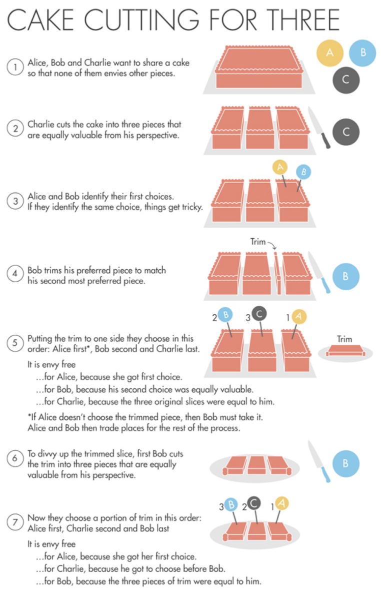
CS480/680: Introduction to Machine Learning

Lec 21: Algorithmic Fairness

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

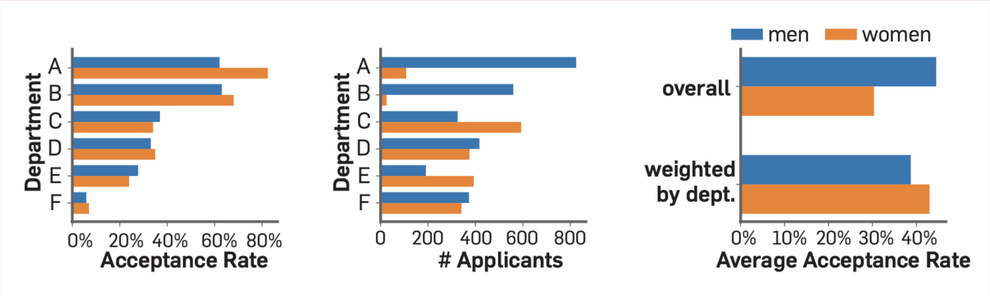
July 22, 2024



H. Aziz and S. Mackenzie. "A bounded and envy-free cake cutting algorithm". *Communications of the ACM*, vol. 63, no. 4 (2020), pp. 119–126.

Simpson's Paradox: Berkeley Admission Statistics (1973 fall)

Figure 2. UC Berkeley admissions statistics for men and women. Left: Acceptance rates. Middle: Number of applicants. Right: Average acceptance rate, either overall or weighted by the total number of applicants (of both groups) for each department.



- Overall acceptance rate for men was higher (44%) than for women (35%)
- For almost all departments, women enjoyed a higher acceptance rate than men

M. Buyl and T. D. Bie. "Inherent Limitations of AI Fairness". *Communications of the ACM*, vol. 67, no. 2 (2024), pp. 48–55.

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

- Developed by Northpointe in 1998, sold to Toronto-based Constellation Software in 2011
- Used in some US criminal justice systems
- Predicts a defendant's risk of committing a misdemeanor or felony *within 2 years*
  - proxy for lack of groundtruth (committing a crime)
- 137 features about an individual and the individual's past criminal record

Example Features in COMPAS

- Prior arrests and convictions
- Address of the defendant
- Whether the defendant a suspected gang member
- Whether the defendant ever violated parole
- If the defendant's parents separated
- If friends/acquaintances of the defendant were ever arrested
- Whether drugs are available in the defendants neighborhood
- How often the defendant has moved residences
- The defendants high school GPA
- How much money the defendant has
- How often the defendant feels bored or sad
- Age at the time of current offense
- Age at the time of first offense

One variable that doesn't appear is the defendant's race

White				Black			
		Actual				Actual	
		NR	R			NR	R
Predicted	NR	999	408	Predicted	NR	873	473
	R	282	414		R	641	1188
FN	0.50			FN	0.28		
FP	0.22			FP	0.42		

- Unequal base rates: $\frac{408+414}{408+414+282+999} \approx 39\%$ vs. $\frac{473+1188}{473+1188+873+641} \approx 52\%$
- Unequal odds: White higher False Negatives while Black higher False Positives
  - positive prediction (i.e., Recidivism) may be used by the judge against the defendent

---

A. W. Flores et al. "False Positives, False Negatives, and False Analyses: A Rejoinder". *Federal Probation*, vol. 80, no. 2 (2016), pp. 38–46.

J. Angwin et al. "Machine bias". 2016.

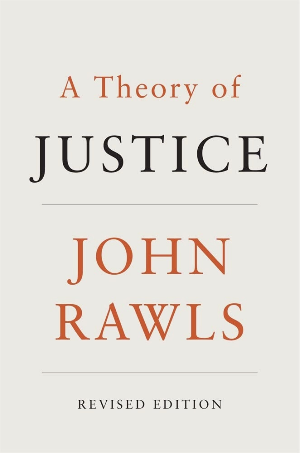
	All	White	Black
Low	32	29	35
Medium	55	53	56
High	75	73	75
Base Rate*	47	39	52
AUC	0.71	0.69	0.70

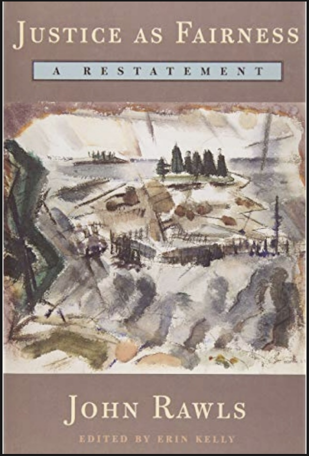
	All	**White**	**Black**
Low	11	9	13
Medium	26	22	27
High	45	38	47
Base Rate*	17	12	21
AUC	0.71	0.68	0.70

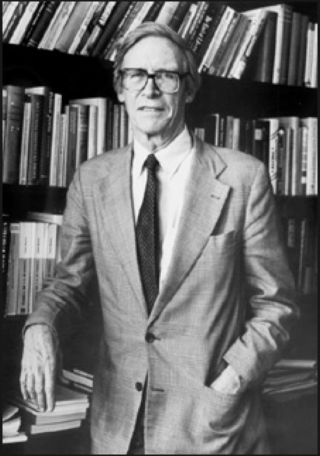
- Pr(Recidivism | race, risk score) roughly calibrated
  - left: any crime; right: violent crime only
- Accuracy parity: $\frac{414+999}{408+414+282+999} \approx 67\%$ vs. $\frac{873+1188}{473+1188+873+641} \approx 65\%$
- No demographic parity: $\frac{282+414}{408+414+282+999} \approx 33\%$ vs. $\frac{641+1188}{473+1188+873+641} \approx 58\%$

A. W. Flores et al. "False Positives, False Negatives, and False Analyses: A Rejoinder". *Federal Probation*, vol. 80, no. 2 (2016), pp. 38–46.

Each person possesses an inviolability founded on justice that even the welfare of society as a whole cannot override.







original position: people select what kind of society they would choose to live under if they did not know which social position they would personally occupy.

Setting

- Features for each individual: $\mathsf{X} \in \mathbb{R}^d$
- Binary labels: $\mathsf{Y} \in \{0, 1\}$
  - $\mathsf{Y} = 1$ being the preferred label, e.g., admission
- Sensitive attributes: $\mathsf{A} \in \{a, b\}$
  - partition individuals into groups
- Prediction (e.g., by an algorithm or human): $\hat{\mathsf{Y}} = \hat{\mathsf{Y}}(\mathsf{X}) \in [0, 1]$
- Disparate Treatment: prediction $\hat{\mathsf{Y}}$ depends on sensitive attribute A
  - often by law or moral: $\mathsf{A} \notin \mathsf{X}$ (Rawls' original position)
  - proxy: may still be able to predict A based on other features in X

Affirmative Action (AA)

- First introduced in US by President JFK in 1961: government contractors "take affirmative action to ensure that applicants are employed, and employees are treated during employment, without regard to their race, creed, color, or national origin."
- By President LBJ in 1965: government employers to take "affirmative action" to "hire without regard to race, religion and national origin."
- In 1965: gender was added to the list
- Grutter v. Bollinger (Supreme Court 2003) permitted educational institutions to consider race as a factor when admitting students
  - California, Michigan, and Washington banned preferential treatment

AA in Action

- Canada: the Canadian Charter of Rights and Freedoms explicitly permits affirmative action but does not require preferential treatment
  - The Canadian Employment Equity Act requires employers in federally-regulated industries to give preferential treatment to Women, persons with disabilities, aboriginal peoples, and visible minorities
- UK: quotas are illegal
- China: lower requirement for minorities in national university entrance exam; quota; dedicated financial aid/scholarship
- India: reservation system for majority (60% college admission or government jobs reserved for 90% majority)

Fairness Definition 1: Statistical/Demographic Parity

$$\mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = a) = \mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = b) = \mathbb{E}(\hat{\mathsf{Y}})$$

- For deterministic classifiers, i.e., $\hat{\mathsf{Y}} \in \{0, 1\}$, demographic parity means $\hat{\mathsf{Y}} \perp\!\!\!\perp \mathsf{A}$
- But, consider the following two scenarios:
  - scenario 1: For $\mathsf{A} = a$, accept top 10%; for $\mathsf{A} = b$, accept random 10%
  - scenario 2: $\mathsf{Y} = [\![\mathsf{A} = a]\!]$; may disallow (almost) perfect classifier...

Estimated Canadian breast cancer statistics (2024)

Category	Women	Men
New cases	30,500	290
Deaths	5,500	60
5-year net survival (estimates for 2015 to 2017)	89%	76%

https://cancer.ca/en/cancer-information/cancer-types/breast/statistics

Disparate Impact

- Griggs v. Duke Power Co. (1971, US Supremum Court)
  - 1950s: Duke Power held policy restricting black employees to its "Labor" dept.
  - 1955: Added requirement of high school diploma for employment in any dept. but Labor, and offered 2/3 training tuition for employee w/o diploma
  - 1965: Added 2 employment tests (mechanical & IQ) to allow employees w/o diploma to transfer to any dept.
  - Blacks were 10 times less likely to pass
- Supremum court ruling: if such tests disparately impact minority groups, businesses must demonstrate that such tests are "reasonably related" to the job for which the test is required

The Fox and the Stork

80% Rule

$$\frac{\mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = a) \ \wedge \ \mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = b)}{\mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = a) \ \vee \ \mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = b)} \geq \tau = 80\%$$

- Recall that $\mathsf{Y} = 1$ is the preferred label, e.g., hire
- Selection rate for the disadvantageous group (min) is at least 80% of that for the advantageous group (max)
- Advocated by the US Equal Employment Opportunity Commission (1979)
- Completely ignores the true label Y (qualification); quota or preferential treatment

M. Feldman et al. "Certifying and Removing Disparate Impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 259–268.

Fairness Definition 2: Equal Odds

$$\mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = a, \mathsf{Y} = y) = \mathbb{E}(\hat{\mathsf{Y}} \mid A = b, \mathsf{Y} = y), \quad \forall y \in \{0, 1\}$$

- For a deterministic classifier, i.e., $\hat{\mathsf{Y}} \in \{0,1\}$, equal odds means $\hat{\mathsf{Y}} \perp\!\!\!\perp \mathsf{A} \mid \mathsf{Y}$
- If true label $\mathsf{Y} = 1$: (generalization of) equal true positives
- If true lable $\mathsf{Y} = 0$: (generalization of) equal false positives

$$\mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A}) = \int \mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A}, \mathsf{Y} = y) \Pr(\mathsf{Y} = y \mid \mathsf{A}) \, \mathrm{d}y$$

- Equal odds implies demographic parity under equal base rates $\Pr(\mathsf{Y} = y \mid \mathsf{A})$

M. Hardt et al. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems 29*. 2016, pp. 3315–3323.

Fairness Definition 3: Equal Opportunity

$$\mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = a, \mathsf{Y} = 1) = \mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = b, \mathsf{Y} = 1)$$

- Recall $\mathsf{Y} = 1$ is the preferred label, e.g., loan approval
- $\mathsf{Y} = 1$: qualified applicants
- Among qualified applicants, equal true positives for different groups
- No requirement on unqualified applicants: maximal utility

M. Hardt et al. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems 29*. 2016, pp. 3315–3323.

Fairness Definition 4: Calibration

$$\Pr(\mathsf{Y} = 1 \mid \hat{\mathsf{Y}}, \mathsf{A} = a) = \hat{\mathsf{Y}} \in [0, 1], \quad \forall a$$

- For a deterministic classifier, i.e., $\hat{\mathsf{Y}} \in \{0, 1\}$, calibrated = perfect
- Among all instances that we predict positive with $\hat{\mathsf{Y}} = 80\%$ probability, indeed $\hat{\mathsf{Y}} = 80\%$ of them have true label 1
- Calibration is often desirable, but it may have little to do with accuracy
  - consider the constant predictor $\hat{\mathsf{Y}} = \mathbb{E}(\mathsf{Y})$: is it calibrated?
- True meaning: $f(\hat{\mathsf{Y}})$ is not more accurate than $\hat{\mathsf{Y}}$ for any post-processing f

G. W. Brier. "Verification of Forecasts Expressed in Terms of Probability". *Monthly Weather Review*, vol. 78, no. 1 (1950), pp. 1–3, M. H. DeGroot and S. E. Fienberg. "The comparison and evaluation of forecasters". *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 32, no. 1-2 (1983), pp. 12–22.

Inherent Tradeoff

Theorem: You can't have everything!

If a probabilistic classifier $\hat{\mathsf{Y}} = \hat{\mathsf{Y}}(\mathsf{X})$ satisfies

$$\text{(calibration)} \quad \mathbb{E}(\mathsf{Y} \mid \hat{\mathsf{Y}}, \mathsf{A} = a) = \mathbb{E}(\mathsf{Y} \mid \hat{\mathsf{Y}}, \mathsf{A} = b) = \hat{\mathsf{Y}}$$

$$\text{(equal odds)} \quad \mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = a, \mathsf{Y} = y) = \mathbb{E}(\hat{\mathsf{Y}} \mid \mathsf{A} = b, \mathsf{Y} = y), \quad \forall y \in \{0, 1\},$$

then either $\hat{\mathsf{Y}}$ is a perfect classifier or the base rates match, i.e.,

$$\forall y \in \{0, 1\}, \quad \Pr(\mathsf{Y} = y \mid \mathsf{A} = a) = \Pr(\mathsf{Y} = y \mid \mathsf{A} = b).$$

- Apply to any probabilistic classifier, algorithm based or human based
- When base rates differ, demographic parity contradicts calibration or equal odds

J. Kleinberg et al. "Inherent trade-offs in the fair determination of risk scores". In: *ITCS*. 2017, 43:1–43:23.

Estimated Canadian breast cancer statistics (2024)

Category	Women	Men
New cases	30,500	290
Deaths	5,500	60
5-year net survival (estimates for 2015 to 2017)	89%	76%

https://cancer.ca/en/cancer-information/cancer-types/breast/statistics

- Base rates clearly differ
- So far, no classifier is perfectly accurate
- Thus, any existing classifier (algorithmic or not) can meet at most one of demographic parity, calibration and equal odds!

Apply the definition of conditional expectation:

$$
\begin{aligned}
\mathbb{E}[\hat{\mathsf{Y}} \mid \mathsf{A}=a, \mathsf{Y}=0] &= \frac{\mathbb{E}\left[\hat{\mathsf{Y}}\,[\![\mathsf{Y}=0]\!] \mid \mathsf{A}=a\right]}{\Pr[\mathsf{Y}=0 \mid \mathsf{A}=a]} \\
&= \frac{\mathbb{E}\left[\hat{\mathsf{Y}}(1-[\![\mathsf{Y}=1]\!]) \mid \mathsf{A}=a\right]}{\Pr[\mathsf{Y}=0 \mid \mathsf{A}=a]} \\
&= \frac{\mathbb{E}[\hat{\mathsf{Y}} \mid \mathsf{A}=a] - \mathbb{E}\left[\hat{\mathsf{Y}}\,[\![\mathsf{Y}=1]\!] \mid \mathsf{A}=a\right]}{\Pr[\mathsf{Y}=0 \mid \mathsf{A}=a]} \\
\text{(follows from calibration)} \quad &= \frac{\mathbb{E}[\mathsf{Y} \mid \mathsf{A}=a] - \mathbb{E}[\hat{\mathsf{Y}} \mid \mathsf{A}=a, \mathsf{Y}=1] \cdot \Pr(\mathsf{Y}=1 \mid \mathsf{A}=a)}{\Pr[\mathsf{Y}=0 \mid \mathsf{A}=a]} \\
&= \frac{\Pr[\mathsf{Y}=1 \mid \mathsf{A}=a]}{\Pr[\mathsf{Y}=0 \mid \mathsf{A}=a]} \cdot \left(1 - \mathbb{E}\left[\hat{\mathsf{Y}} \mid \mathsf{A}=a, \mathsf{Y}=1\right]\right)
\end{aligned}
$$

From equal odds: $\mathbb{E}[\hat{\mathsf{Y}} \mid \mathsf{Y}=1] = 1$ implies $\hat{\mathsf{Y}} \geq \mathsf{Y}$; but from calibration: $\mathbb{E}[\hat{\mathsf{Y}}] = \mathbb{E}[\mathsf{Y}]$.

Fairness Definition 5: Individual Fairness

- Similar individuals should be treated similarly
- Transitivity can easily kill us: if a is similar to b, b is similar to c, ..., then we are forced to call a similar to z, even when they are very different

$$\mathrm{dist}\left(\hat{Y}(X), \hat{Y}(Z)\right) \leq \mathrm{dist}\left(X, Z\right)$$

- In other words, our predictor $\hat{Y}$ needs to be Lipschitz continuous
- But, finding an agreeable distance function is difficult

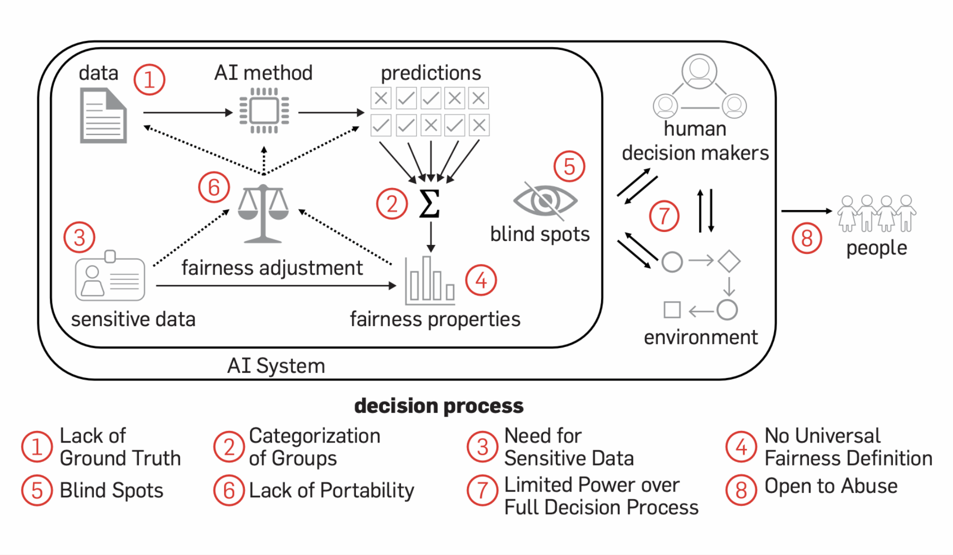
C. Dwork et al. "Fairness Through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 2012, pp. 214–226.

Some Perils of Algorithmic Fairness

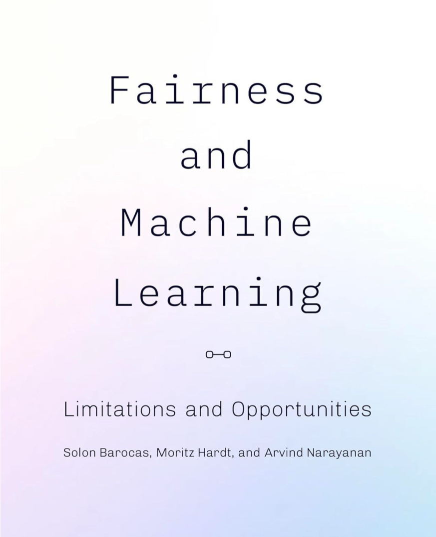
- Limited access to ground-truth label; often resort to questionable proxies
  - commit a crime $\approx$ arrested by police; neither one implies the other
- Need to collect sensitive attributes, something explicitly banned by AA
  - Proposed European AI Act allows processing sensitive data for bias monitoring, detection and correction
- No universally agreed definition (probably never will)
- Limited power over the entire decision pipeline
  - one would be naive to think algorithmic fairness can solve social issues all by itself
- Open to abuse

M. Buyl and T. D. Bie. "Inherent Limitations of AI Fairness". *Communications of the ACM*, vol. 67, no. 2 (2024), pp. 48–55.

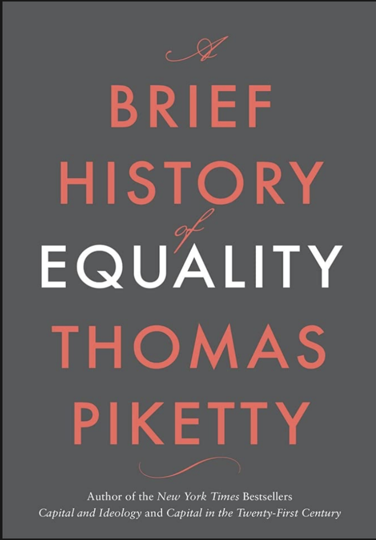
Figure 1. A prototypical fair AI system. Each limitation affects a different component of the full decision process.



 M. Buyl and T. D. Bie. "Inherent Limitations of AI Fairness". *Communications of the ACM*, vol. 67, no. 2 (2024), pp. 48–55.


Fairness
and
Machine
Learning
Limitations and Opportunities
Solon Barocas, Moritz Hardt, and Arvind Narayanan


michael kearns + aaron roth
the
ethical
algorithm
the science of socially aware algorithm design


A
BRIEF
HISTORY
of
EQUALITY
THOMAS
PIKETTY
Author of the New York Times Bestsellers
Capital and Ideology and Capital in the Twenty-First Century

Other Fairness Definitions

- Accuracy parity:

$$\Pr(\hat{\mathsf{Y}} = \mathsf{Y} \mid \mathsf{A} = a) = \Pr(\hat{\mathsf{Y}} = \mathsf{Y} \mid \mathsf{A} = b)$$

  or more generally for a probabilistic classifier:

$$\mathbb{E}\left[\hat{\mathsf{Y}} \cdot \mathsf{Y} + (1 - \hat{\mathsf{Y}})(1 - \mathsf{Y}) \,\middle|\, \mathsf{A} = a\right] = \mathbb{E}\left[\hat{\mathsf{Y}} \cdot \mathsf{Y} + (1 - \hat{\mathsf{Y}})(1 - \mathsf{Y}) \,\middle|\, \mathsf{A} = b\right]$$

- More generally, we can compare the conditional distributions induced by different groups using any risk measure or divergence
- Causality/Counterfactual based

R. Williamson and A. Menon. "Fairness risk measures". In: *Proceedings of the 36th International Conference on Machine Learning*. 2019, pp. 6786–6797.

N. Kilbertus et al. "Avoiding Discrimination through Causal Reasoning". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 656–666, M. J. Kusner et al. "Counterfactual Fairness". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 4066–4076.