What's Fair about Individual Fairness?

Will Fleisher

Northeastern University

Individual Fairness (IF)

- First proposed by Dwork et al (2012).
- Offers alternative to dominant group fairness (GF) paradigm.
- IF is motivated by principle of similar treatment:
  - Similar individuals should be treated similarly.
- IF encodes similar treatment to provide a precise definition.
- IF is a condition imposed using two distance measures:
  - *d(x,y)*: A similarity metric measuring how similar any two individuals *x* and *y* are with respect to some task.
  - *D(Mx,My)*: A measure of how similarly *x* and *y* are treated.
  - Defines fairness as requiring that difference in treatment be no larger than difference in similarity.

Individual Fairness (IF): A mapping $M : V \to \Delta(A)$ satisfies the (D, d)-*Lipschitz* property if for every $x, y \in V$, we have:

$$D(Mx, My) \leq d(x, y)$$

(Dwork et al 2012)

- Crucial for IF is obtaining an appropriate similarity metric *d*.
  - This is what allows IF to be used to evaluate and promote fairness.
  - Primary practical hurdle is obtaining a metric for tasks we wish to evaluate.
- Arguments for IF:
  1. Captures intuitive notion of fairness (similar treatment).
  2. Forbids various kinds of intuitively unfair treatments, including those allowed by GF.

Problem 1: Insufficiency

- IF is insufficient to ensure fairness.
  - There are a variety of counterexamples where IF is satisfied, but the treatment of individuals is clearly unfair.
  - Cases involving systematically making individuals worse-off, for no good reason, while still treating similar individuals similarly.

Universal Rejection Consider a system that offers advice on college admissions decisions. In this case, the system simply recommends denying every application. Here, similar individuals are treated similarly, because everyone is treated similarly: everyone is denied admission. Despite this, individuals who have the ability to succeed in college, but who are denied the opportunity, can rightly complain that the situation is unfair.

- This undermines IF as a definition of fairness and the argument for treating IF as a replacement for GF.

Problem 2: Systematic Bias

- One important method for obtaining similarity metrics is by appeal to judgments made by human arbiters (Ilvento 2019).
- However, it is well known that people harbor implicit and explicit biases.
  - These biases are often systematic, rather than merely being random errors, particularly against marginalized groups.
  - Increasing the number of surveyed arbiters will not wash out the biases.
- This leaves IF vulnerable to inheriting the very biases algorithmic fairness methods are designed to alleviate.

Problem 3: Prior Moral Judgments

- Determining the right distance measure depends on making moral judgments about which features of individuals are fair to consider task-relevant.
  - Whether it is fair to include a feature, and how it is fair to do so, must be determined before building the similarity metric.

Discriminatory Admissions A system *M* is used to assist university admissions. Its task is to select successful students: those who are likely to graduate with high GPAs and obtain good jobs. *M* is trained on historical data from the university's admissions committee, learning to mimic past human decisions. *M* thus learns a preference for white applicants. This turns out to be highly accurate for the task of choosing successful students as the university environment and job market are filled with implicit and explicit racism.

Affirmative Action Admissions *M* is being used by a different university. As before, it is used for admissions with a goal of selecting students who will be successful. It is again trained on historical data. However, this university aims to promote diversity and uses various methods of affirmative action. For instance, it adds some points to black applicants' SAT scores, in response to bias in the test itself, and to the fact that students from this group typically have less access to test preparation. Thus, *M* learns to admit more black applicants for this university.

- The first case is unfair, the second is (at least more) fair.
- Building a distance metric reflecting this difference depends on prior judgments about fairness.
  - Specifically, about how racial differences should be used to inform similarity.
- This undermines IF as an informative, non-circular definition of fairness.
  - IF was supposed to inform us about the nature of fairness.
  - But it turns out to depend on antecedent knowledge of fairness.
- Note: this does not undermine IF's usefulness as an aggregation of such judgments, only claims that it is definitional or uniquely captures the concept.

Problem 4: Incommensurable Values

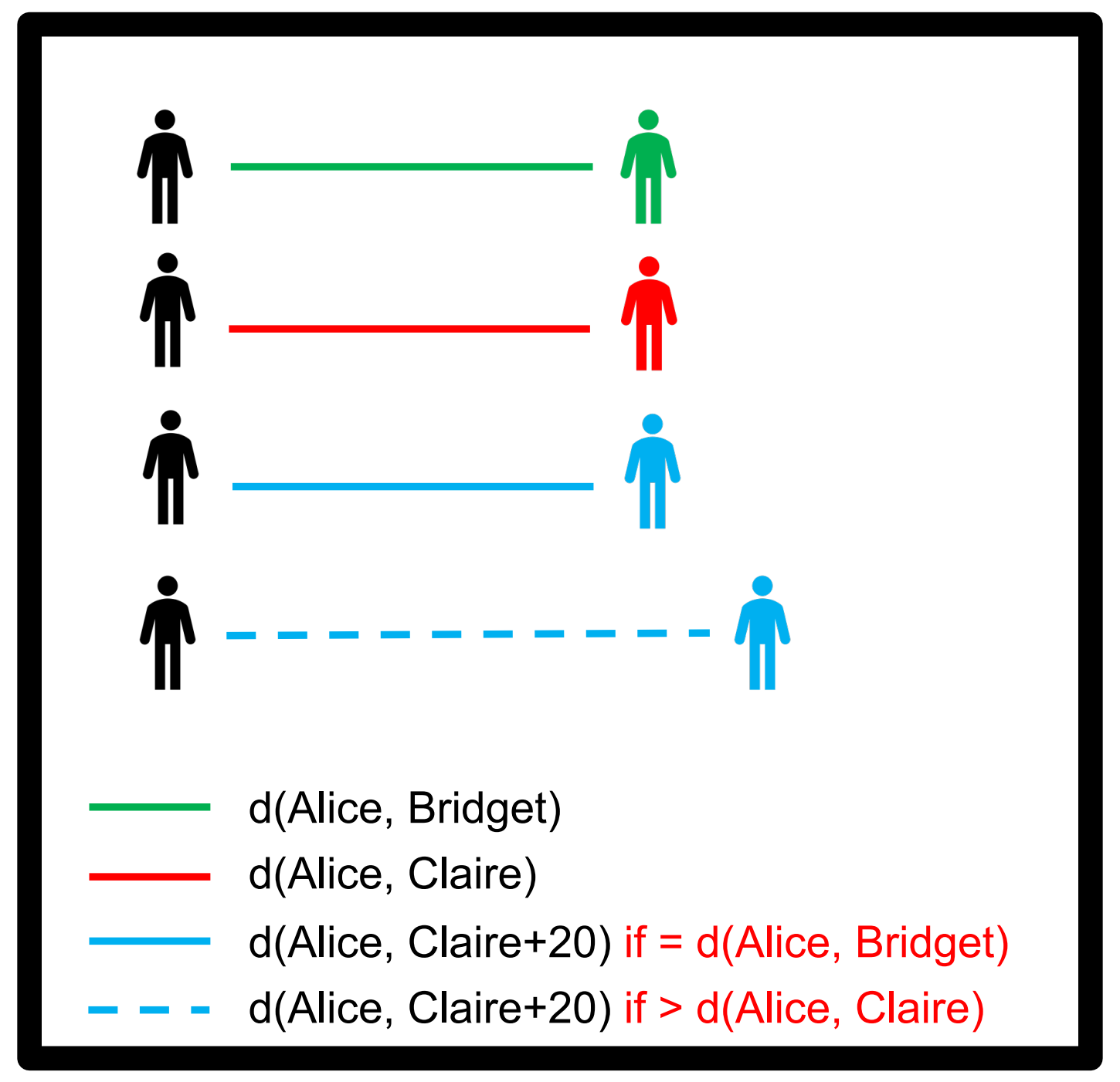
- Some moral values are incommensurable: they cannot be compared along a common measure. They cannot be exchanged in a straightforward manner.
- Incommensurability is illustrated by cases of insensitivity to sweetening: hard choices whose difficulty is not alleviated by small tie-breakers.
  - Imagine you have a hard choice between an academic and private sector career. Then imagine someone offers to "break the tie" by giving you $20 to stay in academia. It is intuitively still rational for you to remain undecided, even though you prefer academia with $20 to academia without it.
- It is impossible to build a similarity metric for incommensurable values:

Hard Admissions Choice: University admissions must decide between Bridget and Claire. The committee is seeking to promote moral values including intellectual achievement, social good, and diversity. The candidates have identical GPAs and SATs. Bridget has a history of charity work. Claire won several student science competitions. The committee finds the choice difficult to make and is unsure which applicant to admit. They end their first meeting undecided. At the next meeting, they learn Claire has taken the SAT again, and this time received a score of 1320, making her a better candidate than before. Must they now choose Claire?

- Value of charity work is incommensurable with value of supporting women scientists.
- The committee is rational (and morally permitted) to be undecided between Bridget and Claire, and between Bridget and Claire + 20 SAT points.
- It is impossible to build a similarity function for this case that reflects the underlying values:
  - Imagine Alice is the ideal student.
  - Bridget and Claire are on a par, so equally far from Alice: d(Alice,Bridget) = d(Alice,Claire).
  - Bridget and Claire+ are still on a par, so d(Alice,Bridget) = d(Alice,Claire+20).
  - Claire+20 is closer to Alice than Claire: d(Alice,Claire+20) > d(Alice,Claire).
  - Assigning real numbers to these distances would require satisfying the following, impossible statement:

d(A,B) = *d(A,C)* > *d(A,C+)* = *d(A,B)*

Illustration of Impossibility of real-valued similarity measure for ***Hard Admissions Choice***


d(Alice, Bridget)
d(Alice, Claire)
d(Alice, Claire+20) if = d(Alice, Bridget)
d(Alice, Claire+20) if > d(Alice, Claire)

Conclusions

These problems support two primary conclusions:

1. IF does not provide an adequate definition of fairness
2. IF should not be used as the sole means for evaluating fairness.

- The point is not that IF methods should be abandoned.
- IF is one important method for promoting fairness among several.

Thanks to Tina Eliassi-Rad, Branden Fitelson, D Black, Frank Wu, David Liu, and Lisa Miracchi.

Will Fleisher. 2021. What's Fair about Individual Fairness?. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), https://doi.org/10.1145/3461702.3462621