Introduction and Algorithmic Fairness (Part 2)

Responsible Data Science
DS-UA 202 and DS-GA 1017

Weeks 3–4

Instructors: Julia Stoyanovich and George Wood

This reader contains links to online materials and excerpts from selected articles on introduction to responsible data science and on algorithmic fairness. For convenience, the readings are organized by course week. Please note that some excerpts end in the middle of a section. Where that is the case, the partial section is not required reading.

Week 3: Fairness and Causality

RUTH FREMSON/NYT/REDUX/EYEVINE

A migrant farm worker has her fingerprints scanned so that she can register for a national identity card in India.

The long road to fairer algorithms

Matt J. Kusner & Joshua R. Loftus

Build models that identify and mitigate the causes of discrimination.

An algorithm deployed across the United States is now known to underestimate the health needs of black patients[1]. The algorithm uses health-care costs as a proxy for health needs. But black patients' health-care costs have historically been lower because systemic racism has impeded their access to treatment – not because they are healthier.

This example illustrates how machine learning and artificial intelligence can maintain and amplify inequity. Most algorithms exploit crude correlations in data. Yet these correlations are often by-products of more salient social relationships (in the health-care example, treatment that is inaccessible is, by definition, cheaper), or chance occurrences that will not replicate.

To identify and mitigate discriminatory relationships in data, we need models that capture or account for the causal pathways that give rise to them. Here we outline what is required to build models that would allow us to explore ethical issues underlying seemingly objective analyses. Only by unearthing the true causes of discrimination can we build algorithms that correct for these.

Causal models

Models that account for causal pathways have three advantages. These 'causal models' are: tailored to the data at hand; allow us to account for quantities that aren't observed; and address shortcomings in current concepts of fairness (see 'Fairness four ways').

A causal model[2] represents how data are generated, and how variables might change in response to interventions. This can be shown as a graph in which each variable is a node and arrows represent the causal connections between them. Take, for example, a data set about who gets a visa to work in a country. There is information about the country each person comes from, the work they do, their religion and whether or not they obtained a visa (see 'Three causal tests', part 1).

This model says that the country of origin directly influences a person's religion and whether they obtain a visa; so, too, do religion and type of work. Having a causal model allows us to address questions related to ethics, such as does religion influence the visa process?

But because many different causal models could have led to a particular observed data set, it is not generally possible to identify the right causal model from that data set alone[3]. For example, without any extra assumptions, data generated from the causal graph described here could seem identical to those from a graph in which religion is no longer linked to visa granting. A modeller must therefore also leverage experiments and expert knowledge, and probe assumptions.

Experiments can help in identifying factors that affect fairness. For example, a modeller wishing to explore whether ethnicity would affect treatment recommendations made online by health-care professionals could create two patient profiles that differ only in some respect that relates to ethnicity. For instance, one profile could have a name common to Americans of Chinese descent, and the other a name common to Americans of African descent. If the treatment recommendations are the same, then names can be ruled out as a source of bias, and the model can be stress-tested in another way.

Few aspects of a deep, multifaceted concept can be tested as easily as changing a name. This means that experimental evidence can underestimate the effects of discrimination. Integration of expert knowledge, particularly from the social sciences and including qualitative methods, can help to overcome such limitations. This knowledge can be used to, for example, inform the modeller of variables that might be influential but unobserved (lighter circles in 'Three causal tests'), or to determine where to put arrows.

Assumptions about unobserved variables that might alter the predictions of a model need to be clearly stated. This is particularly important when experiments cannot be run or more detailed expert knowledge is not available. For example, if 'health-care access' is not observed in a model attempting to predict 'health need', then it is crucial to identify any potential impacts it might have on 'health costs' as well as how it is affected by 'ethnicity'.

This need for context and metadata makes causal models harder to build than non-causal ones. It can also make them a more powerful way to explore ethical questions.

Three tests

Causal models can test the fairness of predictive algorithms in three ways.

Counterfactuals. A causal model allows us to ask and answer questions such as 'Had the past been different, would the present or future have changed?' In the visa example (see 'Three causal tests', part 1), algorithmic biases could be smoked out by tweaking parts of the model to explore, for instance: 'Had individual X been Christian, would this algorithm have granted them a visa?' A researcher could then identify what pieces of information an algorithm could use to achieve counterfactual fairness[4]: the algorithm's output would not change regardless of the individual's religion. For example, if the algorithm used just work and not country of origin or religion, it would satisfy counterfactual fairness.

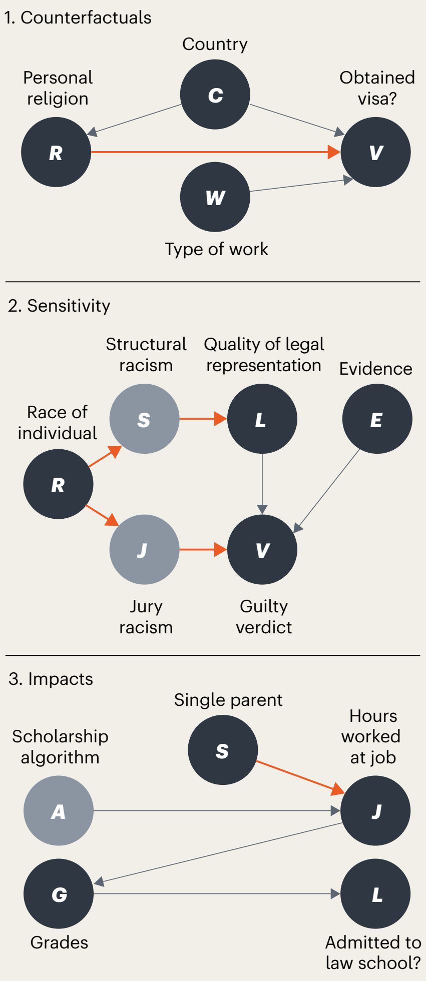
Sensitivity. In many settings, unknowns alter knowns – data we can observe are influenced by data we cannot. Consider a causal model for a trial setting (see 'Three causal tests', part 2).

This model shows how two independent sets of unobserved quantities, structural racism and jury racism, can unfairly lead to a guilty verdict. Although researchers often cannot precisely identify unobserved variables, they can reason about how sensitive a model is to them. For instance, they can explore how sensitive our estimate of the causal link between legal representation and guilty verdict is to different levels of jury racism. Simulations of the worst-case bias scenarios (that is, when jury racism is highest) can then be used to alter jury selection to minimize the bias.

Impacts. Data-driven decisions can have long-term consequences and spillover effects. These effects might not be obvious, especially in the standard machine-learning paradigm of predicting one short-term outcome. But carefully designed causal models can help researchers to use 'interventions' to probe the ripple effects of decisions far into the future[5,6]. For instance, the models can help regulatory agencies to understand how changing a scholarship algorithm influences who is accepted into law school (see 'Three causal tests', part 3). In this example, a single parent might need a scholarship so that they can reduce the hours they need to spend at a job, leaving them more time for study. That boosts their grades and therefore influences their chances of being admitted to law school. This complex chain can be explored using causal models.

THREE CAUSAL TESTS

Algorithmic fairness can be examined in different ways.



Five steps

Causal models are powerful tools, but they must be used appropriately. They are only models, and will thus fail to capture important aspects of the real world. Here we offer some guidelines on using them wisely.

Collaborate across fields. Researchers in statistics and machine learning need to know more about the causes of unfairness in society. They should work closely with those in disciplines such as law, social sciences and the humanities. This will help them to incorporate the context of the data used to train the algorithms. For example, scholars should meet at interdisciplinary workshops and conferences. One such is next year's Association for Computing and Machinery (ACM) conference on Fairness, Accountability and Transparency to derive a set of causal models for setting bail price and for immigration decisions.

A great example of such collaborations is one between information scientist Solon Barocas at Cornell University in Ithaca, New York, and attorney Andrew Selbst at the Data & Society Research Institute in New York City. They described how current law is unable to deal with algorithmic bias[7]. Partly in response to this work, machine-learning researchers have launched a large subfield, known as algorithmic fairness, that looks into ways of removing bias from data. And we and other researchers now use causal models to quantify discrimination due to data.

Partner with stakeholders. Predictive algorithms should be developed with people they are likely to affect. Stakeholders are best placed to provide input on difficult ethical questions and historical context. One example is the work by statistician Kristian Lum at the Human Rights Data Analysis Group in San Francisco, California, which investigates criminal-justice algorithms[8]. Such algorithms decide whether to detain or release arrested individuals and how high to set their bail, yet they are known to be biased. Lum has invited people affected by such decisions to speak at academic conferences attended by people who research these algorithms. This has led to closer collaboration, including the tutorial 'Understanding the context and consequences of pre-trial detention' presented at the 2018 ACM conference on Fairness, Accountability and Transparency in New York. So far, most stakeholder work has focused on criminal justice. Another setting that would benefit from it is mortgage lending.

We propose that a rotating, interdisciplinary panel of stakeholders investigates the impacts of algorithmic decisions, for example as part of a new international regulatory institute.

Make the workforce equitable. Women and people from minority groups are under-represented in the fields of statistics and machine learning. This directly contributes to the creation of unfair algorithms. For example, if facial detection software struggles to detect faces of black people[9], it is likely the algorithm was trained largely on data representing white people. Initiatives such as Black in AI (go.nature.com/38pbcaa) or Women in Machine Learning

Fairness four ways

A flurry of work has conceptualized fairness. Here are some of the most popular, and ways in which causal models offer alternatives.

Fairness through unawareness[12]. This method works by removing any data that are considered *prima facie* to be unfair. For example, for an algorithm used by judges making parole decisions, fairness through unawareness could dictate that data on ethnic origin should be removed when training this algorithm, whereas data on the number of previous offences can be used. But most data are biased. For instance, number of previous offences can bear the stamp of historical racial bias in policing, as can the use of plea bargaining (pleading guilty being more likely to reduce a sentence than arguing innocence)[13]. This can leave researchers with a hard choice: either remove all data or keep biased data.

Alternatively, causal models can directly quantify how data are biased.

Demographic parity[14]. A predictive algorithm satisfies demographic parity if, on average, it gives the same predictions to different groups. For example, a university-admissions algorithm would satisfy demographic parity for gender if 50% of its offers went to women and 50% to men. It is currently more common in law to relax demographic parity so that predictions aren't necessarily equal, but are not too imbalanced. Specifically, the US Equal Employment Opportunity Commission states that fair employment should satisfy the 80% rule: the acceptance rate for any group should be no less than 80% of that of the highest-accepted group. For instance, if 25% of women were offered jobs, and this is the highest acceptance rate, then at least 20% of men must be offered jobs[4]. One criticism of demographic parity is that it might not make sense to use it in certain settings, such as a fair arrest rate for violent crimes (men are significantly more likely to commit acts of violence)[15].

Instead, one could require that counterfactual versions of the same individual should get the same prediction[4].

Equality of opportunity[16]. This is the principle of giving the same beneficial predictions to individuals in each group. Consider a predictive algorithm that grants loans only to individuals who have paid back previous loans. It satisfies 'disability-based equality of opportunity' if it grants loans to the same percentage of individuals who both pay back and have a disability as it does to those who pay back and who do not have a disability. However, being able to pay back a loan in the first place can be affected by bias: discriminatory employers might be less likely to hire a person with a disability, which can make it harder for that person to pay back a loan. This societal unfairness is not captured by equality of opportunity.

A causal model could be used to quantify the bias and estimate an unbiased version of loan repayment.

Individual fairness[17]. This concept states that similar individuals should get similar predictions. If two people are alike except for their sexual orientation, say, an algorithm that displays job advertisements should display the same jobs to both. The main issue with this concept is how to define similar. In this example, training data will probably have been distorted by the fact that one in five individuals from sexual or gender minorities report discrimination against them in hiring, promotions and pay[18]. Thus similarity is hard to define, which makes individual fairness hard to use in practice.

In causal modelling, counterfactuals offer a natural way to define a similar individual. **M.J.K. & J.R.L.**

(go.nature.com/2s5km5g) are positive steps.

And we can go further. Causal models can themselves help to address the field's 'pipeline problem' by identifying where unfairness enters the process and which interventions can increase the participation of under-represented groups without shifting the burden to extra work for role models in those groups. Academic institutions should critically evaluate and use these models for fairer admissions in fields related to artificial intelligence.

Identify when algorithms are inappropriate. Statistics and machine learning are not all-powerful. Some problems should not be solved by expanding data-gathering capabilities and automating decisions. For example, a more accurate model for predictive policing won't solve many of the ethical concerns related to the criminal legal system. In fact, these methods can mask structural issues, including the fact that many neighbourhoods are policed by people who do not live in them[10]. This disconnect means that police officers might not be invested in the community they police or the people they arrest.

There are red flags when demographics, such as ethnic origin, influence nearly every piece of information in a causal graph, or when previous attempts to address a bias failed because people strategically changed behaviours in response. In these cases, an algorithmic solution would paper over a system that needs fundamental change.

Foment criticism. A vibrant culture of feedback is essential. Researchers need to continually question their models, evaluation techniques and assumptions. Useful as causal models are, they should be scrutinized intensely: bad models can make discrimination worse[11]. At the very least, a scientist should check whether a model has the right data to make causal claims, and how much these claims would change when the assumptions are relaxed.

Algorithms are increasingly used to make potentially life-changing decisions about people. By using causal models to formalize our understanding of discrimination, we must build these algorithms to respect the ethical standards required of human decision makers.

The authors

Matt J. Kusner is an associate professor in the Department of Computer Science at University College London, and a fellow at the Alan Turing Institute, London, UK. **Joshua R. Loftus** is an assistant professor in the Department of Technology, Operations, and Statistics at New York University, New York, USA.
e-mails: matt.kusner@gmail.com; loftus@nyu.edu

1. Obermeyer, Z. *et al. Science* **366**, 447–453 (2019).
2. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge Univ. Press, 2000).
3. Spirtes, P. *et al. Causation, Prediction, and Search* (MIT Press, 2000).
4. Kusner, M. J., Loftus, J., Russell, C. & Silva, R. In *Advances in Neural Information Processing Systems* 4066–4076 (MIT Press, 2017).
5. Liu, L. T. *et al.* In *International Conference on Machine Learning* 3150–3158 (ACM, 2018).
6. Kusner, M., Russell, C., Loftus, J. & Silva, R. *Proc. Machine Learning Res.* **97**, 3591–3600 (2019).
7. Barocas, S. & Selbst, A. D. *Calif. L. Rev.* **104**, 671 (2016).
8. Lum, K. *Nature Hum. Behav.* **1**, 0141 (2017).
9. Simon, M. 'HP looking into claim webcams can't see black people.' (CNN Tech, 23 December 2009).
10. McManus, H. D. *et al. Race Justice* https://doi.org/10.1177/2153368719849486 (2019).
11. Kilbertus, N. *et al.* 'The Sensitivity of Counterfactual Fairness to Unmeasured Confounding'. In *Uncertainty in Artificial Intelligence* (AUAI, 2019).
12. Grgic-Hlaca, N. *et al.* 'The case for process fairness in learning: Feature selection for fair decision making.' *NeurIPS Symposium on Machine Learning and the Law* (2016).
13. Wilford, M. M. & Khairalla, A. in *Social Sciences Contributions to the Real Legal System* Ch. 7, 132 (2019).
14. Zafar, M. B., Valera, I., Rogriguez, M. G. & Gummadi, K. P. In *Artificial Intelligence and Statistics* 962–970 (2017).
15. Dobash, R. E., Dobash, R. P., Cavanagh, K. & Lewis, R. *Violence Against Women* **10**, 577–605 (2004).
16. Hardt, M., Price, E. & Srebro, N. 'Equality of opportunity in supervised learning'. In *Advances in Neural Information Processing Systems* 3315–3323 (2016).
17. Dwork, C. *et al.* 'Fairness through awareness'. In *Proc. 3rd Innov. Theoret. Comp. Sci. Conf.* 214–226 (2012).
18. Pizer, J. C. *et al. Loy. LAL Rev.* **45**, 715 (2011).

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (*1*–*3*). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (*4*), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (*5*), and image searches for professions such as CEO produce fewer images of women (*6*). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (*7*, *8*), and natural language processing algorithms encode language in gendered ways (*9*).

Empirical investigations of algorithmic bias, though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with researcher-created algorithms (*10*–*13*). Without an algorithm's training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year. Large health systems and payers rely on this algorithm to target patients for "high-risk care management" programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs (*14*–*17*). Because the programs are themselves expensive—with costs going toward teams of dedicated nurses, extra primary care appointment slots, and other scarce resources—health systems rely extensively on algorithms to identify patients who will benefit the most (*18*, *19*).

Identifying patients who will derive the greatest benefit from these programs is a challenging causal inference problem that requires estimation of individual treatment effects. To solve this problem, health systems make a key assumption: Those with the greatest care needs will benefit the most from the program. Under this assumption, the targeting problem becomes a pure prediction policy problem (*20*). Developers then build algorithms that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (*21*).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard—e.g., number of lives affected, life-and-death consequences of the decision—health is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

[1]School of Public Health, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, USA. [3]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [4]Mongan Institute Health Policy Center, Massachusetts General Hospital, Boston, MA, USA. [5]Booth School of Business, University of Chicago, Chicago, IL, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: sendhil.mullainathan@chicagobooth.edu

and ethnic identities. Our main sample thus consisted of (i) 6079 patients who self-identified as Black and (ii) 43,539 patients who self-identified as White without another race or ethnicity, whom we observed over 11,929 and 88,080 patient-years, respectively (1 patient-year represents data collected for an individual patient in a calendar year). The sample was 71.2% enrolled in commercial insurance and 28.8% in Medicare; on average, 50.9 years old; and 63% female (Table 1).

For these patients, we obtained algorithmic risk scores generated for each patient-year. In the health system we studied, risk scores are generated for each patient during the enrollment period for the system's care management program. Patients above the 97th percentile are automatically identified for enrollment in the program. Those above the 55th percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment.

Many existing metrics of algorithmic bias may apply to this scenario. Some definitions focus on calibration [i.e., whether the realized value of some variable of interest Y matches the risk score R (*2*, *22*, *23*)]; others on statistical parity of some decision D influenced by the algorithm (*10*); and still others on balance of average predictions, conditional on the realized outcome (*22*). Given this multiplicity and the growing recognition that not all conditions can be simultaneously satisfied (*3*, *10*, *22*), we focus on metrics most relevant to the real-world use of the algorithm, which are related to calibration bias [formally, comparing Blacks B and Whites W, $E[Y|R,W] = E[Y|R,B]$ indicates the absence of bias (here, E is the expectation operator)]. The algorithm's stated goal is to predict complex health needs for the purpose of targeting an intervention that manages those needs. Thus, we compare the algorithmic risk score for patient i in year t ($R_{i,t}$), formed on the basis of claims data $X_{i,(t-1)}$ from the prior year, to data on patients' realized health $H_{i,t}$, assessing how well the algorithmic risk score is calibrated across race for health outcomes $H_{i,t}$. We also ask how well the algorithm is calibrated for costs $C_{i,t}$.

To measure H, we link predictions to a wide range of outcomes in electronic health record data, including all diagnoses (in the form of International Classification of Diseases codes) as well as key quantitative laboratory studies and vital signs capturing the severity of chronic illnesses. To measure C, we link predictions to insurance claims data on utilization, including outpatient and emergency visits, hospitalizations, and health care costs. These data, and the rationale for the specific measures of H used in this study, are described in more detail in the supplementary materials.

Health disparities conditional on risk score

We begin by calculating an overall measure of health status, the number of active chronic conditions [or "comorbidity score," a metric used extensively in medical research (*24*) to provide a comprehensive view of a patient's health (*25*)] by race, conditional on algorithmic risk score. Fig. 1A shows that, at the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites. We can quantify these differences by choosing one point on the x axis that corresponds to a very-high-risk group (e.g., patients at the 97th percentile of risk score, at which patients are auto-identified for program enrollment), where Blacks have 26.3% more chronic illnesses than Whites (4.8 versus 3.8 distinct conditions; $P < 0.001$).

What do these prediction differences mean for patients? Algorithm scores are a key input to decisions about future enrollment in a care coordination program. So as we might expect, with less-healthy Blacks scored at similar risk scores to more-healthy Whites, we find evidence

Table 1. Descriptive statistics on our sample, by race. BP, blood pressure; LDL, low-density lipoprotein.

	White	Black
n (patient-years)	88,080	11,929
n (patients)	43,539	6079
Demographics		
Age	51.3	48.6
Female (%)	62	69
Care management program		
Algorithm score (percentile)	50	52
Race composition of program (%)	81.8	18.2
Care utilization		
Actual cost	$7540	$8442
Hospitalizations	0.09	0.13
Hospital days	0.50	0.78
Emergency visits	0.19	0.35
Outpatient visits	4.94	4.31
Mean biomarker values		
HbA1c (%)	5.9	6.4
Systolic BP (mmHg)	126.6	130.3
Diastolic BP (mmHg)	75.5	75.7
Creatinine (mg/dl)	0.89	0.98
Hematocrit (%)	40.7	37.8
LDL (mg/dl)	103.4	103.0
Active chronic illnesses (comorbidities)		
Total number of active illnesses	1.20	1.90
Hypertension	0.29	0.44
Diabetes, uncomplicated	0.08	0.22
Arrythmia	0.09	0.08
Hypothyroid	0.09	0.05
Obesity	0.07	0.18
Pulmonary disease	0.07	0.11
Cancer	0.07	0.06
Depression	0.06	0.08
Anemia	0.05	0.10
Arthritis	0.04	0.04
Renal failure	0.03	0.07
Electrolyte disorder	0.03	0.05
Heart failure	0.03	0.05
Psychosis	0.03	0.05
Valvular disease	0.03	0.02
Stroke	0.02	0.03
Peripheral vascular disease	0.02	0.02
Diabetes, complicated	0.02	0.07
Heart attack	0.01	0.02
Liver disease	0.01	0.02

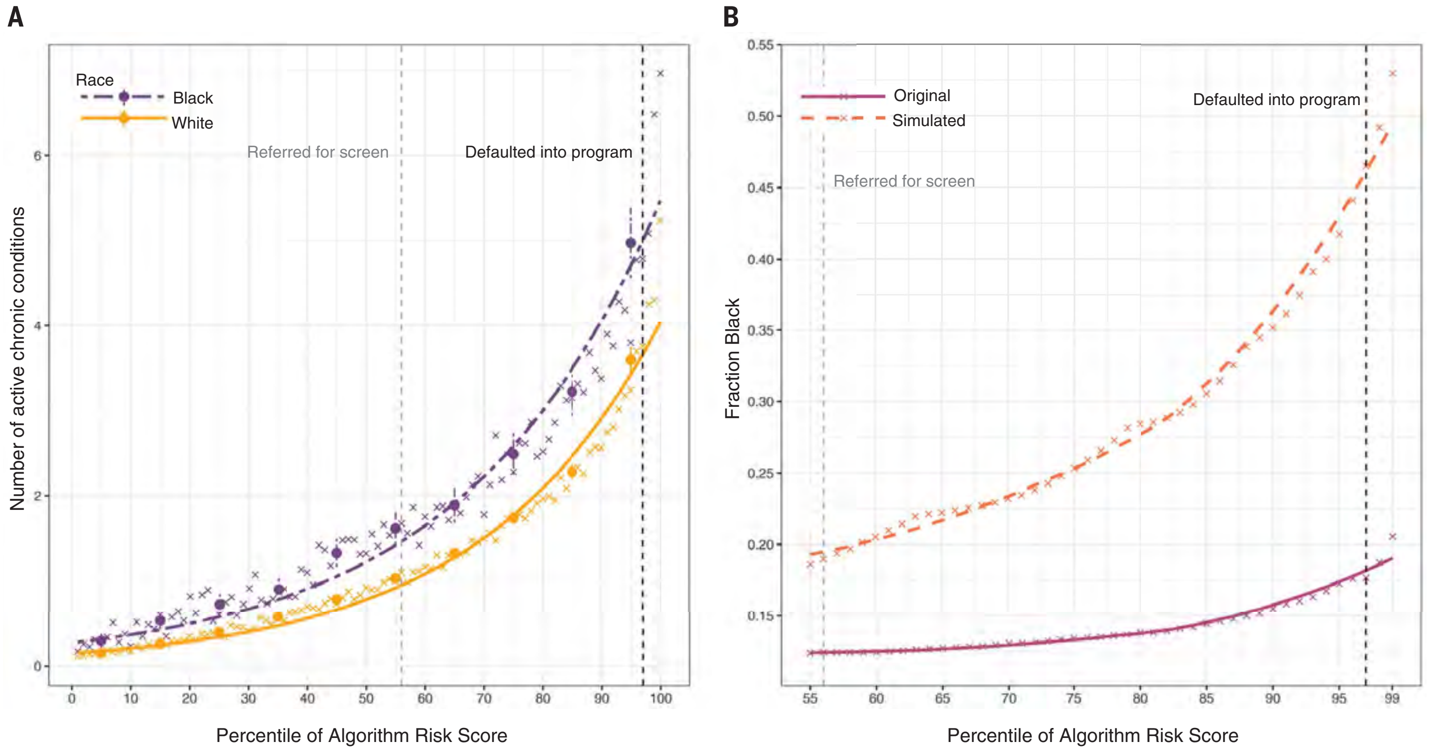


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

of substantial disparities in program screening. We quantify this by simulating a counterfactual world with no gap in health conditional on risk. Specifically, at some risk threshold α, we identify the supramarginal White patient (i) with $R_i > \alpha$ and compare this patient's health to that of the inframarginal Black patient (j) with $R_j < \alpha$. If $H_i > H_j$, as measured by number of chronic medical conditions, we replace the (healthier, but supramarginal) White patient with the (sicker, but inframarginal) Black patient. We repeat this procedure until $H_i = H_j$, to simulate an algorithm with no predictive gap between Blacks and Whites. Fig. 1B shows the results: At all risk thresholds α above the 50th percentile, this procedure would increase the fraction of Black patients. For example, at α = 97th percentile, among those auto-identified for the program, the fraction of Black patients would rise from 17.7 to 46.5%.

We then turn to a more multidimensional picture of the complexity and severity of patients' health status, as measured by biomarkers that index the severity of the most common chronic illnesses in our sample (as shown in Table 1). This allows us to identify patients who might derive a great deal of benefit from care management programs—e.g., patients with severe diabetes who are at risk of catastrophic complications if they do not lower their blood sugar (*18*, *26*). (The materials and methods section describes several experiments to rule out a large effect of the program on these health measures in year t; had there been such an effect, we could not easily use the measures to assess the accuracy of the algorithm's predictions on health, because the program is allocated as a function of algorithm score.) Across all of these important markers of health needs—severity of diabetes, high blood pressure, renal failure, cholesterol, and anemia—we find that Blacks are substantially less healthy than Whites at any level of algorithm predictions, as shown in Fig. 2. Blacks have more-severe hypertension, diabetes, renal failure, and anemia, and higher cholesterol. The magnitudes of these differences are large: For example, differences in severity of hypertension (systolic pressure: 5.7 mmHg) and diabetes [glycated hemoglobin (HbA1c): 0.6%] imply differences in all-cause mortality of 7.6% (*27*) and 30% (*28*), respectively, calculated using data from clinical trials and longitudinal studies.

Mechanism of bias

An unusual aspect of our dataset is that we observe the algorithm's inputs and outputs as well as its objective function, providing us a unique window into the mechanisms by which bias arises. In our setting, the algorithm takes in a large set of raw insurance claims data $X_{i,t-1}$ (features) over the year $t-1$: demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs. Notably, the algorithm specifically excludes race.

The algorithm uses these data to predict $Y_{i,t}$ (i.e., the label). In this instance, the algorithm takes total medical expenditures (for simplicity, we denote "costs" C_t) in year t as the label. Thus, the algorithm's prediction on health needs is, in fact, a prediction on health costs.

As a first check on this potential mechanism of bias, we calculate the distribution of realized costs C versus predicted costs R. By this metric, one could call the algorithm unbiased. Fig. 3A shows that, at every level of algorithm-predicted risk, Blacks and Whites have (roughly) the same costs the following year. In other words, the algorithm's predictions are well calibrated across races. For example, at the median risk score, Black patients had costs of \$5147 versus \$4995 for Whites (U.S. dollars); in the top 5% of algorithm-predicted risk, costs were \$35,541 for Blacks versus \$34,059 for Whites.

Because these programs are used to target patients with high costs, these results are largely inconsistent with algorithmic bias, as measured by calibration: Conditional on risk score, predictions do not favor Whites or Blacks anywhere in the risk distribution.

To summarize, we find substantial disparities in health conditional on risk but little disparity in costs. On the one hand, this is surprising: Health care costs and health needs are highly correlated, as sicker patients need and receive more care, on average. On the other hand, there are many opportunities for a wedge to creep in between needing health care and receiving health care—and crucially, we find that wedge to be correlated with race, as shown in Fig. 3B. At a given level of health (again measured by number of chronic illnesses), Blacks generate lower costs than Whites—on average, $1801 less per year, holding constant the number of chronic illnesses (or $1144 less, if we instead hold constant the specific individual illnesses that contribute to the sum). Table S2 also shows that Black patients generate very different kinds of costs: for example, fewer inpatient surgical and outpatient specialist costs, and more costs related to emergency visits and dialysis. These results suggest that the driving force behind the bias we detect is that Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities. As a result, accurate prediction of costs necessarily means being racially biased on health.

How might these disparities in cost arise? The literature broadly suggests two main potential channels. First, poor patients face substantial barriers to accessing health care, even when enrolled in insurance plans. Although the population we study is entirely insured, there are many other mechanisms by which poverty can lead to disparities in use of health care: geography and differential access to transportation, competing demands from jobs or child care, or knowledge of reasons to seek care (*29*–*31*). To the extent that race and socioeconomic status are correlated, these factors will differentially affect Black patients. Second, race could affect costs directly via several channels: direct (“taste-based”) discrimination, changes to the doctor–patient relationship, or others. A recent trial randomly assigned Black patients to a Black or White primary care provider and found significantly higher uptake of recommended preventive care when the provider was Black (*32*). This is perhaps the most rigorous demonstration of this effect, and it fits with a larger literature on potential mechanisms by which race can affect health care directly. For example, it has long been documented that Black patients have reduced trust in the health care system (*33*), a fact that some studies trace to the revelations of the Tuskegee study and other adverse experiences (*34*). A substantial literature in psychology has documented physicians’ differential perceptions of Black patients, in terms of intelligence, affiliation (*35*), or pain tolerance (*36*). Thus, whether it is communication, trust, or bias, something about the interactions of Black patients with the health care system itself leads to reduced use of health care. The collective effect of these many channels is to lower health spending substantially for Black patients, conditional on need—a finding that has been appreciated for at least two decades (*37*).

Problem formulation

Our findings highlight the importance of the choice of the label on which the algorithm is trained. On the one hand, the algorithm manufacturer’s choice to predict future costs is reasonable: The program’s goal, at least in part, is

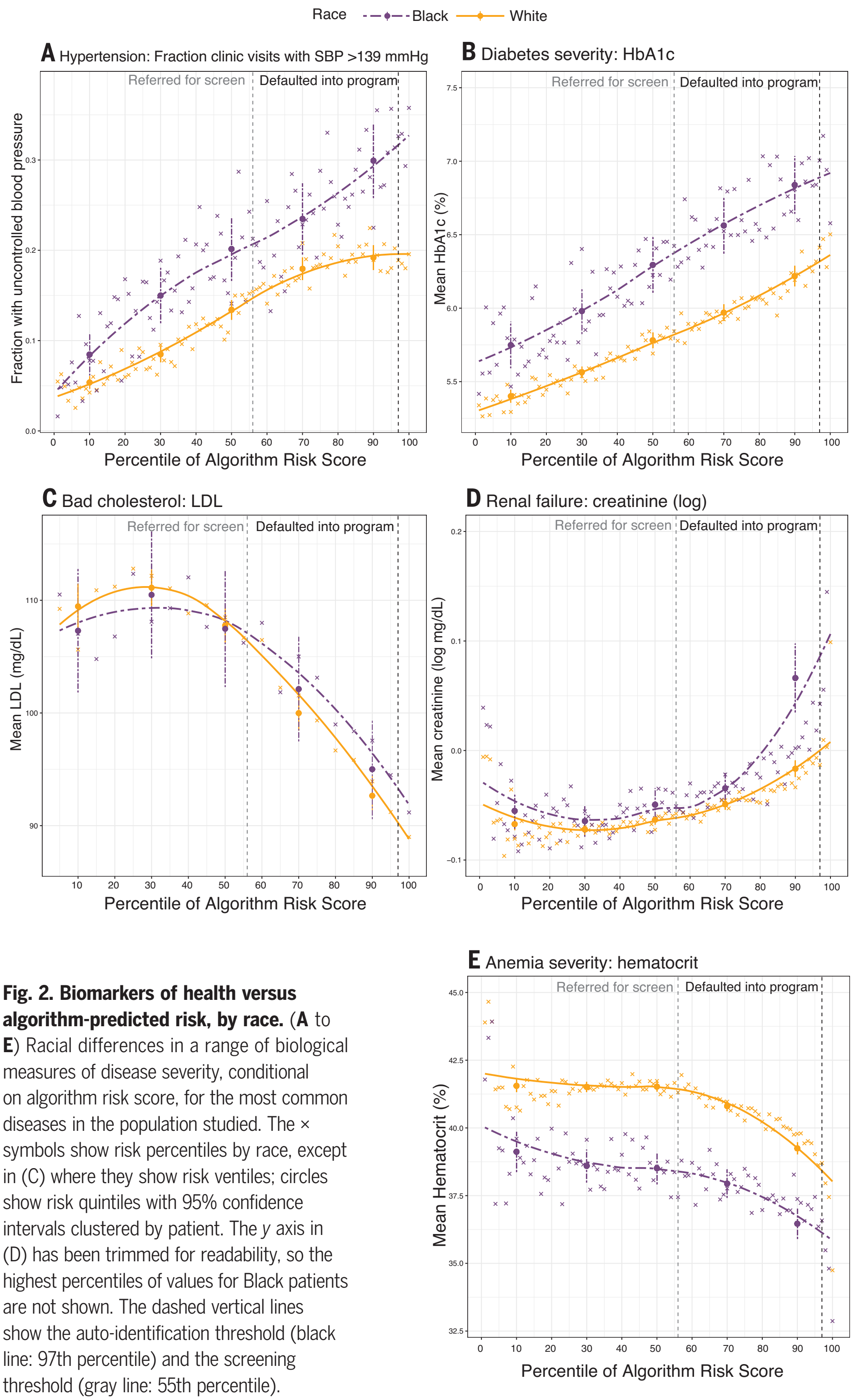


Fig. 2. Biomarkers of health versus algorithm-predicted risk, by race. (**A** to **E**) Racial differences in a range of biological measures of disease severity, conditional on algorithm risk score, for the most common diseases in the population studied. The × symbols show risk percentiles by race, except in (C) where they show risk ventiles; circles show risk quintiles with 95% confidence intervals clustered by patient. The *y* axis in (D) has been trimmed for readability, so the highest percentiles of values for Black patients are not shown. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile).

to reduce costs, and it stands to reason that patients with the greatest future costs could have the greatest benefit from the program. As noted in the supplementary materials, the manufacturer is not alone. Although the details of individual algorithms vary, the cost label reflects the industry-wide approach. For example, the Society of Actuaries's comprehensive evaluation of the 10 most widely used algorithms, including the particular algorithm we study, used cost prediction as its accuracy metric (*21*). As noted in the report, the enthusiasm for cost prediction is not restricted to industry: Similar algorithms are developed and used by non-profit hospitals, academic groups, and governmental agencies, and are often described in academic literature on targeting population health interventions (*18*, *19*).

On the other hand, future cost is by no means the only reasonable choice. For example, the evidence on care management programs shows that they do not operate to reduce costs globally. Rather, these programs primarily work to prevent acute health decompensations that lead to catastrophic health care utilization (indeed, they actually work to increase other categories of costs, such as primary care and home health assistance; see table S2). Thus avoidable future costs, i.e., those related to emergency visits and hospitalizations, could be a useful label to predict. Alternatively, rather than predicting costs at all, we could simply predict a measure of health; e.g., the number of active chronic health conditions. Because the program ultimately operates to improve the management of these conditions, patients with the most encounters related to them could also be a promising group on which to deploy preventative interventions.

The dilemma of which label to choose relates to a growing literature on "problem formulation" in data science: the task of turning an often amorphous concept we wish to predict into a concrete variable that can be predicted in a given dataset (*38*). Problems in health seem particularly challenging: Health is, by nature, holistic and multidimensional, and there is no single, precise way to measure it. Health care costs, though well measured and readily available in insurance claims data, are also the result of a complex aggregation process with a number of distortions due to structural inequality, incentives, and inefficiency. So although the choice of label is perhaps the single most important decision made in the development of a prediction algorithm, in our setting and in many others, there is often a confusingly large array of different options, each with its own profile of costs and benefits.

Experiments on label choice

Through a series of experiments with our dataset, we can gain some insight into how label choice affects both predictive performance and racial bias. We develop three new predictive algorithms, all trained in the same way, to predict the following outcomes: total cost in year *t* (this tailors cost predictions to our own dataset rather than the national training set), avoidable cost in year *t* (due to emergency visits and hospitalizations), and health in year *t* (measured by the number of chronic conditions that flare up in that year). We train all models in a random ⅔ training set and show all results only from the ⅓ holdout set. Furthermore, as with the original algorithm, we exclude race from the feature set (more details are in the materials and methods).

Table 2 shows the results of these experiments. The first finding is that all algorithms perform reasonably well for predicting not only the outcome on which they were trained but also the other outcomes: The concentration of realized outcomes in those at or above the 97th percentile is notably similar for all algorithms across all outcomes. The largest difference in performance across algorithms is seen for cost prediction: Of all costs in the holdout set, the fraction generated by those at or above the 97th percentile is 16.5% for the cost predictor versus 12.1% for the predictor

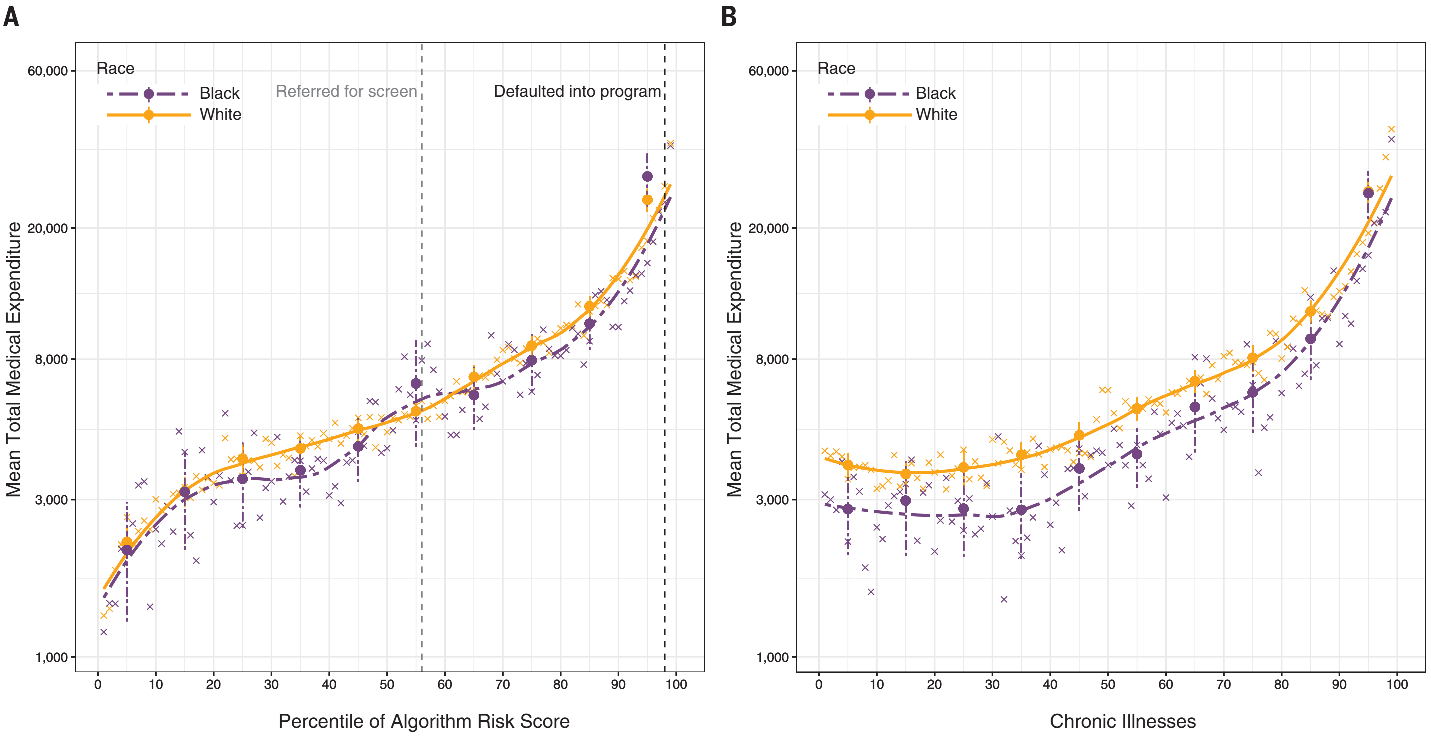


Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. (**A**) Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). (**B**) Total medical expenditures by race, conditional on number of chronic conditions. The × symbols show risk percentiles; circles show risk deciles with 95% confidence intervals clustered by patient. The *y* axis uses a log scale.

of chronic conditions. We then test for label choice bias, defined analogously to calibration bias above: For two algorithms trained to predict Y and Y', and using a threshold τ indexing a (similarly sized) high-risk group, we would test $p[B|R > \tau] = p[B|R' > \tau]$ (here, p denotes probability and B represents Black patients).

We find that the racial composition of this highest-risk group varies far more across algorithms: The fraction of Black patients at or above these risk levels ranges from 14.1% for the cost predictor to 26.7% for the predictor of chronic conditions. Thus, although there could be many reasonable choices of label—all predictions are highly correlated, and any could be justified as a measure of patients' likely benefit from the program—they have markedly different implications in terms of bias, with nearly twofold variation in composition of Black patients in the highest-risk groups.

Relation to human judgment

As noted above, the algorithm is not used for program enrollment decisions in isolation. Rather, it is used as a screening tool, in part to alert primary care doctors to high-risk patients. Specifically, for patients at or above a certain level of predicted risk (the 55th percentile), doctors are presented with contextual information from patients' electronic health records and insurance claims and are prompted to consider enrolling them in the program. Thus, realized enrollment decisions largely reflect how doctors respond to algorithmic predictions, along with other administrative factors related to eligibility (for instance, primary care practice site, residence outside of a nursing home, and continual enrollment in an insurance plan).

Table 3 shows statistics on those enrolled in the program, accounting for 1.3% of observations in our sample: The enrolled individuals are 19.2% Black (versus 11.9% Black in our entire sample) and account for 2.9% of all costs and 3.3% of all active chronic conditions in the population as a whole. We then perform four counterfactual simulations to put these numbers in context; naturally, these simulations use only observable factors, not the many unobserved administrative and human factors that also affect enrollment. First, we calculate the realized program enrollment rate within each percentile of the original algorithm's predicted risk bins and randomly sample patients in each bin for enrollment. This simulation, which mimics "race-blind" enrollment conditional on algorithm score, would yield an enrolled population that is 18.3% Black (versus 19.2% observed; P = 0.8348). Second, rather than randomly sampling, we sample those with the highest predicted number of active chronic conditions within a risk bin (using our experimental algorithm described above); this would yield a population that is 26.9% Black. Finally, we compare this to simply assigning those with the highest predicted costs, or the highest number of active chronic conditions, to the program (also using our own algorithms detailed above), which would yield 17.2 and 29.2% Black patients, respectively. Thus, although doctors do redress a small part of the algorithm's bias, they do so far less than an algorithm trained on a different label.

Discussion

Bias attributable to label choice—the difference between some unobserved optimal prediction and the prediction of an algorithm trained on an observed label—is a useful framework through which to understand bias in algorithms, both

Table 2. Performance of predictors trained on alternative labels. For each new algorithm, we show the label on which it was trained (rows) and the concentration of a given outcome of interest (columns) at or above the 97th percentile of predicted risk. We also show the fraction of Black patients in each group.

Algorithm training label	Concentration in highest-risk patients (SE)						Fraction of Black patients in group with highest risk (SE)	
	Total costs		Avoidable costs		Active chronic conditions			
Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141	(0.003)
Avoidable costs	0.142	(0.003)	0.215	(0.003)	0.130	(0.003)	0.210	(0.003)
Active chronic conditions	0.121	(0.003)	0.182	(0.003)	0.148	(0.003)	0.267	(0.003)
Best-to-worst difference	0.044		0.033		0.043		0.126	

Table 3. Doctors' decisions versus algorithmic predictions. For those enrolled in the high-risk care management program (1.3% of our sample), we first show the fraction of the population that is Black, as well as the fraction of all costs and chronic conditions accounted for by these observations. We also show these quantities for four alternative program enrollment rules, which we simulate in our dataset (using the holdout set when we use our experimental predictors). We first calculate the program enrollment rate within each percentile bin of predicted risk from the original algorithm and either (i) randomly sample patients or (ii) sample those with the highest predicted number of active chronic conditions within a bin and assign them to the program. The resultant values are then compared with values obtained by simply assigning the aforementioned 1.3% of our sample with (iii) the highest predicted cost or (iv) the highest number of active chronic conditions to the program.

Population	Fraction Black (SE)		Fraction of all costs (SE)		Fraction of all active chronic conditions (SE)	
Observed program enrollment (1.3%)	0.192	(0.003)	0.029	(0.001)	0.033	(0.001)
Simulated alternative enrollment rules						
Random, in predicted-cost bin	0.183	(0.003)	0.044	(0.002)	0.034	(0.001)
Predicted health, in predicted-cost bin	0.269	(0.003)	0.044	(0.002)	0.064	(0.002)
Highest predicted cost	0.172	(0.003)	0.100	(0.002)	0.047	(0.002)
Worst predicted health	0.292	(0.004)	0.067	(0.002)	0.076	(0.002)

in the health sector and further afield. This is because labels are often measured with errors that reflect structural inequalities (*39*). Within the health sector, using mortality or readmission rates to measure hospital performance penalizes those serving poor or non-White populations (*40*, *41*). Outside of the health arena, credit-scoring algorithms predict outcomes related to income, thus incorporating disparities in employment and salary (*2*). Policing algorithms predict measured crime, which also reflects increased scrutiny of some groups (*42*). Hiring algorithms predict employment decisions or supervisory ratings, which are affected by race and gender biases (*43*). Even retail algorithms, which set pricing for goods at the national level, penalize poorer households, which are subjected to increased prices as a result (*44*).

This mechanism of bias is particularly pernicious because it can arise from reasonable choices: Using traditional metrics of overall prediction quality, cost seemed to be an effective proxy for health yet still produced large biases. After completing the analyses described above, we contacted the algorithm manufacturer for an initial discussion of our results. In response, the manufacturer independently replicated our analyses on its national dataset of 3,695,943 commercially insured patients. This effort confirmed our results—by one measure of predictive bias calculated in their dataset, Black patients had 48,772 more active chronic conditions than White patients, conditional on risk score—illustrating how biases can indeed arise inadvertently.

To resolve the issue, we began to experiment with solutions together. As a first step, we suggested using the existing model infrastructure—sample, predictors (excluding race, as before), training process, and so forth—but changing the label: Rather than future cost, we created an index variable that combined health prediction with cost prediction. This approach reduced the number of excess active chronic conditions in Blacks, conditional on risk score, to 7758, an 84% reduction in bias. Building on these results, we are establishing an ongoing (unpaid) collaboration to convert the results of Table 3 into a better, scaled predictor of multidimensional health measures, with the goal of rolling these improvements out in a future round of algorithm development. Of course, our experience may not be typical of all algorithm developers in this sector. But because the manufacturer of the algorithm we study is widely viewed as an industry leader in data and analytics, we are hopeful that this endeavor will prompt other manufacturers to implement similar fixes.

These results suggest that label biases are fixable. Changing the procedures by which we fit algorithms (for instance, by using a new statistical technique for decorrelating predictors with race or other similar solutions) is not required. Rather, we must change the data we feed the algorithm—specifically, the labels we give it. Producing new labels requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment. But there is precedent for all of these functions in the literature and, more concretely, in the private companies that invest heavily in developing new and improved labels to predict factors such as consumer behavior (*45*). In addition, although health—as well as criminal justice, employment, and other socially important areas—presents substantial challenges to measurement, the importance of these sectors emphasizes the value of investing in such research. Because labels are the key determinant of both predictive quality and predictive bias, careful choice can allow us to enjoy the benefits of algorithmic predictions while minimizing their risks.

REFERENCES AND NOTES

1. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine Bias," *ProPublica* (23 May 2016); www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
2. S. Barocas, A. D. Selbst, *Calif. Law Rev.* **104**, 671 (2016).
3. A. Chouldechova, A. Roth, arXiv:1810.08810 [cs.LG] (20 October 2018).
4. A. Datta, M. C. Tschantz, A. Datta, *Proc. Privacy Enhancing Technol.* **2015**, 92–112 (2015).
5. L. Sweeney, *Queue* **11**, 1–19 (2013).
6. M. Kay, C. Matuszek, S. A. Munson, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM, 2015), pp. 3819–3828.
7. B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, A. K. Jain, *IEEE Trans. Inf. Forensics Security* **7**, 1789–1801 (2012).
8. J. Buolamwini, T. Gebru, in *Proceedings of the Conference on Fairness, Accountability and Transparency* (PMLR, 2018), pp. 77–91.
9. A. Caliskan, J. J. Bryson, A. Narayanan, *Science* **356**, 183–186 (2017).
10. S. Corbett-Davies, S. Goel, arXiv:1808.00023 [cs.CY] (31 July 2018).
11. M. De-Arteaga *et al*., arXiv:1901.09451 [cs.IR] (27 January 2019).
12. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 259–268.
13. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, *Q. J. Econ.* **133**, 237–293 (2018).
14. C. S. Hong, A. L. Siegel, T. G. Ferris, *Issue Brief (Commonwealth Fund)* **19**, 1–19 (2014).
15. N. McCall, J. Cromwell, C. Urato, "Evaluation of Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Massachusetts General Hospital and Massachusetts General Physicians Organization (MGH)" (RTI International, 2010).
16. J. Hsu *et al*., *Health Aff.* **36**, 876–884 (2017).
17. L. Nelson, "Lessons from Medicare's demonstration projects on disease management and care coordination" (Working Paper 2012-01, Congressional Budget Office, 2012).
18. C. Vogeli *et al*., *J. Gen. Intern. Med.* **22** (suppl. 3), 391–395 (2007).
19. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, *Health Aff.* **33**, 1123–1131 (2014).
20. J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **105**, 491–495 (2015).
21. G. Hileman, S. Steele, "Accuracy of claims-based risk scoring models" (Society of Actuaries, 2016).
22. J. Kleinberg, S. Mullainathan, M. Raghavan, arXiv:1609.05807 [cs.LG] (19 September 2016).
23. A. Chouldechova, *Big Data* **5**, 153–163 (2017).
24. V. de Groot, H. Beckerman, G. J. Lankhorst, L. M. Bouter, *J. Clin. Epidemiol.* **56**, 221–229 (2003).
25. J. J. Gagne, R. J. Glynn, J. Avorn, R. Levin, S. Schneeweiss, *J. Clin. Epidemiol.* **64**, 749–759 (2011).
26. A. K. Parekh, M. B. Barton, *JAMA* **303**, 1303–1304 (2010).
27. D. Ettehad *et al*., *Lancet* **387**, 957–967 (2016).
28. K.-T. Khaw *et al*., *BMJ* **322**, 15 (2001).
29. K. Fiscella, P. Franks, M. R. Gold, C. M. Clancy, *JAMA* **283**, 2579–2584 (2000).
30. N. E. Adler, K. Newman, *Health Aff.* **21**, 60–76 (2002).
31. N. E. Adler, W. T. Boyce, M. A. Chesney, S. Folkman, S. L. Syme, *JAMA* **269**, 3140–3145 (1993).
32. M. Alsan, O. Garrick, G. C. Graziani, "Does diversity matter for health? Experimental evidence from Oakland" (National Bureau of Economic Research, 2018).
33. K. Armstrong, K. L. Ravenell, S. McMurphy, M. Putt, *Am. J. Public Health* **97**, 1283–1289 (2007).
34. M. Alsan, M. Wanamaker, *Q. J. Econ.* **133**, 407–455 (2018).
35. M. van Ryn, J. Burke, *Soc. Sci. Med.* **50**, 813–828 (2000).
36. K. M. Hoffman, S. Trawalter, J. R. Axt, M. N. Oliver, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4296–4301 (2016).
37. J. J. Escarce, F. W. Puffer, in *Racial and Ethnic Differences in the Health of Older Americans* (National Academies Press, 1997), chap. 6; www.ncbi.nlm.nih.gov/books/NBK109841/.
38. S. Passi, S. Barocas, arXiv:1901.02547 [cs.CY] (8 January 2019).
39. S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **107**, 476–480 (2017).
40. K. E. Joynt Maddox *et al*., *Health Serv. Res.* **54**, 327–336 (2019).
41. K. E. Joynt Maddox, M. Reidhead, A. C. Qi, D. R. Nerenz, *JAMA Intern. Med.* **179**, 769–776 (2019).
42. K. Lum, W. Isaac, *Significance* **13**, 14–19 (2016).
43. I. Ajunwa, "The Paradox of Automation as Anti-Bias Intervention," available at SSRN (2016); https://ssrn.com/abstract=2746078.
44. S. DellaVigna, M. Gentzkow, "Uniform pricing in US retail chains" (National Bureau of Economic Research, 2017).
45. C. A. Gomez-Uribe, N. Hunt, *ACM Trans. Manag. Inf. Syst.* **6**, 13 (2016).

ACKNOWLEDGMENTS

We thank S. Lakhtakia, Z. Li, K. Lin, and R. Mahadeshwar for research assistance and D. Buefort and E. Maher for data science expertise. **Funding:** This work was supported by a grant from the National Institute for Health Care Management Foundation. **Author contributions:** Z.O. and S.M. designed the study, obtained funding, and conducted the analyses. All authors contributed to reviewing findings and writing the manuscript. **Competing interests:** The analysis was completely independent: None of the authors had any contact with the algorithm's manufacturer until after it was complete. No authors received compensation, in any form, from the manufacturer or have any commercial interests in the manufacturer or competing entities or products. There were no confidentiality agreements that limited reporting of the work or its results, no material transfer agreements, no oversight in the preparation of this article (besides ethical oversight from the approving IRB, which was based at a non-profit academic health system), and no formal relationship of any kind between any of the authors and the manufacturer. **Data and materials availability:** Because the data used in this analysis are protected health information, they cannot be made publicly available. We provide instead a synthetic dataset (using the R package synthpop) and all code necessary to reproduce our analyses at https://gitlab.com/labsysmed/dissecting-bias.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/366/6464/447/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S4
References (*46*–*51*)

8 March 2019; accepted 4 October 2019
10.1126/science.aax2342

Causal Reasoning for Algorithmic Fairness

Joshua R. Loftus[1], Chris Russell[2,5], Matt J. Kusner[3,5], and Ricardo Silva[4,5]
[1]New York University [2]University of Surrey [3]University of Warwick
[4]University College London [5]Alan Turing Institute

Abstract

In this work, we argue for the importance of causal reasoning in creating fair algorithms for decision making. We give a review of existing approaches to fairness, describe work in causality necessary for the understanding of causal approaches, argue why causality is necessary for any approach that wishes to be fair, and give a detailed analysis of the many recent approaches to causality-based fairness.

1 Introduction

The success of machine learning algorithms has created a wave of excitement about the problems they could be used to solve. Already we have algorithms that match or outperform humans in non-trivial tasks such as image classification [18], the game of Go [37], and skin cancer classification [15]. This has spurred the use of machine learning algorithms in predictive policing [25], in loan lending [17], and to predict whether released people from jail will re-offend [9]. In these life-changing settings however, it has quickly become clear that machine learning algorithms can unwittingly perpetuate or create discriminatory decisions that are biased against certain individuals (for example, against a particular race, gender, sexual orientation, or other protected attributes). Specifically, such biases have already been demonstrated in natural language processing systems [5] (where algorithms associate men with technical occupations like ‘computer programmer’ and women with domestic occupations like ‘homemaker’), and in online advertising [41] (where Google showed advertisements suggesting that a person had been arrested when that person had a name more often associated with black individuals).

As machine learning is deployed in an increasingly wide range of human scenarios, it is more important than ever to understand what biases are present in a decision making system, and what constraints we can put in place to guarantee that a learnt system never exhibits such biases. Research into these problems is referred to as *algorithmic fairness*. It is a particularly challenging area of research for two reasons: many different features are intrinsically linked to protected classes such as race or gender. For example, in many scenarios, knowledge of someone's address makes it easy to predict their race with relatively high accuracy; while their choice of vocabulary might reveal much about their upbringing or gender. As such it is to easy to accidentally create algorithms that make decisions without knowledge of a persons race or gender, but still exhibit a racial or gender bias. The second issue is more challenging still, there is fundamental disagreement in the field as to what *algorithmic fairness* really means. Should algorithms be fair if they always make similar decisions for similar individuals? Should we instead call algorithms that make beneficial decisions for all genders at roughly the same rate fair? Or should we use a third different criteria? This question is of fundamental importance as many of these different criteria can not be satisfied at the same time [22].

In this work we argue that it is important to understand where these sources of bias come from in order to rectify them, and that causal reasoning is a powerful tool for doing this. We review existing notions of fairness in prediction problems; the tools of causal reasoning; and show how these can be combined together using techniques such as counterfactual fairness [23].

2 Current Work in Algorithmic Fairness

To discuss the existing measures of fairness, we use capital letters to refer to variables and lower case letters to refer to a value a variable takes. For example, we will always use A for a protected attribute such as gender, and a or a' to refer to the different values the attribute can take such as *man* or *woman.* We use Y to refer to the true state of a variable we wish to predict, for example the variable might denote whether a person defaults on a loan or if they will violate parole conditions. We will use $\hat{Y}$ to denote our prediction of the true variable Y. The majority of definitions of fairness in prediction problems are statements about probability of a particular prediction occurring given that some prior conditions hold. In what follows, we will use $P(\cdot \mid \cdot)$ to represent either conditional probability of events, probability mass functions or density functions, as required by the context.

2.0.1 Equalised Odds

Two definitions of fairness that have received much attention are equalised odds and calibration. Both were heavily used in the ProPublica investigation into Northpointe's COMPAS score, designed to gauge the propensity of a prisoner to re-offend upon release [16]. The first measure is equalised odds, which says that if a person truly has state y, the classifier will predict this at the same rate regardless of the value of their protected attribute. This can be written as an equation in the following form:

$$P(\hat{Y} = y \mid A = a, Y = y) = P(\hat{Y} = y \mid A = a', Y = y) \tag{1}$$

for all y, a, a'. Another way of stating this property is by saying that $\hat{Y}$ is independent of A given Y, which we will denote by $\hat{Y} \perp\!\!\!\perp A \mid Y$.

2.0.2 Calibration

The second condition is referred to as calibration (or 'test fairness' in [9]). This reverses the previous condition of equalised odds, and says that if the classifier predicts that a person has state y, their probability of actually having state y should be the same for all choices of attribute.

$$P(Y = y \mid A = a, \hat{Y} = y) = P(Y = y \mid A = a', \hat{Y} = y) \tag{2}$$

for all choices of y, a, and a', that is, $Y \perp\!\!\!\perp A \mid \hat{Y}$.

Although the two measures sound very similar, they are fundamentally incompatible. These two measures achieved some notoriety when Propublica showed that Northpointe's COMPAS score violated equalised odds, accusing them of racial discrimination. In response, Northpointe claimed that their COMPAS score satisfied calibration and that they did not discriminate. Kleinberg et al. [22] and Chouldechova [8] showed that both conditions cannot be satisfied at the same time except in special cases such as zero prediction error or if $Y \perp\!\!\!\perp A$.

The use of calibration and equalised odds has another major limitation. If $Y \not\perp\!\!\!\perp A$, the true scores Y typically have some inherent bias. This happens, for example, if the police are more likely to unfairly decide that minorities are violating their parole. The definitions of calibration or equalised odds do not explicitly forbid the classifier from preserving an existing bias.

2.0.3 Demographic Parity/Disparate Impact

Perhaps the most common non-causal notion of fairness is *demographic parity*, defined as follows:

$$P(\hat{Y} = y \mid A = a) = P(\hat{Y} = y \mid A = a'), \tag{3}$$

for all y, a, a', that is, $\hat{Y} \perp\!\!\!\perp A$. If unsatisfied, this notion is also referred to as *disparate impact*. Demographic parity has been used, for several purposes, in the following works: [14, 19, 20, 24, 42, 43].

Satisfying demographic parity can often require positive discrimination, where certain individuals who are otherwise very similar are treated differently due to having different protected attributes. Such *disparate treatment* can violate other intuitive notions of fairness or equality, contradict equalised odds or calibration, and in some cases is prohibited by law.

2.0.4 Individual Fairness

Dwork et al. [12] proposed the concept of *individual fairness* as follows.

$$P(\hat{Y}^{(i)} = y \mid X^{(i)}, A^{(i)}) \approx P(\hat{Y}^{(j)} = y \mid X^{(j)}, A^{(j)}), \text{ if } d(i,j) \approx 0, \tag{4}$$

where i, j refer to two different individuals and the superscripts $(i), (j)$ are their associated data. The function $d(\cdot, \cdot)$ is a 'task-specific' metric that describes how any pair of individuals should be treated similarly in a fair world. The work suggests that this metric could be defined by ‘a regulatory body, or . . . a civil rights organization’. While this notion mitigates the issues with individual predictions that arose from demographic parity, it replaces the problem of defining fairness with defining a fair metric $d(\cdot, \cdot)$. As we observed in the introduction, many variables vary along with protected attributes such as race or gender, making it challenging to find a distance measure that will not allow some implicit discrimination.

2.0.5 Causal Notions of Fairness

A number of recent works use causal approaches to address fairness [1, 7, 21, 23, 35, 44], which we review in more detail in Section 5. We describe selected background on causal reasoning in Section 3. These works depart from the previous approaches in that they are not wholly data-driven but require additional knowledge of the structure of the world, in the form of a causal model. This additional knowledge is particularly valuable as it informs us how changes in variables propagate in a system, be it natural, engineered or social. Explicit causal assumptions remove ambiguity from methods that just depend upon statistical correlations. For instance, causal methods provide a recipe to express assumptions on how to recover from sampling biases in the data (Section 4) or how to describe mixed scenarios where we may believe that certain forms of discrimination should be allowed while others should not (e.g., how gender influences one’s field of study in college, as in Section 5).

3 Causal Models

We now review causality in sufficient detail for our analysis of causal fairness in Section 5. It is challenging to give a self-contained definition of causality, as many working definitions reveal circularities on close inspection. For two random variables X and Y, informally we say that X *causes* Y when there exist at least two different *interventions* on X that result in two different probability distributions of Y. This does not mean we will be able to define what an “intervention” is without using causal concepts, hence circularities appear.

Nevertheless, it is possible to formally express causal assumptions and to compute the consequences of such assumptions if one is willing to treat some concepts, such as interventions, as primitives. This is just an instance of the traditional axiomatic framework of mathematical modelling, dating back to Euclid. In particular, in this paper we will make use primarily of the *structural causal model* (SCM) framework advocated by [29], which shares much in common with the approaches by [33] and [39].

3.1 Structural Causal Models

We define a causal model as a triplet (U, V, F) of sets such that:

is set to intervention levels a and a'. A joint distribution for $Y(a)$ and $Y(a')$ is implied by the model. Conditional distributions, such as $P(Y(a) = y_a, Y(a') = y_{a'} \mid A = a, Y = y, Z = z)$ are also defined. Figure 1(c) shows the case for interventions on Y. It is not difficult to show, as Y is not an ancestor of A in the graph, that $A(y, u) = A(y', u) = A(u)$ for all u, y, y'. This captures the notion the Y does not cause A.

3.3 Counterfactuals Require Untestable Assumptions

Unless structural equations depend on observed variables only, they cannot be tested for correctness (unless other untestable assumptions are imposed). We can illustrate this problem by noting that a conditional density function $P(V_j \mid V_i = v)$ can be written as an equation $V_j = f_1(v, U) \equiv F^{-1}_{V_i=v}(U) = F^{-1}_{V_i=v}(g^{-1}(g(U))) \equiv f_2(v, U')$, where $F^{-1}_{V_i=V}(\cdot)$ is the inverse cumulative distribution function corresponding to $P(V_j \mid V_i = v)$, U is an uniformly distributed random variable on $[0, 1]$, $g(\cdot)$ is some arbitrary invertible function on $[0, 1]$, and $U' \equiv g(U)$. While this is not fundamental for effects of causes, which depend solely on predictive distributions that at least in theory can be estimated from RCTs, different structural equations with the same interventional distributions will imply different joint distributions over the counterfactuals.

The traditional approach for causal inference in statistics tries to avoid any estimand that cannot be expressed by the marginal distributions of the counterfactuals (i.e., all estimands in which marginals $P(Y(a) = y_a)$ and $P(Y(a') = y_{a'})$ would provide enough information, such as the *average causal effect* $\mathsf{E}[Y(a) - Y(a')] = \mathsf{E}[Y \mid do(A = a)] - \mathsf{E}[Y \mid do(A = a')]$). Models that follow this approach and specify solely the univariate marginals of a counterfactual joint distribution are sometimes called *single-world* models [32]. However, as we will see, *cross-world* models seem a natural fit to algorithmic fairness. In particular, they are required for non-trivial statements that concern fairness at an individual level as opposed to fairness measures averaged over groups of individuals.

4 Why Causality is Critical For Fairness

Ethicists and social choice theorists recognise the importance of causality in defining and reasoning about fairness. Terminology varies, but many of their central questions and ideas, such as the role of agency in justice, responsibility-sensitive egalitarianism, and luck egalitarianism [10, 13, 31] involve causal reasoning. Intuitively, it is unfair for individuals to experience different outcomes caused by factors outside of their control. Empirical studies of attitudes about distributive justice [6, 26] have found that most participants prefer redistribution to create fairer outcomes, and do so in ways that depend on how much control individuals have on their outcomes. Hence, when choosing policies and designing systems that will impact people, we should minimise or eliminate the causal dependence on factors outside an individual's control, such as their perceived race or where they were born. Since such factors have influences on other aspects of peoples' lives that may also be considered relevant for determining what is fair, applying this intuitive notion of fairness requires careful causal modelling as we describe here.

Is it necessary that models attempting to remove such factors be causal? Many other notions of algorithmic fairness have also attempted to control or adjust for covariates. While it is possible to produce identical predictions or decisions with a model that is equivalent mathematically but without overt causal assumptions or interpretations, the design decisions underlying a covariate adjustment are often based on implicit causal reasoning. There is a fundamental benefit from an explicit statement of these assumptions. To illustrate this, we consider a classic example of bias in graduate admissions.

4.1 Revisiting Gender Bias In Berkeley Admissions

The Berkeley admissions example [3] is often used to explain Simpson's paradox [38] and highlight the importance of adjusting for covariates. In the fall of 1973, about 34.6% of women and 44.3% of men who applied to graduate studies at Berkeley were admitted. However, this was not evidence that the admissions decisions were biased against women. Decisions were made on a departmental basis, and each department admitted proportions of men and women at approximately the same rate. However, a greater proportion of women applied to the most selective departments, resulting in a lower overall acceptance rate for women.

While the overall outcome is seemingly unfair, after controlling for choice of department it appears to be fair, at least in some sense. In fact, while the presentation of this example to illustrate Simpson's paradox often ends there, the authors in [3] conclude, "Women are shunted by their socialisation and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects." The outcome can still be judged to be unfair, not due to biased admissions decisions, but rather to the causes of differences in choice of department, such as socialisation. Achieving or defining fairness requires addressing those root causes and applying value judgements. Applicants certainly have some agency over which department they apply to, but that decision is not made free of outside influences. They had no control over what kind of society they had been born into, what sort of gender norms that society had during their lifetime, or the scarcity of professional role models, and so on.

The quote above suggests that the authors in [3] were reasoning about causes even if they did not make explicit use of causal modelling. Indeed, conditioning on the choice of the department only makes sense because we understand it has a causal relationship with the outcome of interest and is not just a spurious correlation. Pearl [29] provides a detailed account of the causal basis of Simpson's paradox.

4.2 Selection Bias and Causality

Unfairness can also arise from bias in how data is collected or sampled. For instance, if the police stop individuals on the street to check for the possession of illegal drugs, and if the stopping protocol is the result of discrimination that targets individuals of a particular race, this can create a feedback loop that justifies discriminatory practice. Namely, if data gathered by the police suggests that $P(Drugs = yes \mid Race = a) > P(Drugs = yes \mid Race = a')$, this can be exploited to justify an unbalanced stopping process when police resources are limited. How then can we assess its fairness? It is possible to postulate structures analogous to the Berkeley example, where a mechanism such as $Race \rightarrow Economic\ status \rightarrow Drugs$ explains the pathway. Debate would focus on the level of agency of an individual on finding himself or herself at an economic level that leads to increased drug consumption.

But selection bias cuts deeper than that, and more recently causal knowledge has been formally brought in to understand the role of such biases [2, 40]. This is achieved by representing a selection variable *Selected* as part of our model, and carrying out inference by acknowledging that, in the data, all individuals are such that "$Selected = true$". The association between race and drug is expressed as $P(Drugs = yes \mid Race = a, Selected = true) > P(Drugs = yes \mid Race = a', Selected = true)$, which may or may not be representative of the hypothetical population in which everyone has been examined. The data cannot directly tell whether $P(Drugs = yes \mid Race = a, do(Selected = true)) > P(Drugs = yes \mid Race = a', do(Selected = true))$. As an example, it is possible to postulate the two following causal structures that cannot be distinguished on the basis of data already contaminated with selection bias: (i) the structure $Race \rightarrow Drugs$, with *Selected* being a disconnected vertex; (ii) the structure $Race \rightarrow Selected \leftarrow H \rightarrow Drugs$, where H represents hidden variables not formally logged in police records.

In the latter case, we can check that drugs and race and unrelated. However, $P(Drugs = yes \mid Race = a, Selected = true) \neq P(Drugs = yes \mid Race = a', Selected = true)$, as conditioning on *Selected* means that both of its causes *Race* and H "compete" to explain the selection. This induces an association between *Race* and H, which carries over the association between *Race* and *Drugs*. At the same time, $P(Drugs = yes \mid Race = a, do(Selected = true)) = P(Drugs = yes \mid Race = a', do(Selected = true))$, a conclusion that cannot be reached without knowledge of the causal graph or a controlled experiment making use of interventions. Moreover, if the actual structure is a combination of (i) and (ii), standard statistical adjustments that remove the association between *Race* and *Drugs* cannot disentangle effects due to selection bias from those due to the causal link $Race \rightarrow Drugs$, harming any arguments that can be constructed around the agency behind the direct link.

4.3 Fairness Requires Intervention

Approaches to algorithmic fairness usually involve imposing some kind of constraints on the algorithm (such as those formula given by Section 2). We can view this as an intervention on the predicted outcome $\hat{Y}$. And, as argued in [1], we can also try to understand the causal implications for the system we are intervening on. That is, we can use an SCM to model the causal relationships between variables in the data, between those and the predictor $\hat{Y}$ that we are intervening on, and between $\hat{Y}$ and other aspects of the system that will be impacted by decisions made based on the output of the algorithm.

To say that fairness is an intervention is not a strong statement considering that any decision can be considered to be an intervention. Collecting data, using models and algorithms with that data to predict some outcome variable, and making decisions based on those predictions are all intentional acts motivated by a causal hypothesis about the consequences of that course of action. In particular, *not* imposing fairness can also be a deliberate intervention, albeit one of inaction.

We should be clear that prediction problems do not tell the whole story. Breaking the causal links between A and a prediction $\hat{Y}$ is a way of avoiding some unfairness in the world, but it is only one aspect of the problem. Ideally, we would like that no paths from A to Y existed, and the provision of fair predictions is predicated on the belief that it will be a contributing factor for the eventual change in the generation of Y. We are not, however, making any formal claims of modelling how predictive algorithmic fairness will lead to this ideal stage where causal paths from A to Y themselves disappear.

5 Causal Notions of Fairness

In this section we discuss some of the emerging notions of fairness formulated in terms of SCMs, focusing in particular on a notion introduced by us in [23], *counterfactual fairness*. We explain how counterfactual fairness relates to some of the more well-known notions of statistical fairness and in which ways a causal perspective contributes to their interpretation. The remainder of the section will discuss alternative causal notions of fairness and how they relate to counterfactual fairness.

5.1 Counterfactual Fairness

A predictor $\hat{Y}$ is said to satisfy *counterfactual fairness* if

$$P(\hat{Y}(a, U) = y \mid X = x, A = a) = P(\hat{Y}(a', U) = y \mid X = x, A = a), \qquad (8)$$

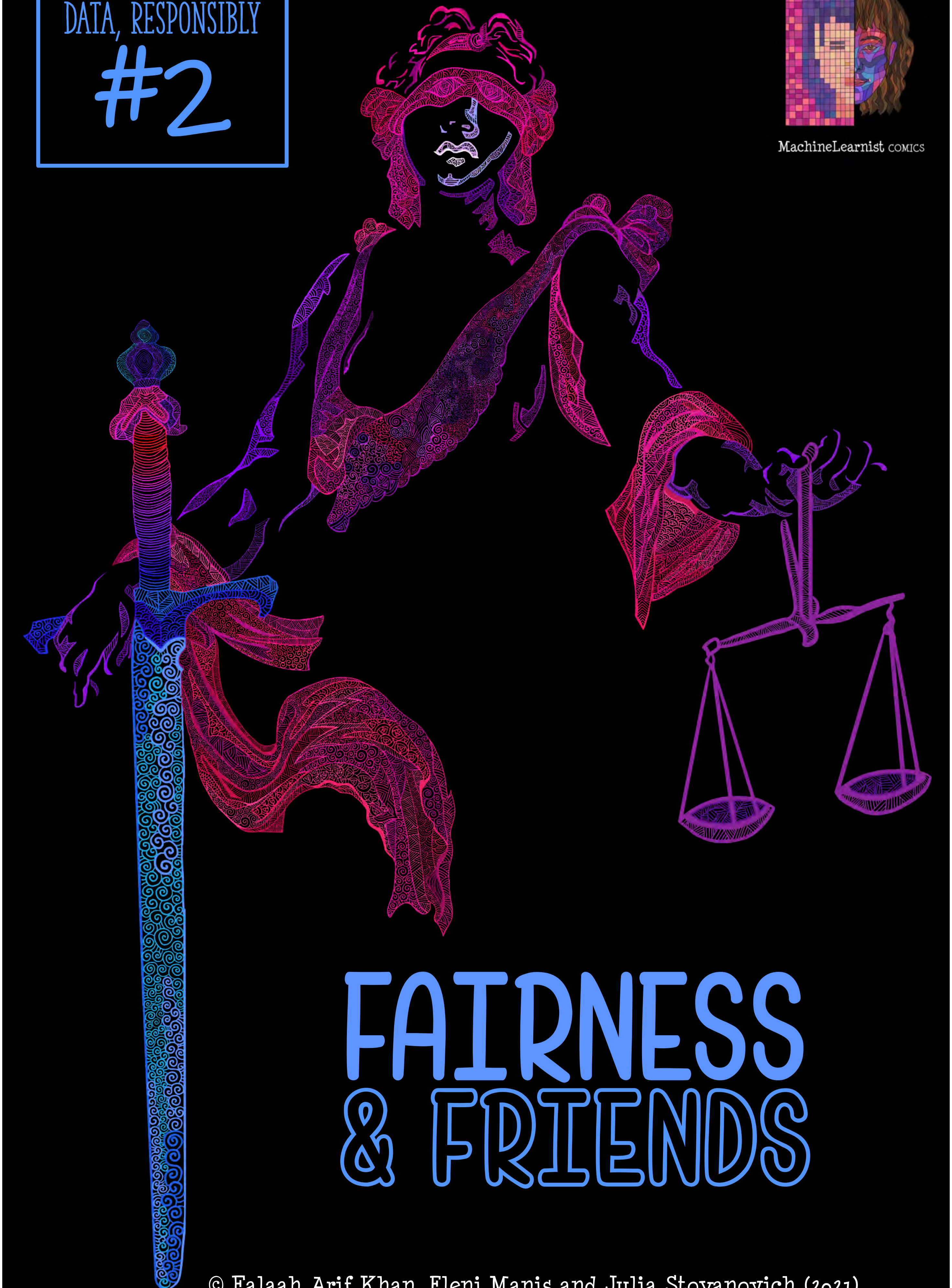
for all y, x, a, a' in the domain of the respective variables [23]. The randomness here is on U (recall that background variables U can be thought of as describing a particular individual person at some point in time). In practice, this means we can build $\hat{Y}$ from any variable Z in the system which is not caused by A

Week 4: Fairness as Equality of Opportunity

DATA, RESPONSIBLY

#2

MachineLearnist COMICS



FAIRNESS & FRIENDS

TERMS OF USE

All the panels in this comic book are licensed CC BY-NC-ND 4.0. Please refer to the license page for details on how you can use this artwork.

TL;DR: Feel free to use panels/groups of panels in your presentations/articles, as long as you

1. Provide the proper citation
2. Do not make modifications to the individual panels themselves

Cite as:

Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. "Fairness and Friends". *Data, Responsibly Comics*, Volume 2 (2021)
https://dataresponsibly.github.io/comics/vol2/fairness_en.pdf

Contact:

Please direct any queries about using elements from this comic to themachinelearnist@gmail.com and cc stoyanovich@nyu.edu

ACCESSIBILITY STATEMENT

The purpose of scientific publication is the presentation of ideas and dissemination of findings. In the course of our (ongoing) work on creating a comic series about Responsible AI, we have found that relatable cartoons and visual humor are a rich but underappreciated source of clarity and accessibility that enable effective communication to a broad audience. Comic books are a particularly prescient medium for literature reviews and critical surveys, and for bridging insights from different disciplines such as philosophy, law, sociology, and computer science. Given the inherently interdisciplinary nature of machine learning, we see comics and other technical artwork as a promising new medium of scholarship. We hope to demonstrate their utility through our work and to popularize their adoption more broadly in the scientific community.

We care deeply about making our comics as digitally accessible as possible. Towards this end, we have taken the following measures:

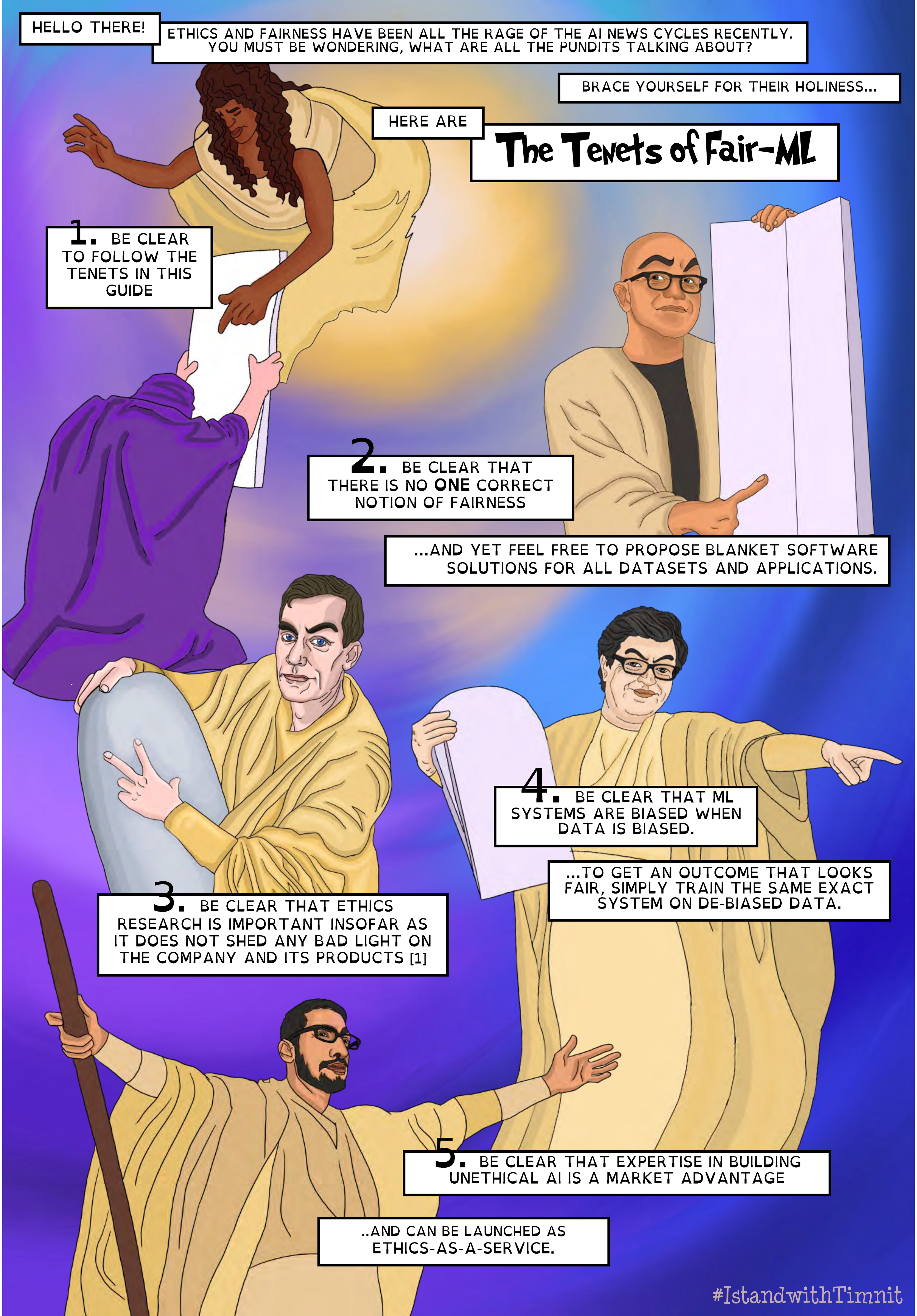
1. We've chosen a typeface that was developed specially for dyslexic readers. All of the major text in the comic is in the "Open Dyslexic" font.

2. The comic book is fully alt-texted and can be read entirely using a screen reader. We are also releasing a complete transcript of the comic book, including all of the text and image descriptions.

3. We will be translating the comic into different languages to cater to speakers of languages other than English, as we have done with previous volumes of the Data, Responsibly comic series.

We would like to thank Amy Hurst and Chancey Fleet for guiding us on the Accessibility front.

Please feel free to reach out to us if you have any recommendations on how we can further improve the accessibility of our comics.



HELLO THERE!
ETHICS AND FAIRNESS HAVE BEEN ALL THE RAGE OF THE AI NEWS CYCLES RECENTLY. YOU MUST BE WONDERING, WHAT ARE ALL THE PUNDITS TALKING ABOUT?
BRACE YOURSELF FOR THEIR HOLINESS...
HERE ARE
The Tenets of Fair-ML
1. BE CLEAR TO FOLLOW THE TENETS IN THIS GUIDE
2. BE CLEAR THAT THERE IS NO ONE CORRECT NOTION OF FAIRNESS
...AND YET FEEL FREE TO PROPOSE BLANKET SOFTWARE SOLUTIONS FOR ALL DATASETS AND APPLICATIONS.
4. BE CLEAR THAT ML SYSTEMS ARE BIASED WHEN DATA IS BIASED.
...TO GET AN OUTCOME THAT LOOKS FAIR, SIMPLY TRAIN THE SAME EXACT SYSTEM ON DE-BIASED DATA.
3. BE CLEAR THAT ETHICS RESEARCH IS IMPORTANT INSOFAR AS IT DOES NOT SHED ANY BAD LIGHT ON THE COMPANY AND ITS PRODUCTS [1]
5. BE CLEAR THAT EXPERTISE IN BUILDING UNETHICAL AI IS A MARKET ADVANTAGE
..AND CAN BE LAUNCHED AS ETHICS-AS-A-SERVICE.
#IstandwithTimnit

TIME OUT. WELCOME TO THE FAIR-ML CLUB.

THERE'S ONLY ONE TENET OF FAIR-ML AND IT'S THAT THERE ARE NO TENETS OF FAIR-ML

FAIRNESS IS **NOT** A TECHNICAL OR STATISTICAL CONCEPT AND THERE CAN NEVER BE A TOOL OR SOFTWARE THAT CAN FULLY 'DE-BIAS' YOUR DATA OR MAKE YOUR MODEL 'FAIR'.

FAIRNESS IS AN ETHICAL CONCEPT, AND A CONTESTED ONE AT THAT. AT BEST, WE CAN SELECT SOME IDEAL OF WHAT IT MEANS TO BE 'FAIR' AND THEN MAKE PROGRESS TOWARDS SATISFYING IT IN OUR PARTICULAR SETTING.

LET'S BACK UP FURTHER, SHALL WE? WHAT ARE WE EVEN TRYING TO MAKE 'FAIR' ? WHAT ARE ALGORITHMS AND WHEN ARE THEY BIASED?

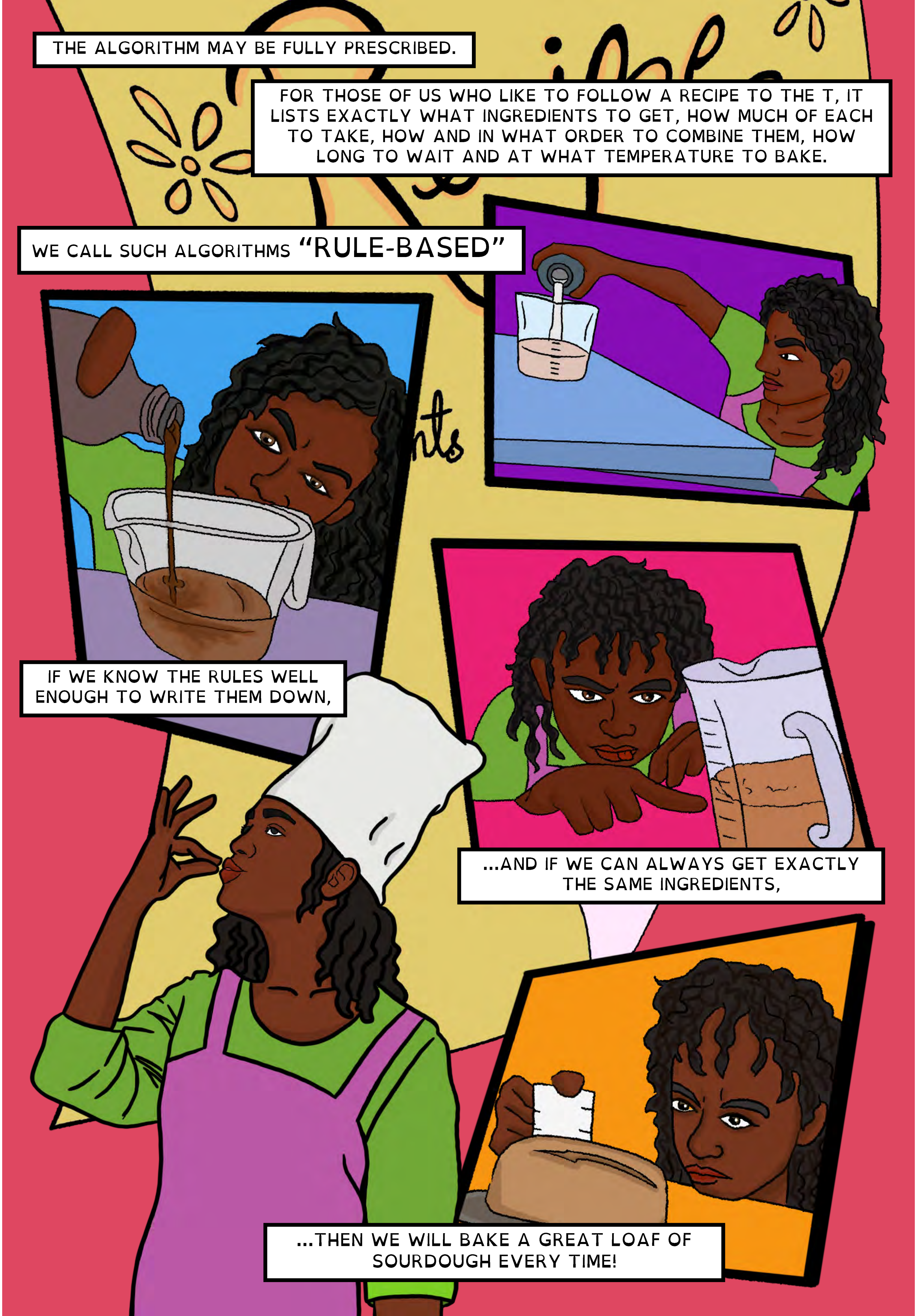
WHAT IS AN ALGORITHM?

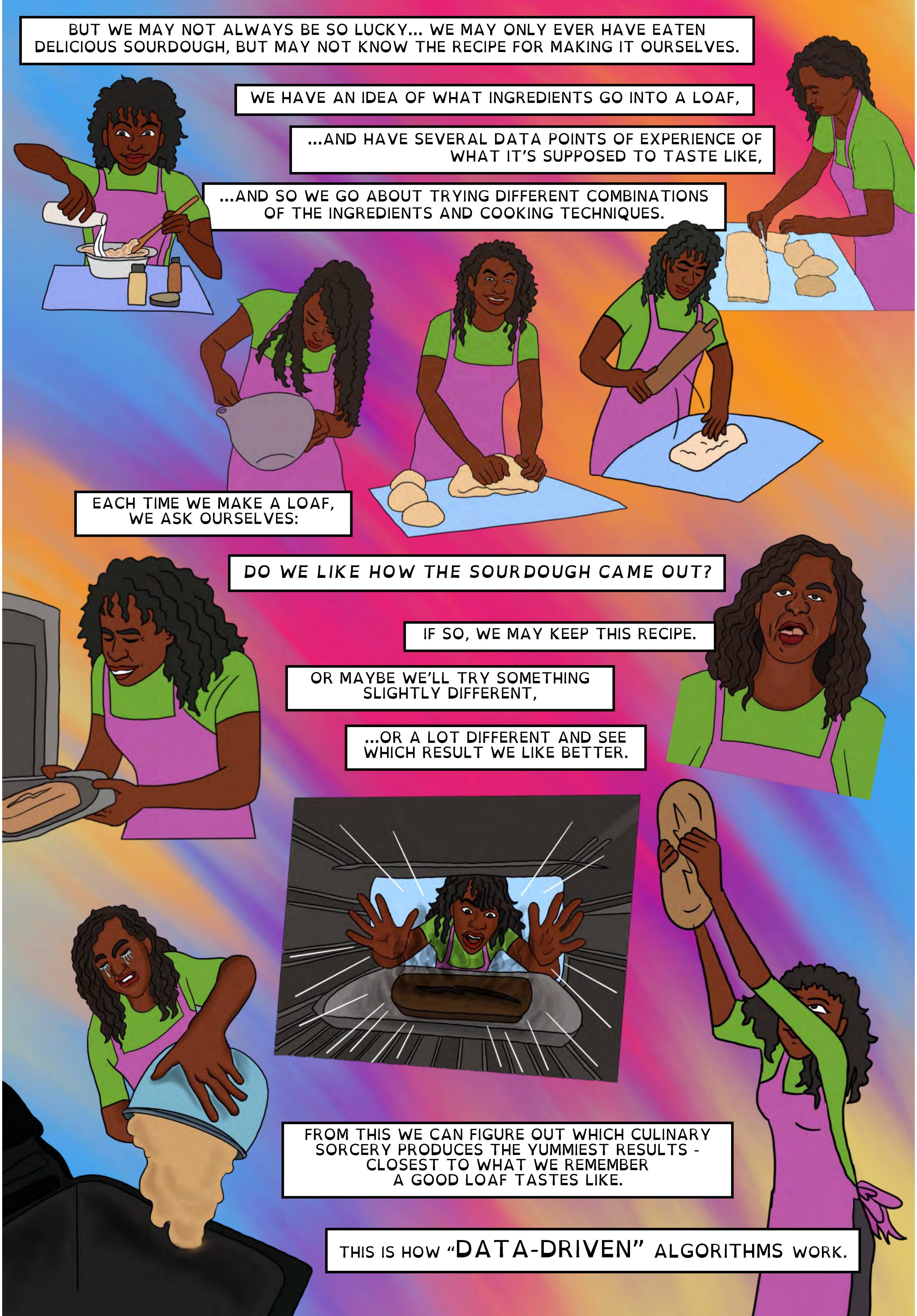
HERE'S A THROWBACK TO THE PREHISTORIC DAYS OF EARLY 2020. REMEMBER THE HOBBY THAT MANY OF US ATTEMPTED TO MASTER - WITH MIXED RESULTS - DURING THE PANDEMIC LOCKDOWN?

BAKING!

THE RECIPE IS THE ALGORITHM: IT LISTS THE INGREDIENTS AND THEIR PROPORTIONS, AND THE STEPS TO TAKE TO TRANSFORM THEM INTO A SCRUMPTIOUS LOAF.

AKIN TO HOW WE EACH HAVE OUR OWN COOKING STYLES, ALGORITHMS ARE OF DIFFERENT TYPES...


THE ALGORITHM MAY BE FULLY PRESCRIBED.
FOR THOSE OF US WHO LIKE TO FOLLOW A RECIPE TO THE T, IT LISTS EXACTLY WHAT INGREDIENTS TO GET, HOW MUCH OF EACH TO TAKE, HOW AND IN WHAT ORDER TO COMBINE THEM, HOW LONG TO WAIT AND AT WHAT TEMPERATURE TO BAKE.
WE CALL SUCH ALGORITHMS "RULE-BASED"
IF WE KNOW THE RULES WELL ENOUGH TO WRITE THEM DOWN,
...AND IF WE CAN ALWAYS GET EXACTLY THE SAME INGREDIENTS,
...THEN WE WILL BAKE A GREAT LOAF OF SOURDOUGH EVERY TIME!



BUT WE MAY NOT ALWAYS BE SO LUCKY… WE MAY ONLY EVER HAVE EATEN DELICIOUS SOURDOUGH, BUT MAY NOT KNOW THE RECIPE FOR MAKING IT OURSELVES.

WE HAVE AN IDEA OF WHAT INGREDIENTS GO INTO A LOAF,

…AND HAVE SEVERAL DATA POINTS OF EXPERIENCE OF WHAT IT'S SUPPOSED TO TASTE LIKE,

…AND SO WE GO ABOUT TRYING DIFFERENT COMBINATIONS OF THE INGREDIENTS AND COOKING TECHNIQUES.

EACH TIME WE MAKE A LOAF, WE ASK OURSELVES:

DO WE LIKE HOW THE SOURDOUGH CAME OUT?

IF SO, WE MAY KEEP THIS RECIPE.

OR MAYBE WE'LL TRY SOMETHING SLIGHTLY DIFFERENT,

…OR A LOT DIFFERENT AND SEE WHICH RESULT WE LIKE BETTER.

FROM THIS WE CAN FIGURE OUT WHICH CULINARY SORCERY PRODUCES THE YUMMIEST RESULTS - CLOSEST TO WHAT WE REMEMBER A GOOD LOAF TASTES LIKE.

THIS IS HOW "DATA-DRIVEN" ALGORITHMS WORK.

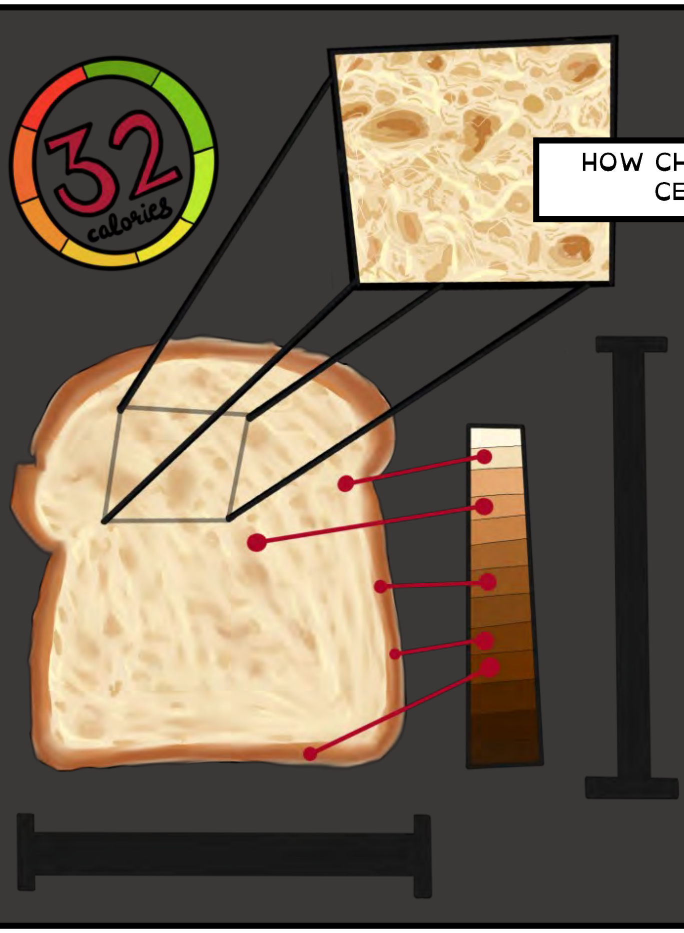
THE RECIPE IS THE ALGORITHM, NOW WHAT ABOUT **THE DATA?**

THE **INPUT** DATA IS THE INGREDIENTS AND THEIR RELATIVE PROPORTIONS.

ANOTHER FORM OF DATA IS THE PARAMETER SETTINGS OF YOUR COOKING EQUIPMENT SUCH AS OVEN TEMPERATURE OR WAIT TIMES.

THEY ARE THE KNOBS YOU CAN TURN TO ADJUST THE RECIPE.

THEN THERE'S DATA THAT DESCRIBES THE **OUTPUT**: THAT SCRUMPTIOUS SOURDOUGH THAT WE REMEMBER DEMOLISHING AND ARE HOPING TO BAKE OURSELVES.



HOW CHEWY IS THE CENTER?

HOW WELL-DONE IS THE CRUST?

WHAT IS IT'S NUTRITIONAL VALUE?

HOW MUCH DOES IT WEIGH?

THESE ARE ALL 'OBJECTIVELY' MEASURABLE FACTORS.

THE FINAL KIND OF DATA IS OUR **REACTION TO THE OUTPUT**

IS IT TASTY?

DOES THE LOAF MEET OUR EXPECTATIONS?

THESE FACTORS BOIL DOWN TO PERSONAL PREFERENCE AND, MORE OFTEN THAN NOT, ARE MORE IMPORTANT THAN THE NUMERICALLY QUANTIFIABLE PROPERTIES OF THE OUTPUT.

WHAT ABOUT **DECISIONS?**

IN THE PROCESS WE DESCRIBED, IN THE COURSE OF EXECUTION OF THE ALGORITHM, WE ARE FACED WITH SEVERAL DECISIONS.

DOES THE DOUGH LOOK GOOD ENOUGH TO PUT INTO THE OVEN?

HAS THE LOAF RISEN ENOUGH AND SHALL WE TAKE IT OUT OF THE OVEN?

IS THE RESULT INSTAGRAM-WORTHY?

ARE WE GIVING IT A THUMBS UP OR A THUMBS DOWN?

A MORE CONSEQUENTIAL DECISION IS - NOW THAT WE'VE TRIED A BUNCH OF RECIPES, WHICH WILL WE CONSIDER A SUCCESS?

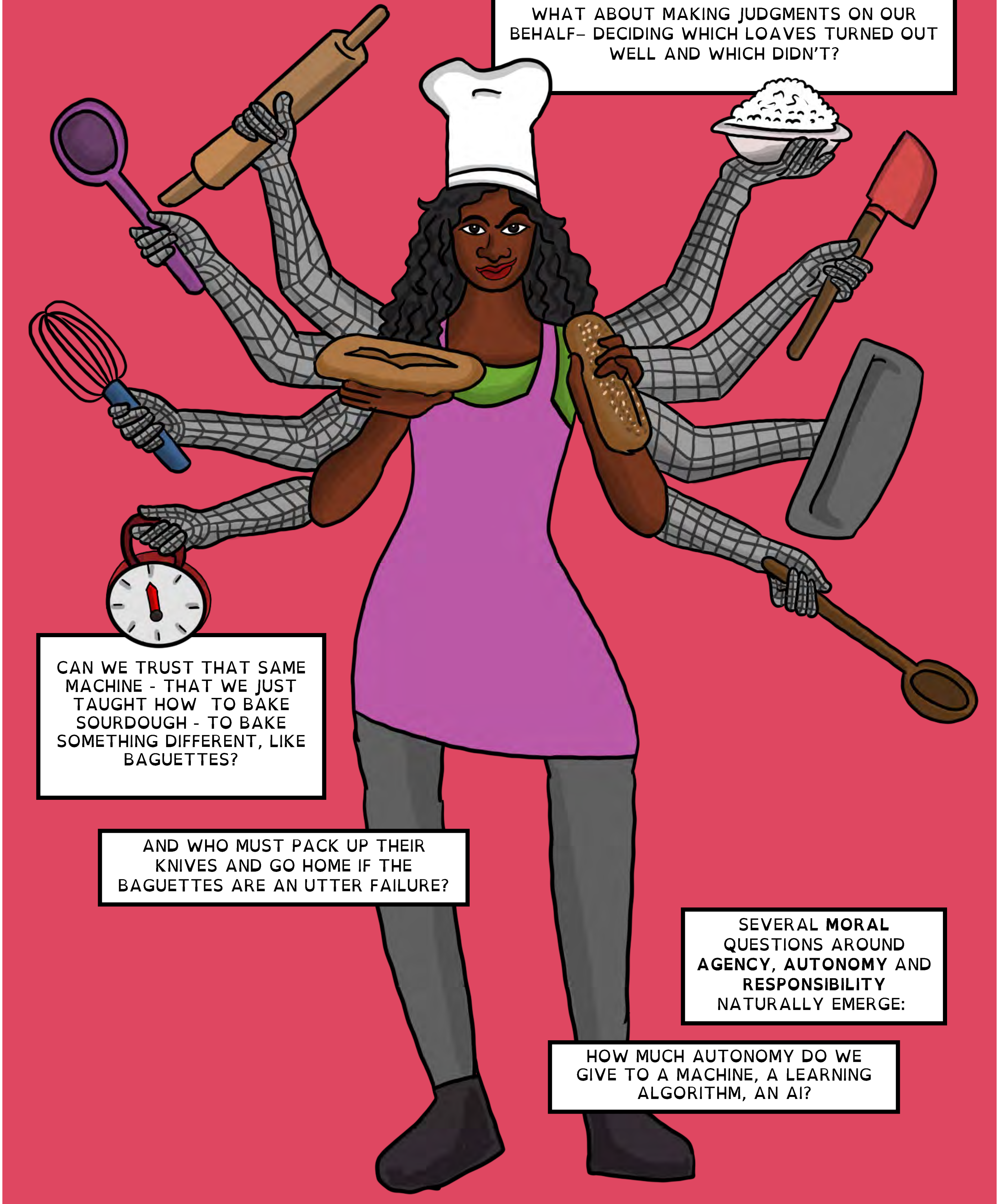
WILL WE SAY THAT IT'S MORE IMPORTANT TO HAVE AN APPETIZING-LOOKING LOAF OR ONE THAT CONSISTENTLY COMES OUT CHEWY ON THE INSIDE AND CRUSTY ON THE OUTSIDE?

WILL WE DECIDE TO ALWAYS - OR NEVER - USE SOME SPECIFIC INGREDIENTS OR COOKING TECHNIQUES?

AN EVEN MORE IMPORTANT DECISION IS - DO WE THINK THAT WE'VE TRIED OUT ENOUGH RECIPES TO PASS OUR EXPERIENCE ON TO A MACHINE,

AND TRUST IT TO BAKE ON OUR BEHALF?

WHAT ABOUT MAKING JUDGMENTS ON OUR BEHALF– DECIDING WHICH LOAVES TURNED OUT WELL AND WHICH DIDN'T?



CAN WE TRUST THAT SAME MACHINE - THAT WE JUST TAUGHT HOW TO BAKE SOURDOUGH - TO BAKE SOMETHING DIFFERENT, LIKE BAGUETTES?

AND WHO MUST PACK UP THEIR KNIVES AND GO HOME IF THE BAGUETTES ARE AN UTTER FAILURE?

SEVERAL **MORAL** QUESTIONS AROUND **AGENCY**, **AUTONOMY** AND **RESPONSIBILITY** NATURALLY EMERGE:

HOW MUCH AUTONOMY DO WE GIVE TO A MACHINE, A LEARNING ALGORITHM, AN AI?

WHAT IS AN ADS?

SO, AN ALGORITHM IS A RECIPE. THEN, WHAT IS AN AUTOMATED DECISION SYSTEM (ADS) ? IS IT LIKE A SELF-BAKING OVEN?

EASY THERE, MUSK-ETEER.

WE DON'T REALLY HAVE A CONSENSUS ON WHAT AN ADS ACTUALLY IS (OR ISN'T).

THE LAW SEEMS TO HAVE TAKEN A PAGE OUT OF THE 'PAULA ABDUL PLAYBOOK OF JUDGING', GOING OVERLY LENIENT AND VAGUE IN ITS DEFINITION.

NEW YORK CITY'S LOCAL LAW 49 DEFINES AN ADS AS *"COMPUTERIZED IMPLEMENTATIONS OF ALGORITHMS, INCLUDING THOSE DERIVED FROM MACHINE LEARNING OR OTHER DATA PROCESSING OR ARTIFICIAL INTELLIGENCE TECHNIQUES, WHICH ARE USED TO MAKE OR ASSIST IN MAKING DECISIONS."* [2]

USING THIS DEFINITION, ONE COULD ARGUE THAT SPREADSHEETS OR EVEN INTERNET SEARCHES COULD BE ADS, BECAUSE THEY ARE, IN FACT, COMPUTERIZED AND DO, IN FACT, GUIDE DECISION-MAKING. [3]

A PRECISE DEFINITION WILL BE CRUCIAL FOR THE EFFICACY OF ANY ATTEMPT AT REGULATING THESE SYSTEMS. AN ALTERNATE APPROACH WOULD BE TO DEFINE ADS BY EXTENSION. [4]

SO YOU THINK YOU'RE AN ADS?

DO YOU:

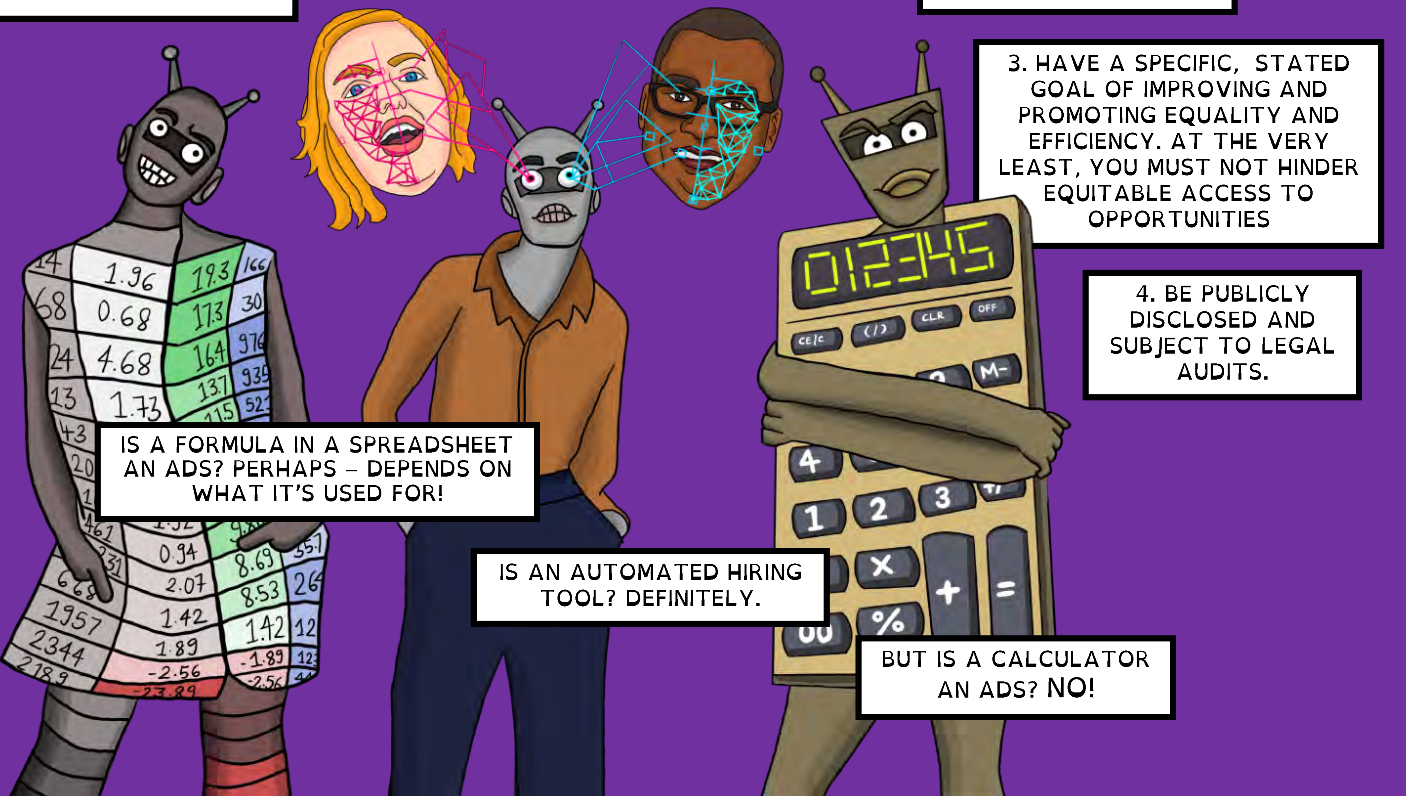
1. PROCESS DATA ABOUT PEOPLE

2. ASSIST - EITHER IN COMBINATION WITH HUMAN DECISION MAKING OR AUTONOMOUSLY - IN MAKING CONSEQUENTIAL DECISIONS THAT IMPACT PEOPLE'S LIVES.

ADDITIONALLY, WE WOULD LIKE IT IF YOU WOULD:

3. HAVE A SPECIFIC, STATED GOAL OF IMPROVING AND PROMOTING EQUALITY AND EFFICIENCY. AT THE VERY LEAST, YOU MUST NOT HINDER EQUITABLE ACCESS TO OPPORTUNITIES

4. BE PUBLICLY DISCLOSED AND SUBJECT TO LEGAL AUDITS.

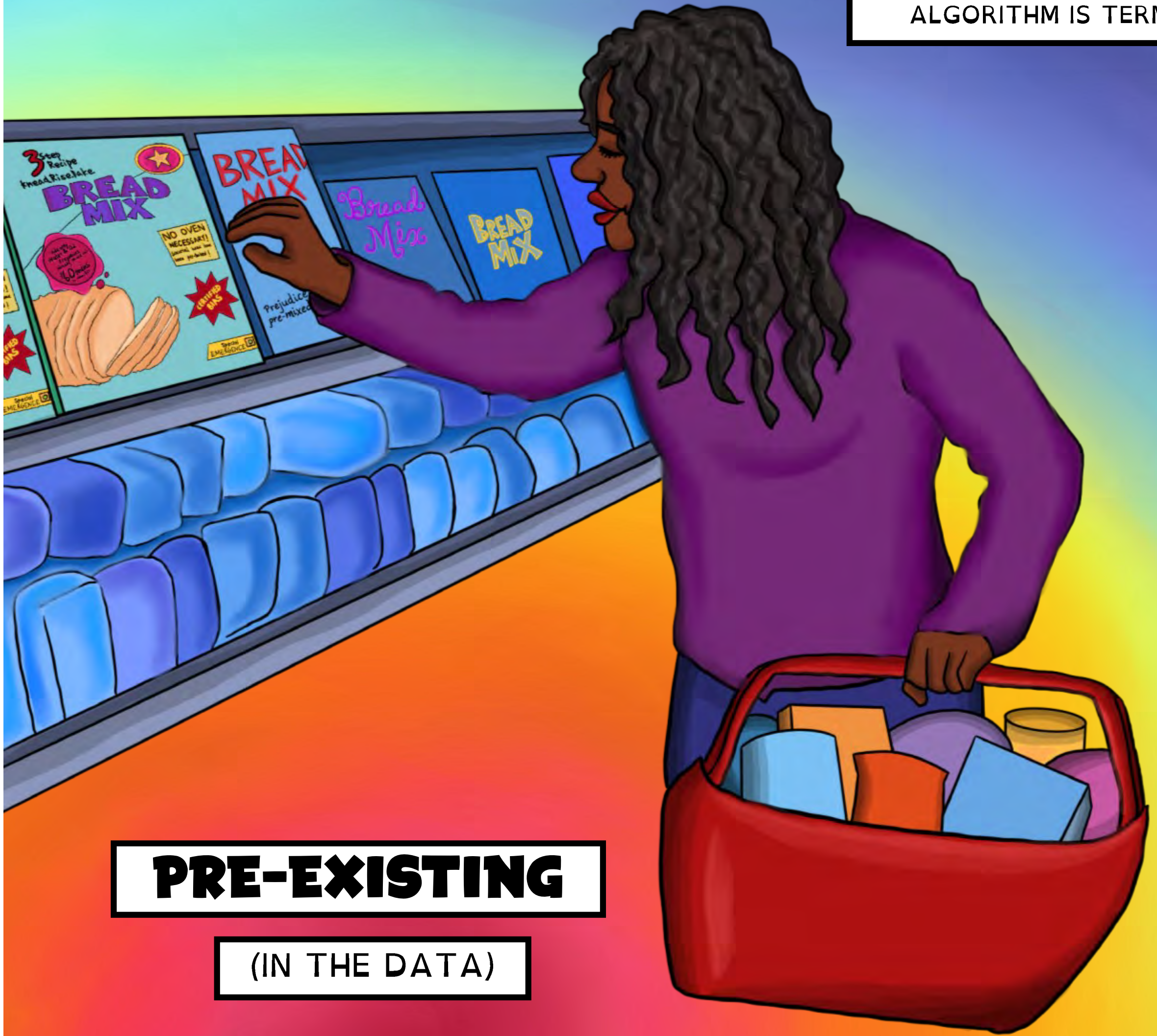

IS A FORMULA IN A SPREADSHEET AN ADS? PERHAPS – DEPENDS ON WHAT IT'S USED FOR!
IS AN AUTOMATED HIRING TOOL? DEFINITELY.
BUT IS A CALCULATOR AN ADS? NO!
012345

ALL ABOUT THAT BIAS...

WITH THAT IN MIND, NOW LET'S LOOK AT WHAT WE MEAN BY BIAS IN AN ADS AND HOW IT ARISES. [5]

IN THE CONTEXT OF DATA-DRIVEN SYSTEMS, BIASES ARE 'HARMFUL' ASSOCIATIONS PICKED UP BY THE ALGORITHM - EITHER FROM THE DATA ITSELF, OR FROM HOW THE ALGORITHM IS DESIGNED, OR FROM THE OBJECTIVES THAT WE SPECIFIED FOR IT, OR FROM HOW WE USE IT.

SYSTEMATIC DISCRIMINATION BY AN ALGORITHM IS TERMED 'BIAS'.



PRE-EXISTING

(IN THE DATA)

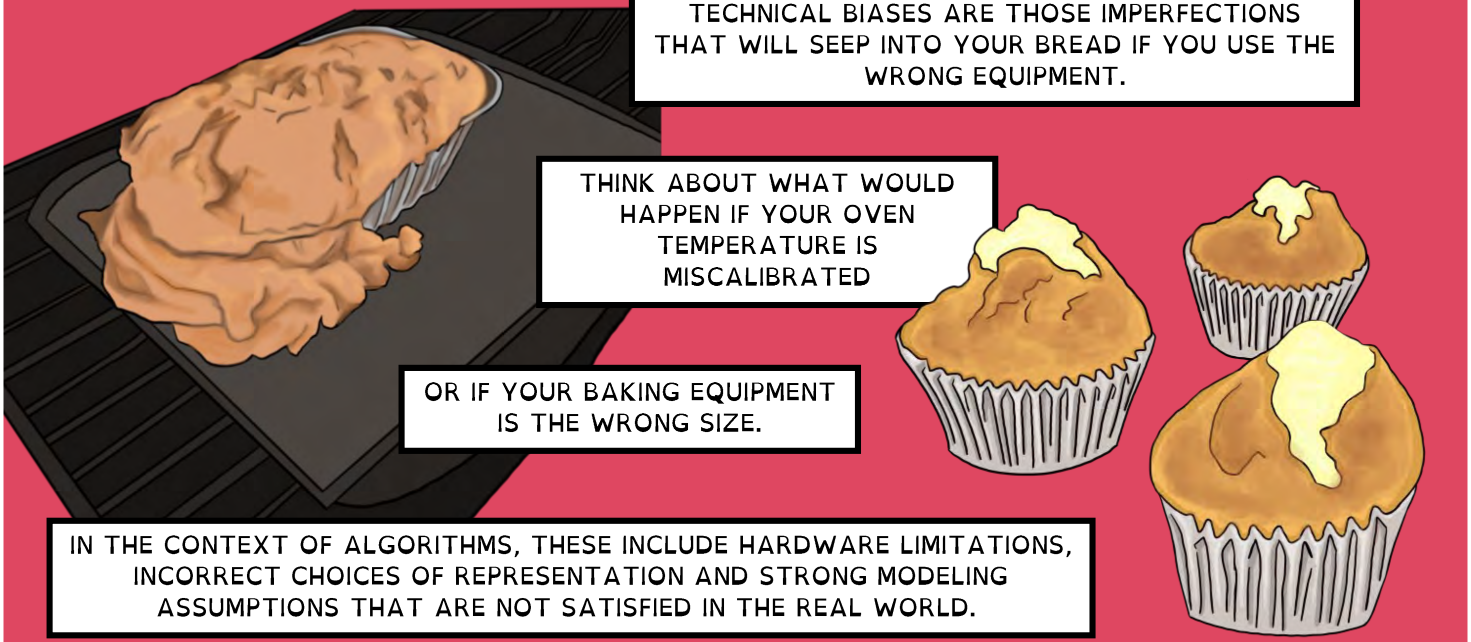
PRE-EXISTING BIASES EXIST IN SOCIETY AND COME 'PRE-BAKED' INTO THE MODEL AS A RESULT OF THE UNDERLYING DISCRIMINATORY SYSTEM THAT THE DATA WAS GENERATED FROM.

THESE WOULD BE THE FLAVOR NOTES THAT WILL SEEP INTO YOUR BREAD IF YOU DON'T PRIORITIZE THE PURITY/FRESHNESS OF YOUR INGREDIENTS OR IF YOU DECIDE TO USE PREMIXED OFF-THE-SHELF BATTER.

A NOTORIOUS EXAMPLE IS THE GENDER AND RACIAL STEREOTYPES THAT LANGUAGE MODELS PICK UP WHEN TRAINED ON DATA FROM SOCIAL MEDIA PLATFORMS.

TECHNICAL

(IN THE TECHNICAL SYSTEM)



TECHNICAL BIASES ARE THOSE IMPERFECTIONS THAT WILL SEEP INTO YOUR BREAD IF YOU USE THE WRONG EQUIPMENT.

THINK ABOUT WHAT WOULD HAPPEN IF YOUR OVEN TEMPERATURE IS MISCALIBRATED

OR IF YOUR BAKING EQUIPMENT IS THE WRONG SIZE.

IN THE CONTEXT OF ALGORITHMS, THESE INCLUDE HARDWARE LIMITATIONS, INCORRECT CHOICES OF REPRESENTATION AND STRONG MODELING ASSUMPTIONS THAT ARE NOT SATISFIED IN THE REAL WORLD.

EMERGENT

(DUE TO DECISIONS)

THE PATTERNS THAT EMERGE AS A RESULT OF YOUR BAKING COMPRISE 'EMERGENT' BIAS.

WHAT IF YOU BECOME SUCH A MAESTRO AT BAKING THAT YOU INADVERTENTLY MAKE BREAD A STEADY PART OF YOUR DIET!

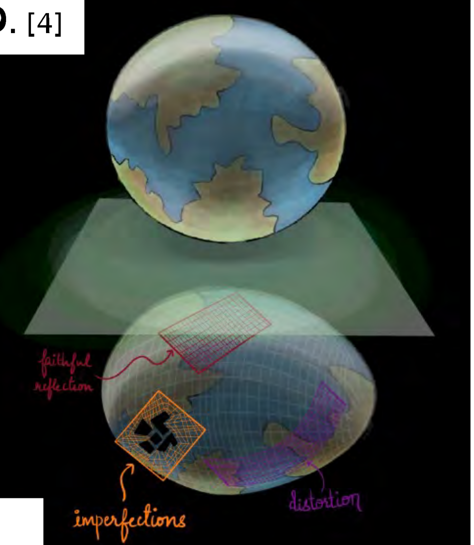
OR MAKE IT SO OFTEN, THAT YOU TURN EVERYONE AROUND YOU OFF THE THOUGHT OF EVER EATING ANOTHER SLICE!

OR THINK ABOUT HOW YOUR IDEA OF 'WHAT BREAD SHOULD TASTE LIKE' IS SHAPED BY THE POPULARITY OF PRODUCTS LIKE 'WONDER BREAD'.

DATA IS A MIRROR REFLECTION OF THE **WORLD**. [4]

ALL WE HAVE IS A DISTORTED (BIASED) REFLECTION.

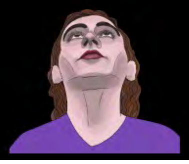
WITHOUT KNOWLEDGE OR ASSUMPTIONS ABOUT THE PROPERTIES OF THE MIRROR AND OF THE WORLD IT REFLECTS, WE CANNOT KNOW WHETHER WE ARE LOOKING AT A DISTORTED REFLECTION OF A PERFECT WORLD OR A PERFECT REFLECTION OF A DISTORTED WORLD OR WHETHER THESE DISTORTIONS COMPOUND. [6]



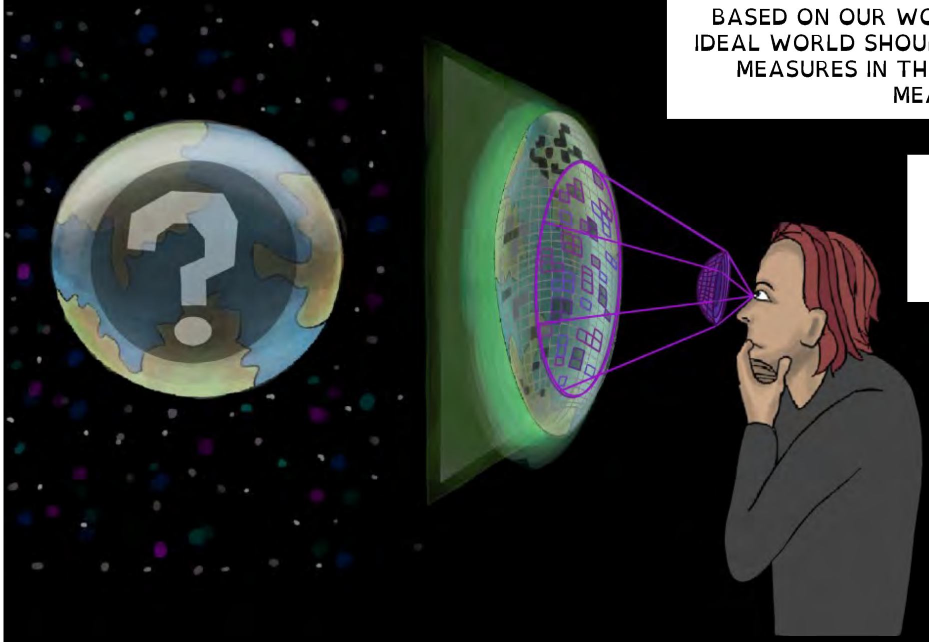
WHAT IS ALGORITHMIC FAIRNESS?

ALGORITHMIC FAIRNESS IS THE CORRECTIVE LENS THAT WE WEAR IN ORDER TO SEE THE WORLD CLOSER TO WHAT WE WANT IT TO LOOK LIKE THAN WHAT IT ACTUALLY IS.

CORRECTIVE LENSES ARE TAILORED TO THE WEARER AND, SIMILARLY, DIFFERENT INDIVIDUALS JUDGE DIFFERENT FAIRNESS IDEALS TO MATTER, FOR DIFFERENT REASONS.



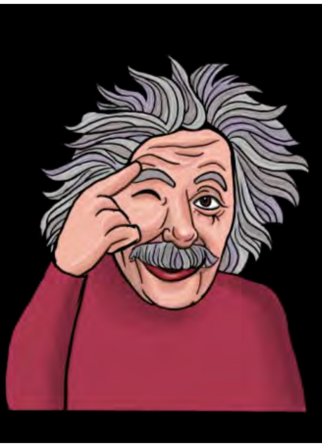
BASED ON OUR WORLDVIEW (BELIEFS ABOUT WHAT THE IDEAL WORLD SHOULD LOOK LIKE), WE APPLY CORRECTIVE MEASURES IN THE FORM OF DIFFERENT STATISTICAL MEASURES OF 'FAIRNESS'.



HOWEVER, WEARING THESE LENSES ONLY CHANGES HOW WE VIEW THE REFLECTION - IT DOES NOT AND CANNOT FIX DISTORTIONS IN THE MIRROR OR FIX DISTORTIONS IN THE WORLD.

UNLESS SUCH FIXES ARE SUPPLEMENTED BY SYSTEMIC CHANGE, WE CAN QUICKLY CONFUSE THE WORLD SEEN THROUGH ROSE-COLORED GLASSES WITH THE REAL WORLD.

ALGORITHMIC DECISIONS ARE MAPPINGS BETWEEN THREE 'SPACES', NAMELY - THE CONSTRUCT SPACE (THE REAL WORLD), THE OBSERVED SPACE (THE REFLECTION) AND THE DECISION SPACE (THE OUTCOMES OR ALLOCATIONS). [7]



"INTELLIGENCE" IS THE CONSTRUCT.

TEST SCORES ARE THE OBSERVATIONS THAT WE ARE ACTUALLY ABLE TO MEASURE.

THE DECISION IS WHETHER OR NOT TO CERTIFY ONE'S INTELLECTUAL ABILITY BY CONFERRING UPON THEM A DIPLOMA

IN A PERFECT WORLD, WHERE THERE IS NEITHER A DISTORTION IN THE WORLD NOR IN THE REFLECTION, OUR CONSTRUCTS AND OUR OBSERVATIONS WOULD BE THE SAME.

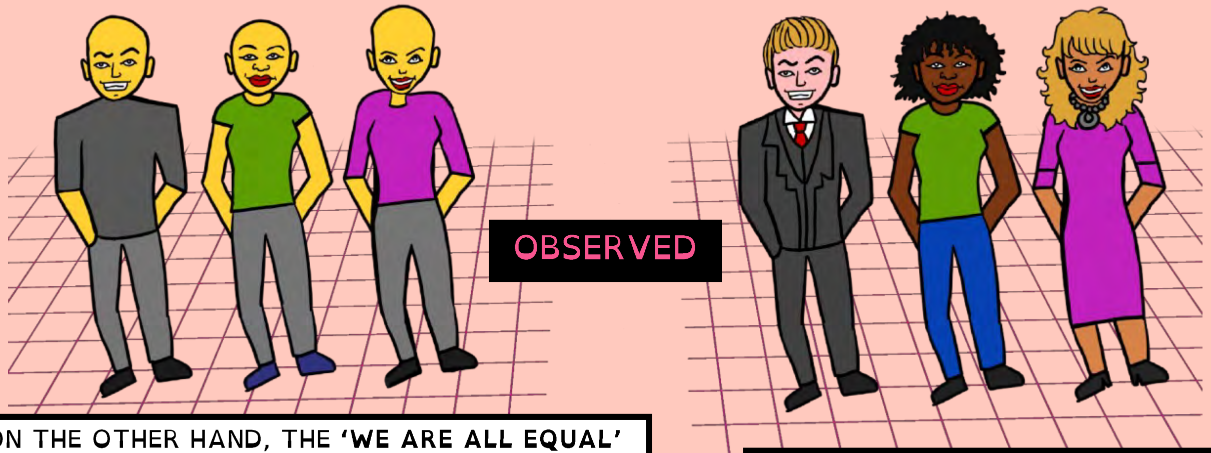
IN REALITY, THE CONSTRUCT SPACE IS UNOBSERVABLE AND SO WE NEED TO MAKE ASSUMPTIONS ABOUT ITS NATURE AND ABOUT THE MAPPING FROM CONSTRUCT TO OBSERVATION. THESE ASSUMPTIONS COLOR OUR JUDGMENTS ABOUT WHETHER ALLOCATIONS OF BENEFITS ARE ‘FAIR’ (BY SOME SPECIFIC NOTION).

DIFFERENT WORLDVIEWS AFFECT OUR INTUITIONS ABOUT ‘FAIRNESS’. [7]

THE **‘WHAT YOU SEE IS WHAT YOU GET’** WORLDVIEW ASSUMES THAT RELEVANT CHARACTERISTICS ARE CORRECTLY CAPTURED IN THE DATA

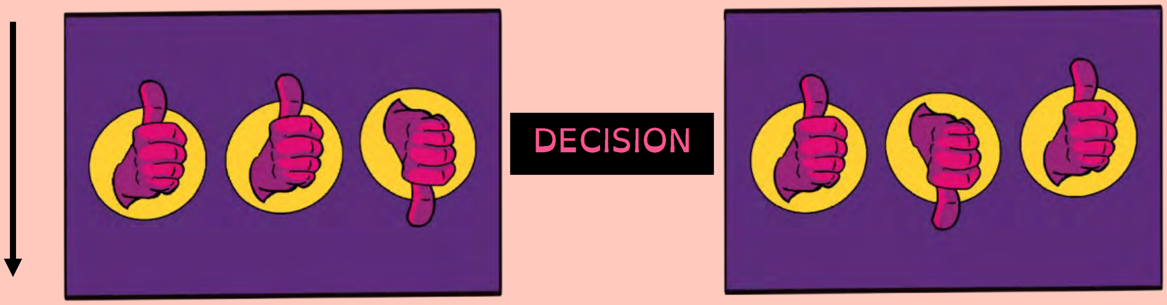
AND THAT DIFFERENCES AMONG PEOPLE’S ABILITIES (BY SOME TASK-SPECIFIC DISTANCE METRIC) ARE PRESERVED FROM THE CONSTRUCT SPACE TO THE OBSERVED SPACE.

CONSTRUCT
We're All Equal
What You See Is What You Get



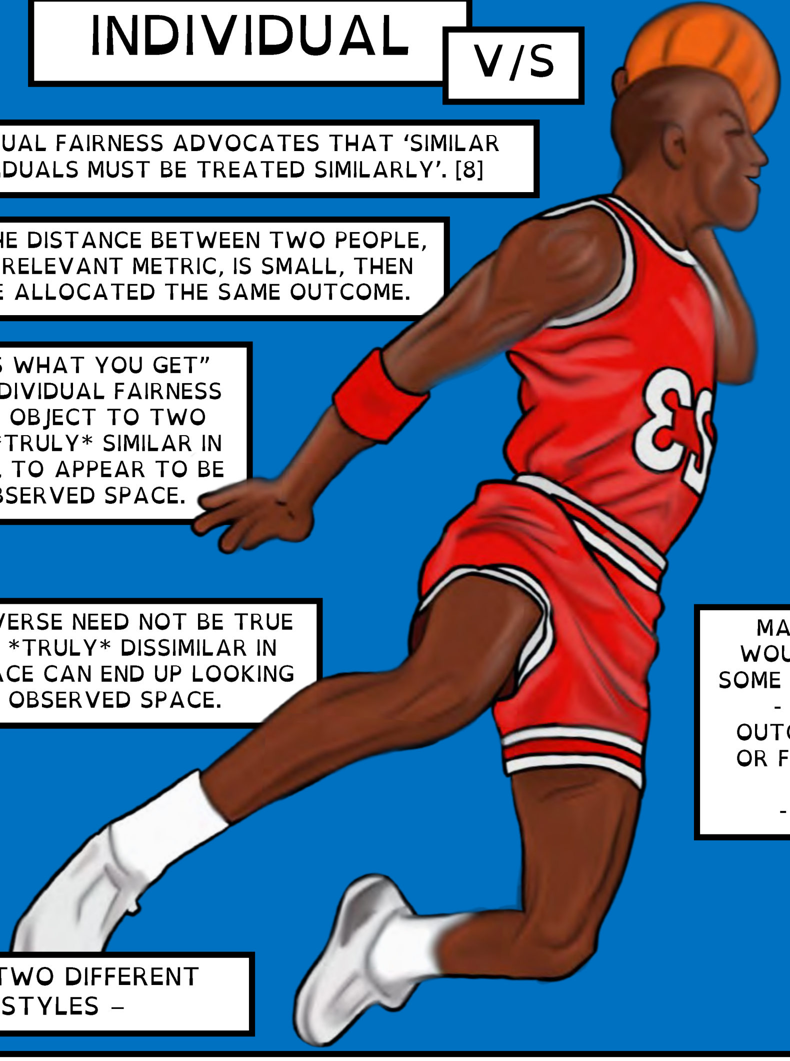
ON THE OTHER HAND, THE **‘WE ARE ALL EQUAL’** WORLDVIEW IS BASED ON THE IDEA THAT DIFFERENCES IN PEOPLE’S OBSERVED ABILITIES ARE ATTRIBUTABLE TO FACTORS OUTSIDE OF THEIR CONTROL.

THIS IS GOOD NEWS! IF WE CAN CORRECTLY MEASURE PEOPLE’S *TRUE* ABILITIES, WE CAN MAKE ‘FAIR’ DECISIONS.



IN SO FAR THAT PEOPLE’S ABILITIES CAN BE MEASURED IN A MANNER THAT IS INDEPENDENT OF THEIR PROTECTED CHARACTERISTICS SUCH AS SEX AND RACE, WE CAN MAKE ‘FAIR’ DECISIONS.

INDIVIDUAL V/S GROUP



INDIVIDUAL FAIRNESS ADVOCATES THAT 'SIMILAR INDIVIDUALS MUST BE TREATED SIMILARLY'. [8]

MATHEMATICALLY, IF THE DISTANCE BETWEEN TWO PEOPLE, BASED ON SOME TASK-RELEVANT METRIC, IS SMALL, THEN THEY SHOULD BOTH BE ALLOCATED THE SAME OUTCOME.

THE "WHAT YOU SEE IS WHAT YOU GET" WORLDVIEW TRACKS INDIVIDUAL FAIRNESS INSOFAR THAT IT WILL OBJECT TO TWO INDIVIDUALS WHO ARE *TRULY* SIMILAR IN THE CONSTRUCT SPACE, TO APPEAR TO BE DISSIMILAR IN THE OBSERVED SPACE.

HOWEVER, THE CONVERSE NEED NOT BE TRUE – PEOPLE WHO ARE *TRULY* DISSIMILAR IN THE CONSTRUCT SPACE CAN END UP LOOKING SIMILAR IN THE OBSERVED SPACE.

GROUP FAIRNESS TRIES TO ENSURE SOME NOTION OF PARITY IN OUTCOMES FOR MEMBERS OF DIFFERENT PROTECTED GROUPS.

MATHEMATICALLY, WE WOULD AIM TO EQUALIZE SOME STATISTICAL MEASURE - SUCH AS POSITIVE OUTCOMES, ERROR RATES OR FALSE POSITIVE/FALSE NEGATIVE RATES - ACROSS GROUPS.

THINK OF IT AS TWO DIFFERENT COACHING STYLES –

ARE YOU THE DOUG COLLINS OF THE '86-'88 BULLS, DESIGNING YOUR ENTIRE OFFENSE AROUND YOUR MOST TALENTED PLAYER - EAGER TO SEE HIM EARN HIS PLACE AMONG THE ALL-TIME GREATS?

OR ARE YOU THE PHIL JACKSON OF THE BULLS, IDENTIFYING THE DIFFERENT STRENGTHS OF DIFFERENT PLAYERS AND ORGANIZING THE TRIANGLE OFFENSE TO PERFECTION,

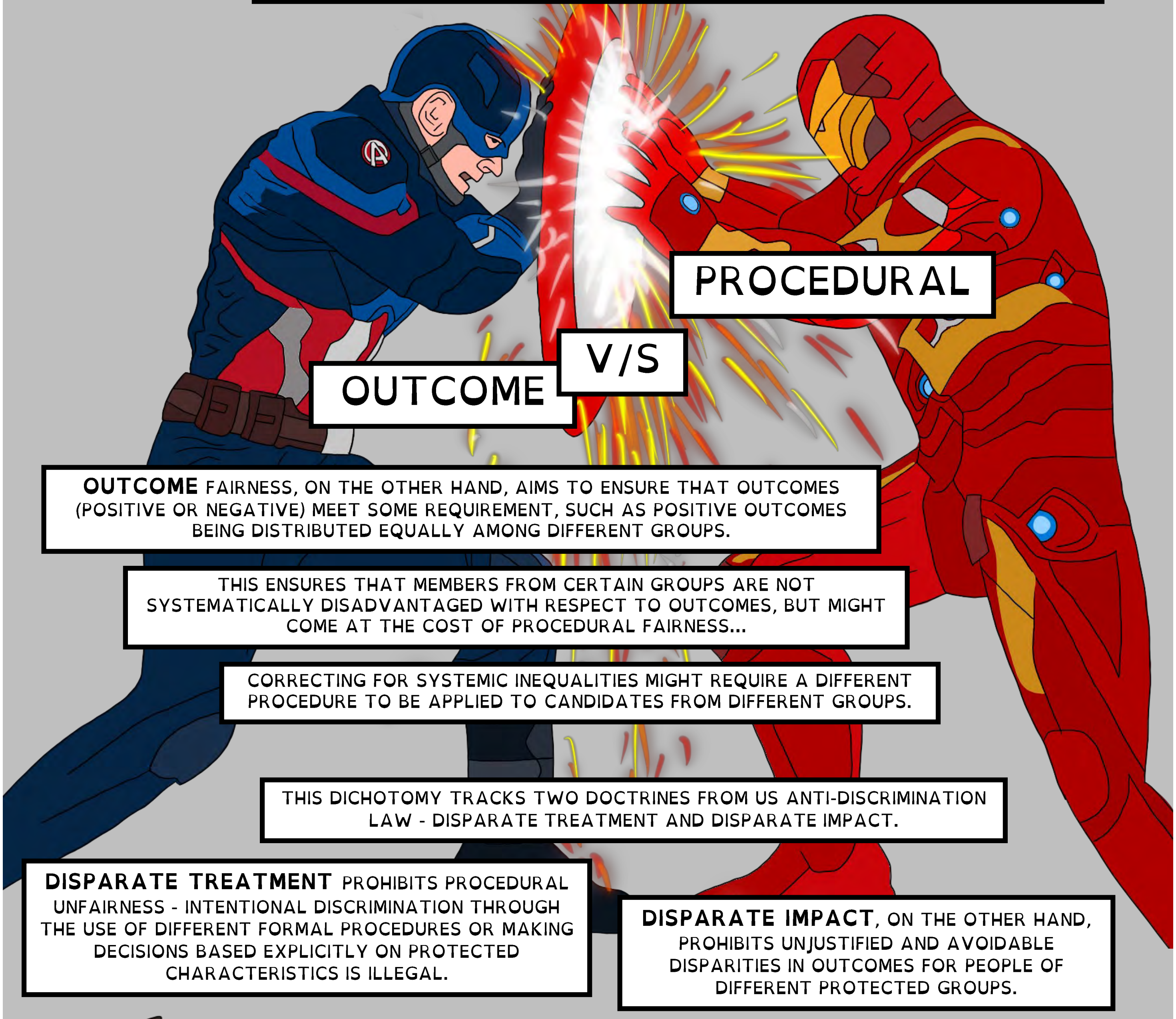
...THEREBY TAKING THE BULLS – LED BY THE INIMITABLE JORDAN, OF COURSE - TO THEIR FIRST CHAMPIONSHIP VICTORY.

IN PRINCIPLE, INDIVIDUAL AND GROUP FAIRNESS NEED NOT BE INCOMPATIBLE [9] – YOU CAN PULL OFF TWO 'THREEPEAT' CHAMPIONSHIP WINS, WHILE HAVING JORDAN WIN LEAGUE MVP EACH YEAR.

A SECOND DICHOTOMY ARISES FROM THE WAY IN WHICH WE ARRIVE AT A 'FAIR' DECISION.

PROCEDURAL FAIRNESS EMPHASIZES THAT THE SAME PROCESS BE APPLIED TO ALL INDIVIDUALS,

IRRESPECTIVE OF THE SOCIETAL FACTORS THAT MIGHT ADVANTAGE SOME AND DISADVANTAGE OTHERS IN GETTING A 'FAIR' SHOT IN THE SELECTION PROCESS.



OUTCOME FAIRNESS, ON THE OTHER HAND, AIMS TO ENSURE THAT OUTCOMES (POSITIVE OR NEGATIVE) MEET SOME REQUIREMENT, SUCH AS POSITIVE OUTCOMES BEING DISTRIBUTED EQUALLY AMONG DIFFERENT GROUPS.

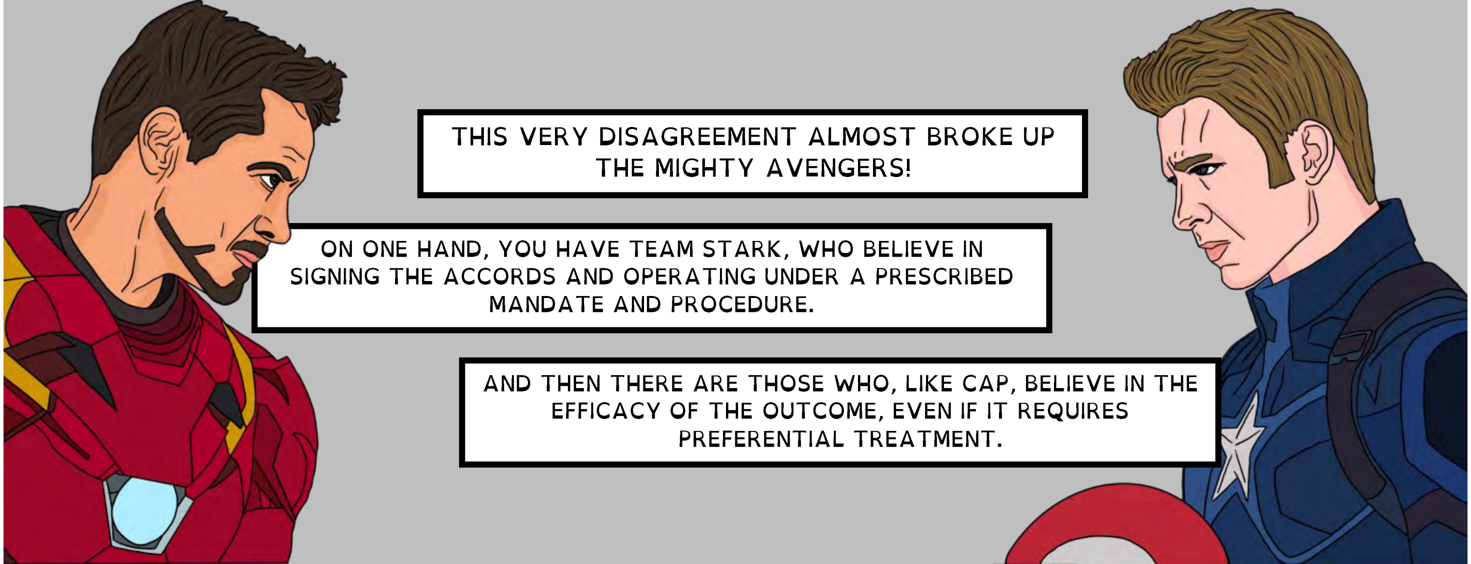
THIS ENSURES THAT MEMBERS FROM CERTAIN GROUPS ARE NOT SYSTEMATICALLY DISADVANTAGED WITH RESPECT TO OUTCOMES, BUT MIGHT COME AT THE COST OF PROCEDURAL FAIRNESS...

CORRECTING FOR SYSTEMIC INEQUALITIES MIGHT REQUIRE A DIFFERENT PROCEDURE TO BE APPLIED TO CANDIDATES FROM DIFFERENT GROUPS.

THIS DICHOTOMY TRACKS TWO DOCTRINES FROM US ANTI-DISCRIMINATION LAW - DISPARATE TREATMENT AND DISPARATE IMPACT.

DISPARATE TREATMENT PROHIBITS PROCEDURAL UNFAIRNESS - INTENTIONAL DISCRIMINATION THROUGH THE USE OF DIFFERENT FORMAL PROCEDURES OR MAKING DECISIONS BASED EXPLICITLY ON PROTECTED CHARACTERISTICS IS ILLEGAL.

DISPARATE IMPACT, ON THE OTHER HAND, PROHIBITS UNJUSTIFIED AND AVOIDABLE DISPARITIES IN OUTCOMES FOR PEOPLE OF DIFFERENT PROTECTED GROUPS.



THIS VERY DISAGREEMENT ALMOST BROKE UP THE MIGHTY AVENGERS!

ON ONE HAND, YOU HAVE TEAM STARK, WHO BELIEVE IN SIGNING THE ACCORDS AND OPERATING UNDER A PRESCRIBED MANDATE AND PROCEDURE.

AND THEN THERE ARE THOSE WHO, LIKE CAP, BELIEVE IN THE EFFICACY OF THE OUTCOME, EVEN IF IT REQUIRES PREFERENTIAL TREATMENT.

THE FAMOUS IMPOSSIBILITY RESULTS [10, 11] HAVE DECREED THAT DIFFERENT MEASURES OF 'FAIRNESS' IN PREDICTIONS ARE MUTUALLY INCOMPATIBLE.

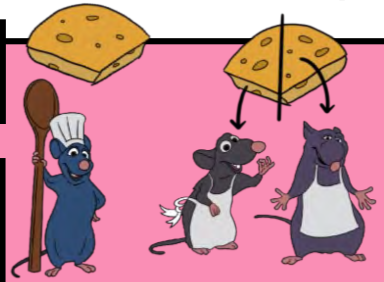
HERE'S AN EXAMPLE OF IMPOSSIBILITY IN 'FAIR' RESOURCE ALLOCATION-

SAY YOU NEED TO REWARD YOUR HUNGRY, HUNGRY HELPERS. AND SAY YOUR HELPERS ARE OF DIFFERENT AGES AND CULINARY EXPERTISE. HOW DO YOU GO ABOUT MAKING THIS ALLOCATION?

(EXECUTIVE)
(SOUS CHEFS)
(LINE CHEFS)

IF YOU DECIDE THAT THE 'FAIR' WAY TO DO THIS WOULD BE TO ENSURE THAT YOU WILL SPLIT THE PIE INTO THREE EQUAL PARTS

- ONE FOR EACH LEVEL OF CULINARY EXPERTISE,



THEN EACH ROOKIE WOULD GET LESS THAN EACH EXECUTIVE CHEF

(OR)

- PURELY DUE TO THAT FACT THAT THERE ARE MORE ROOKIES!

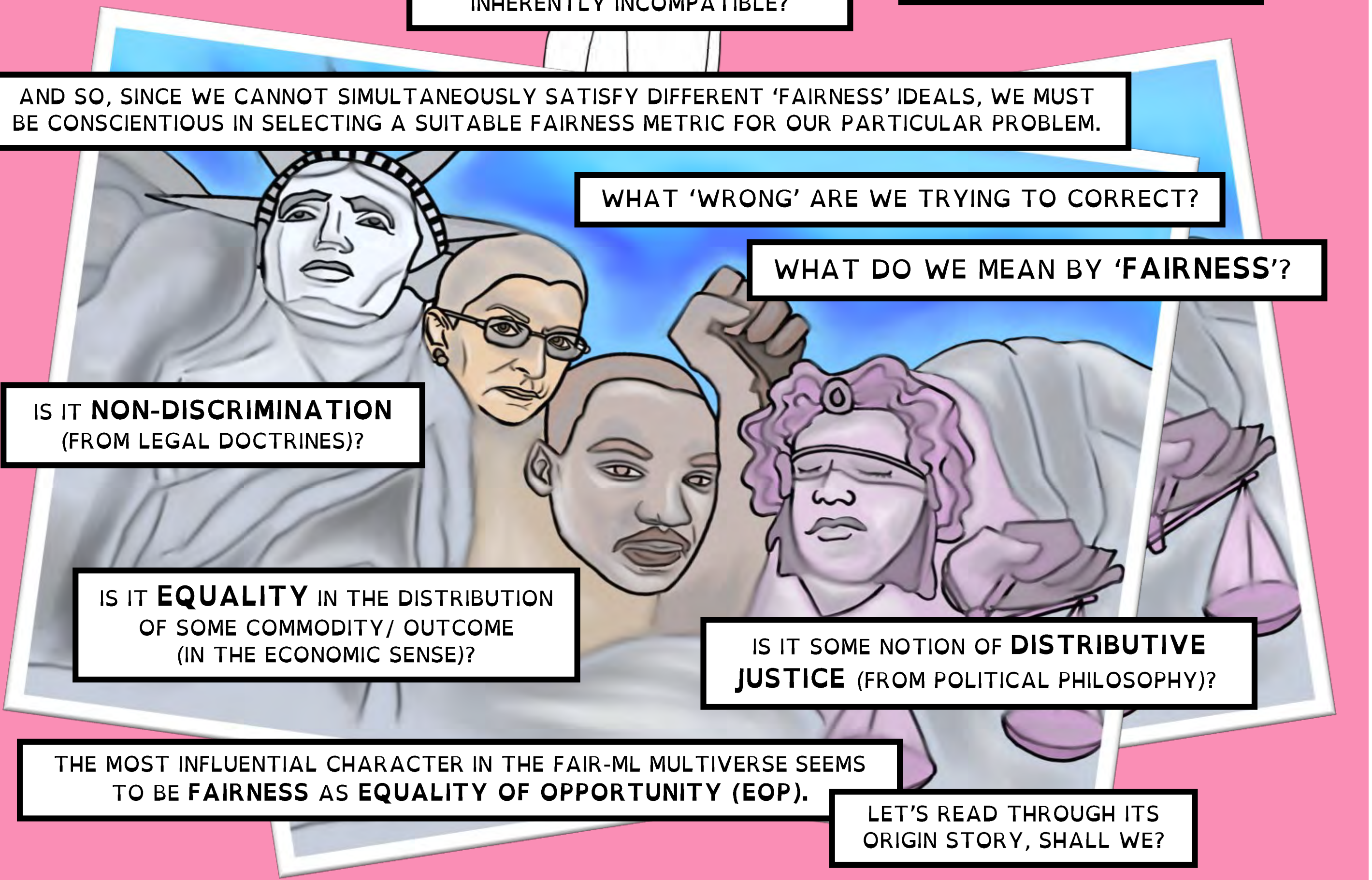
2
3
1
5
6
4
1
2
3
4
5
6

IF YOU DECIDE INSTEAD TO GIVE EACH CHEF THE SAME AMOUNT OF FOOD,

THEN IT WOULD BE IMPOSSIBLE TO HAVE PARITY IN OUTCOMES FOR ALL GROUPS

THERE WOULD BE MUCH MORE FOOD OVERALL GIVEN TO THE ROOKIE GROUP.

SEE, HOW DIFFERENT METRICS ARE INHERENTLY INCOMPATIBLE?



AND SO, SINCE WE CANNOT SIMULTANEOUSLY SATISFY DIFFERENT 'FAIRNESS' IDEALS, WE MUST BE CONSCIENTIOUS IN SELECTING A SUITABLE FAIRNESS METRIC FOR OUR PARTICULAR PROBLEM.

WHAT 'WRONG' ARE WE TRYING TO CORRECT?

WHAT DO WE MEAN BY **'FAIRNESS'**?

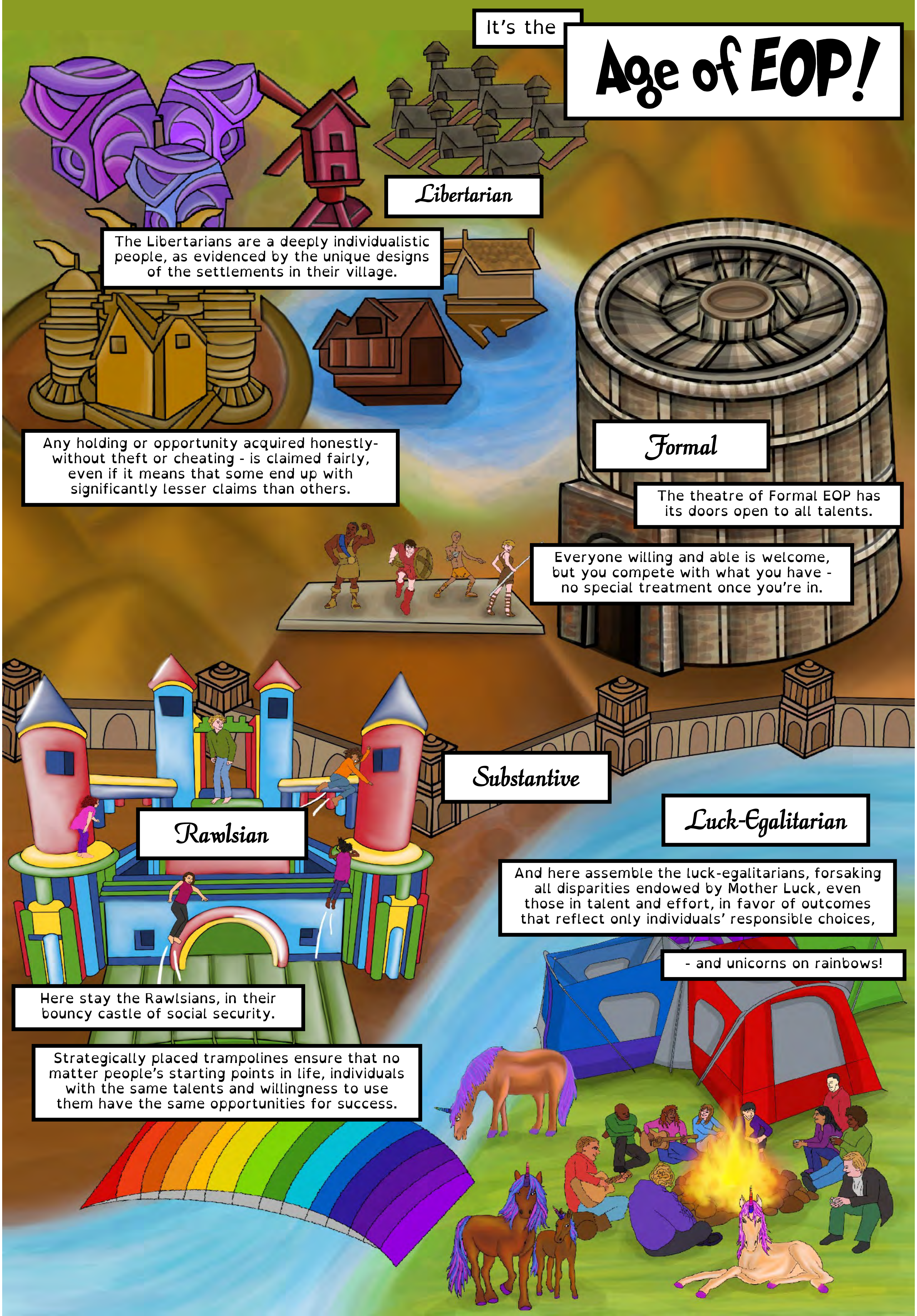
IS IT **NON-DISCRIMINATION** (FROM LEGAL DOCTRINES)?

IS IT **EQUALITY** IN THE DISTRIBUTION OF SOME COMMODITY/ OUTCOME (IN THE ECONOMIC SENSE)?

IS IT SOME NOTION OF **DISTRIBUTIVE JUSTICE** (FROM POLITICAL PHILOSOPHY)?

THE MOST INFLUENTIAL CHARACTER IN THE FAIR-ML MULTIVERSE SEEMS TO BE **FAIRNESS** AS **EQUALITY OF OPPORTUNITY (EOP).**

LET'S READ THROUGH ITS ORIGIN STORY, SHALL WE?


It's the
Age of EOP!
Libertarian
The Libertarians are a deeply individualistic people, as evidenced by the unique designs of the settlements in their village.
Any holding or opportunity acquired honestly- without theft or cheating - is claimed fairly, even if it means that some end up with significantly lesser claims than others.
Formal
The theatre of Formal EOP has its doors open to all talents.
Everyone willing and able is welcome, but you compete with what you have - no special treatment once you're in.
Substantive
Rawlsian
Luck-Egalitarian
And here assemble the luck-egalitarians, forsaking all disparities endowed by Mother Luck, even those in talent and effort, in favor of outcomes that reflect only individuals' responsible choices,
- and unicorns on rainbows!
Here stay the Rawlsians, in their bouncy castle of social security.
Strategically placed trampolines ensure that no matter people's starting points in life, individuals with the same talents and willingness to use them have the same opportunities for success.

Libertarian

THE LIBERTARIAN VIEW FOCUSES ON THE INDIVIDUAL’S FREEDOMS AND LIBERTIES.

LIKE IN A GAME OF MONOPOLY, PLAYERS ARE FREE TO CAPITALIZE ON WHATEVER OPPORTUNITIES THEY HAVE ACCESS TO - SUCH AS ROLLING DOUBLES AND GETTING TO MOVE TWICE, OR PICKING UP THAT CHANCE CARD THAT ADVANCES YOU TO BOARDWALK!

- PROVIDED THEY GAIN SUCH ACCESS FAIR AND SQUARE - NO CHEATING BY ROLLING BIASED DICE, STEALING FROM THE BANK OR FORCING PLAYERS TO TRADE PROPERTIES.

ALL PLAYERS ARE FREE TO DECIDE WHICH PROPERTY TO CHASE. WHETHER THEY ACTUALLY GET THE OPPORTUNITY TO BUY AND DEVELOP ON THAT SPOT IS NOT ENTIRELY DEVOID OF CHANCE, BUT THE GAME DOES NOT ATTEMPT TO CORRECT FOR IT.

INSTEAD, THE EMPHASIS IS ON RESPECT FOR PLAYERS’ LIBERTY TO BUY AND SELL PROPERTY AND THEIR FREEDOM TO EXERCISE THEIR INDIVIDUAL SKILLS OF NEGOTIATION AND DICE-THROWING.

THIS DOESN’T APPEAR TO BE A FORM OF EOP AT ALL: THERE’S NOTHING BEING EQUALIZED.

A LIBERTARIAN ADS IS ONLY CONCERNED ABOUT ENSURING A VERY LIMITED NOTION OF PROCEDURAL FAIRNESS.

Formal EOP

"CAREERS OPEN TO TALENTS"

FORMAL EOP SAYS A COMPETITION IS FAIR WHEN COMPETITORS ARE ONLY EVALUATED ON THE BASIS OF THEIR RELEVANT QUALIFICATIONS - IN ANY CONTEST, THE MOST QUALIFIED PERSON WINS.

THIS IS A VIEW THAT REJECTS HEREDITARY PRIVILEGE AS THE BASIS FOR WINNING POSITIONS: BEING AN ARISTOCRAT WON'T GET YOU THE JOB.

STILL, FORMAL EOP MAKES NO ATTEMPT TO CORRECT FOR ARBITRARY PRIVILEGES AND DISADVANTAGES THAT CAN LEAD TO DISPARITIES IN INDIVIDUALS' OPPORTUNITIES TO BUILD QUALIFICATIONS.

FORMAL EOP ADVOCATES **'SEE NOTHING IRRELEVANT, SPEAK NOTHING IRRELEVANT, HEAR NOTHING IRRELEVANT'.**

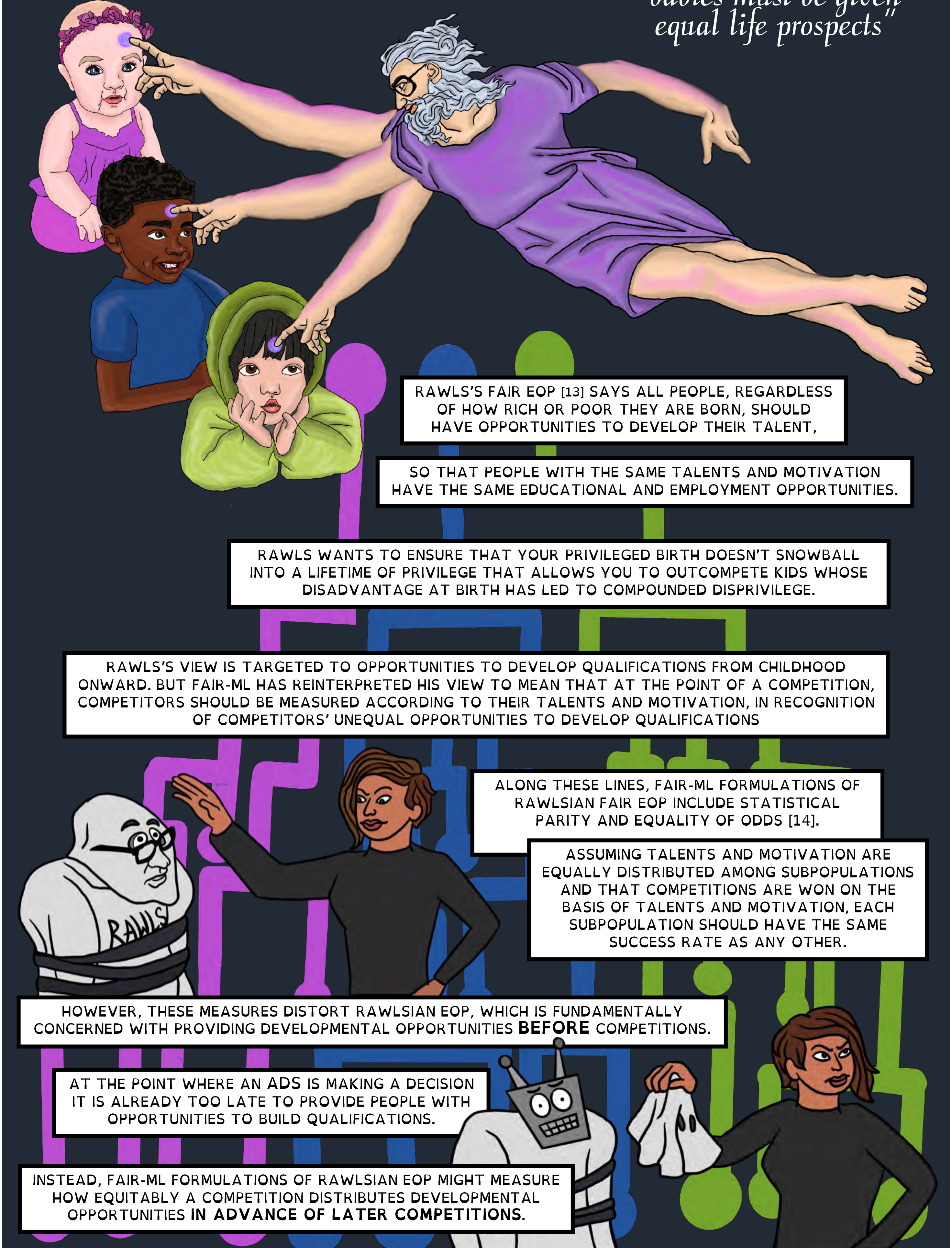
DECISION MAKERS ARE TAUGHT TO IGNORE IRRELEVANT TRAITS LIKE SOCIAL STATUS AND TO FOCUS ONLY ON RELEVANT QUALIFICATIONS IN ADJUDICATING A CONTEST

IN FAIR-ML, THIS HAS BEEN CODIFIED AS 'FAIRNESS THROUGH BLINDNESS', WHERE ANY PROTECTED ATTRIBUTES - THOSE THAT CAN IDENTIFY GROUP MEMBERSHIP - ARE STRIPPED AWAY FROM THE DATA.

BUT THERE'S MORE TO FORMAL EOP, IF WE CONSIDER ITS MOTIVATION. A TEST THAT IS MORE INACCURATE FOR MEMBERS OF A PROTECTED CLASS - THAT BADLY MISMEASURES THE QUALIFICATIONS OF WOMEN CANDIDATES COMPARED TO MEN, FOR EXAMPLE - ALSO VIOLATES THE SPIRIT OF FORMAL EOP, EVEN IF THE TEST DOES NOT TAKE GENDER INTO ACCOUNT. [12]

(Substantive)

Rawls' Fair EOP



"Equally talented babies must be given equal life prospects"

RAWLS'S FAIR EOP [13] SAYS ALL PEOPLE, REGARDLESS OF HOW RICH OR POOR THEY ARE BORN, SHOULD HAVE OPPORTUNITIES TO DEVELOP THEIR TALENT,

SO THAT PEOPLE WITH THE SAME TALENTS AND MOTIVATION HAVE THE SAME EDUCATIONAL AND EMPLOYMENT OPPORTUNITIES.

RAWLS WANTS TO ENSURE THAT YOUR PRIVILEGED BIRTH DOESN'T SNOWBALL INTO A LIFETIME OF PRIVILEGE THAT ALLOWS YOU TO OUTCOMPETE KIDS WHOSE DISADVANTAGE AT BIRTH HAS LED TO COMPOUNDED DISPRIVILEGE.

RAWLS'S VIEW IS TARGETED TO OPPORTUNITIES TO DEVELOP QUALIFICATIONS FROM CHILDHOOD ONWARD. BUT FAIR-ML HAS REINTERPRETED HIS VIEW TO MEAN THAT AT THE POINT OF A COMPETITION, COMPETITORS SHOULD BE MEASURED ACCORDING TO THEIR TALENTS AND MOTIVATION, IN RECOGNITION OF COMPETITORS' UNEQUAL OPPORTUNITIES TO DEVELOP QUALIFICATIONS

ALONG THESE LINES, FAIR-ML FORMULATIONS OF RAWLSIAN FAIR EOP INCLUDE STATISTICAL PARITY AND EQUALITY OF ODDS [14].

ASSUMING TALENTS AND MOTIVATION ARE EQUALLY DISTRIBUTED AMONG SUBPOPULATIONS AND THAT COMPETITIONS ARE WON ON THE BASIS OF TALENTS AND MOTIVATION, EACH SUBPOPULATION SHOULD HAVE THE SAME SUCCESS RATE AS ANY OTHER.

HOWEVER, THESE MEASURES DISTORT RAWLSIAN EOP, WHICH IS FUNDAMENTALLY CONCERNED WITH PROVIDING DEVELOPMENTAL OPPORTUNITIES **BEFORE** COMPETITIONS.

AT THE POINT WHERE AN ADS IS MAKING A DECISION IT IS ALREADY TOO LATE TO PROVIDE PEOPLE WITH OPPORTUNITIES TO BUILD QUALIFICATIONS.

INSTEAD, FAIR-ML FORMULATIONS OF RAWLSIAN EOP MIGHT MEASURE HOW EQUITABLY A COMPETITION DISTRIBUTES DEVELOPMENTAL OPPORTUNITIES **IN ADVANCE OF LATER COMPETITIONS.**

(Substantive)

Luck-Egalitarian EOP

"Nothing that you did not choose for yourself should affect your life prospects"

THE LUCK EGALITARIAN SAYS THAT RAWLS DOESN'T GO FAR ENOUGH IN CONTROLLING FOR FACTORS THAT PROVIDE UNFAIR ADVANTAGE OR DISADVANTAGE.

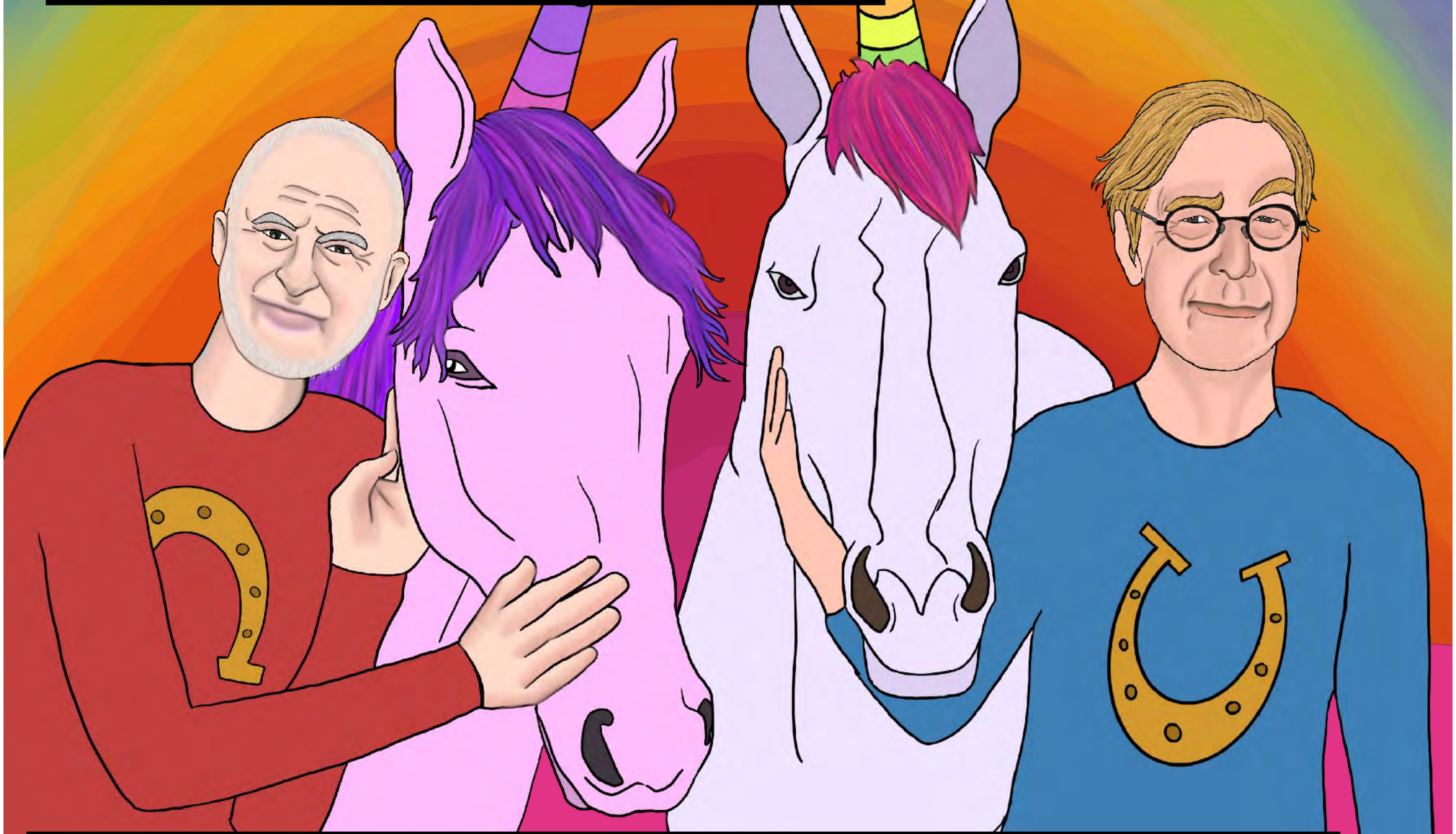
OUR OUTCOMES SHOULD ONLY BE AFFECTED BY OUR "CHOICE LUCK" (RESPONSIBLE CHOICES); NO EFFECTS OF "BRUTE LUCK" (FROM HAVING RICH PARENTS TO GETTING STRUCK BY LIGHTNING) SHOULD BE ALLOWED TO STAND.

HOW DO WE SEPARATE THE EFFECTS OF **LUCK** FROM THE EFFECTS OF **RESPONSIBLE CHOICES?**

ONE POPULAR FORMULATION IN FAIR-ML IS **ROEMER'S EOP** [15], WHICH MEASURES A PERSON'S EFFORT COMPARED TO OTHERS IN SIMILAR CIRCUMSTANCES. [16]

THIS DIALS BACK ON THE IDEA OF CONTROLLING FOR ALL BRUTE LUCK. INSTEAD, WE FOCUS ON A FEW BRUTE LUCK FACTORS, SUCH AS RACE AND SEX, THAT TRACK SIGNIFICANT UNDESERVED PRIVILEGE AND DISPRIVILEGE AND AFFECT PEOPLE'S OPPORTUNITIES TO DEVELOP QUALIFICATIONS.

WE CREATE BRACKETS BASED ON MATTERS OF BRUTE LUCK AND THEN COMPARE CANDIDATES TO OTHERS IN THEIR OWN BRACKETS.



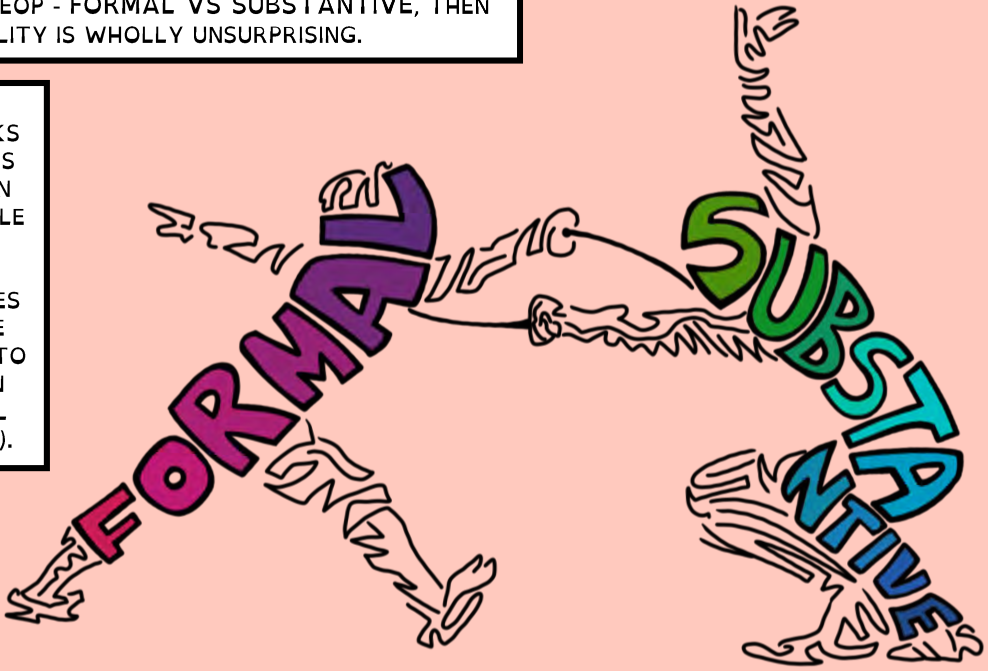
THIS FORMULATION IN FAIR-ML CAPTURES THE ESSENCE OF LUCK EGALITARIANISM AND IS APPEALING BECAUSE IT ALSO MEETS THE DECISION MAKER'S OBJECTIVE TO FIND QUALIFIED CANDIDATES

- THE ADS CONSIDERS ALL OF A CANDIDATE'S QUALIFICATIONS, NOT JUST THOSE THAT ARE ATTRIBUTABLE TO NATIVE TALENT/MOTIVATION (RAWLS) OR RESPONSIBLE CHOICES (OTHER LUCK EGALITARIANS)

THE IMPOSSIBILITY RESULTS IN FAIR-ML ARE COMMONLY INTERPRETED TO MEAN THAT 'FAIRNESS IS IMPOSSIBLE'.

BUT, IF WE LOOK AT DIFFERENT STATISTICAL MEASURES AS PROMOTING DIFFERENT CONCEPTIONS OF EOP - FORMAL VS SUBSTANTIVE, THEN THIS INCOMPATIBILITY IS WHOLLY UNSURPRISING.

WE WOULD NOT EXPECT A WORLD VIEW THAT ONLY LOOKS AT 'RELEVANT' QUALIFICATIONS AT THE POINT OF COMPETITION (FORMAL EOP) TO BE COMPATIBLE WITH ONE THAT AIMS TO PROVIDE COMPARABLE DEVELOPMENTAL OPPORTUNITIES FOR INDIVIDUALS AND, AT THE POINT OF COMPETITION, SEEKS TO CORRECT FOR INEQUALITIES IN CANDIDATES' DEVELOPMENTAL OPPORTUNITIES (SUBSTANTIVE).



WE CAN INTERPRET THIS INCOMPATIBILITY AS THE DIFFERENCE IN PHILOSOPHICAL VIEWPOINTS AND INCENTIVES OF DECISION MAKERS.

THIS GROUNDING GIVES US SOME MUCH-NEEDED GUIDANCE IN CHOOSING A SUITABLE 'FAIRNESS' MEASURE FOR OUR GIVEN CONTEXT.

IF WE BELIEVE THAT INEQUALITIES OF BIRTH DO NOT AFFECT A PERSON'S QUALIFICATIONS, THEN THE FORMAL APPROACH MIGHT BE SUFFICIENT TO MODEL A 'FAIR' FOOTRACE BETWEEN A KING AND A PEASANT.

FORMAL EOP ALSO OFFERS FAIRNESS IN THE FORM OF 'BLIND AUDITIONS'.

WHEN WE WORRY THAT JUDGES WILL BE SWAYED BY IRRELEVANT TRAITS LIKE GENDER, RACE AND APPEARANCE, BLIND AUDITIONS FORCE JUDGES TO EVALUATE CONTESTANTS SOLELY ON THEIR SINGING CHOPS.

SIMILARLY, MAKING EMPLOYERS BLIND TO JOB APPLICANTS' CREDIT SCORES OR CRIMINAL CONVICTIONS DURING INITIAL APPLICANT SCREENINGS CAN HELP PEOPLE OVERCOME STUBBORN OBSTACLES TO EMPLOYMENT!

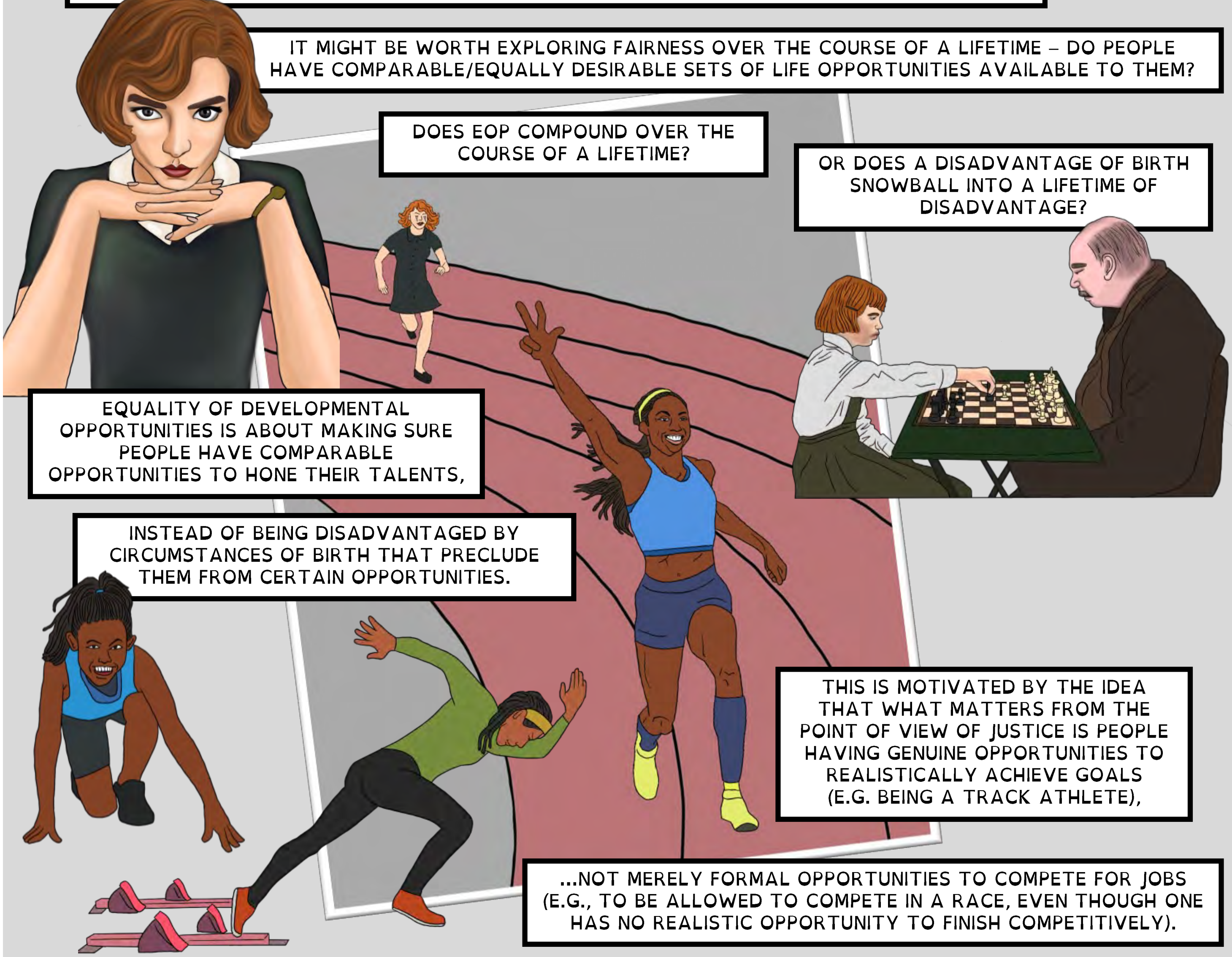
WHEN WE HAVE REASON TO BELIEVE THAT STRUCTURAL INEQUALITIES PRECLUDE SWATHS OF PEOPLE FROM DEVELOPING COMPETITIVE QUALIFICATIONS, WE MIGHT DECIDE TO MODEL A MORE SUBSTANTIVE CONCEPTION OF EOP.

IN A FOOTRACE, IF HURDLES – IN THE FORM OF SYSTEMIC DISCRIMINATION AND INEQUITABLE ACCESS – ABOUND IN THE PATH OF CERTAIN COMPETITORS,

WE MIGHT WANT TO COMPARE THE HURDLE JUMPERS WITH OTHER HURDLE JUMPERS AND THE SMOOTH-TRACK RUNNERS WITH SMOOTH-TRACK RUNNERS.

THE SUBSTANTIVE APPROACH IS DESIRABLE HERE IN ORDER TO ENSURE THAT WE DON'T END UP INFLICTING A SISYPHEAN STRUGGLE UPON CERTAIN CANDIDATES BY OVERLOOKING DISADVANTAGES THEY'VE HAD RELATIVE TO OTHER COMPETITORS.

ONE IMPORTANT IDEA FROM POLITICAL PHILOSOPHY THAT IS OVERLOOKED IN FAIR-ML IS THE DISTINCTION BETWEEN EQUALITY OF DEVELOPMENTAL OPPORTUNITIES, EOP OVER A LIFETIME, AND EOP AT A DECISION POINT (FAIR-ML'S FOCUS).



IT MIGHT BE WORTH EXPLORING FAIRNESS OVER THE COURSE OF A LIFETIME – DO PEOPLE HAVE COMPARABLE/EQUALLY DESIRABLE SETS OF LIFE OPPORTUNITIES AVAILABLE TO THEM?

DOES EOP COMPOUND OVER THE COURSE OF A LIFETIME?

OR DOES A DISADVANTAGE OF BIRTH SNOWBALL INTO A LIFETIME OF DISADVANTAGE?

EQUALITY OF DEVELOPMENTAL OPPORTUNITIES IS ABOUT MAKING SURE PEOPLE HAVE COMPARABLE OPPORTUNITIES TO HONE THEIR TALENTS,

INSTEAD OF BEING DISADVANTAGED BY CIRCUMSTANCES OF BIRTH THAT PRECLUDE THEM FROM CERTAIN OPPORTUNITIES.

THIS IS MOTIVATED BY THE IDEA THAT WHAT MATTERS FROM THE POINT OF VIEW OF JUSTICE IS PEOPLE HAVING GENUINE OPPORTUNITIES TO REALISTICALLY ACHIEVE GOALS (E.G. BEING A TRACK ATHLETE),

...NOT MERELY FORMAL OPPORTUNITIES TO COMPETE FOR JOBS (E.G., TO BE ALLOWED TO COMPETE IN A RACE, EVEN THOUGH ONE HAS NO REALISTIC OPPORTUNITY TO FINISH COMPETITIVELY).

OUR STROLL THROUGH EOP-VILLE HAS SHOWN US A RANGE OF INTERPRETATIONS OF 'FAIRNESS'. BUT IS 'FAIRNESS' ALL THAT'S REQUIRED FOR AN ALGORITHM TO BE 'JUST'?

RAWLS SANDWICHES HIS EOP PRINCIPLE BETWEEN TWO OTHER PRINCIPLES THAT ALSO MUST BE SATISFIED FOR A DEMOCRATIC SOCIETY TO BE 'JUST'.

HE ARRIVES AT THESE PRINCIPLES VIA **THE ORIGINAL POSITION**- A THOUGHT EXPERIMENT ABOUT HOW CITIZENS WOULD NEGOTIATE THE SET-UP OF SOCIETY, UNDER THE **'VEIL OF IGNORANCE'**

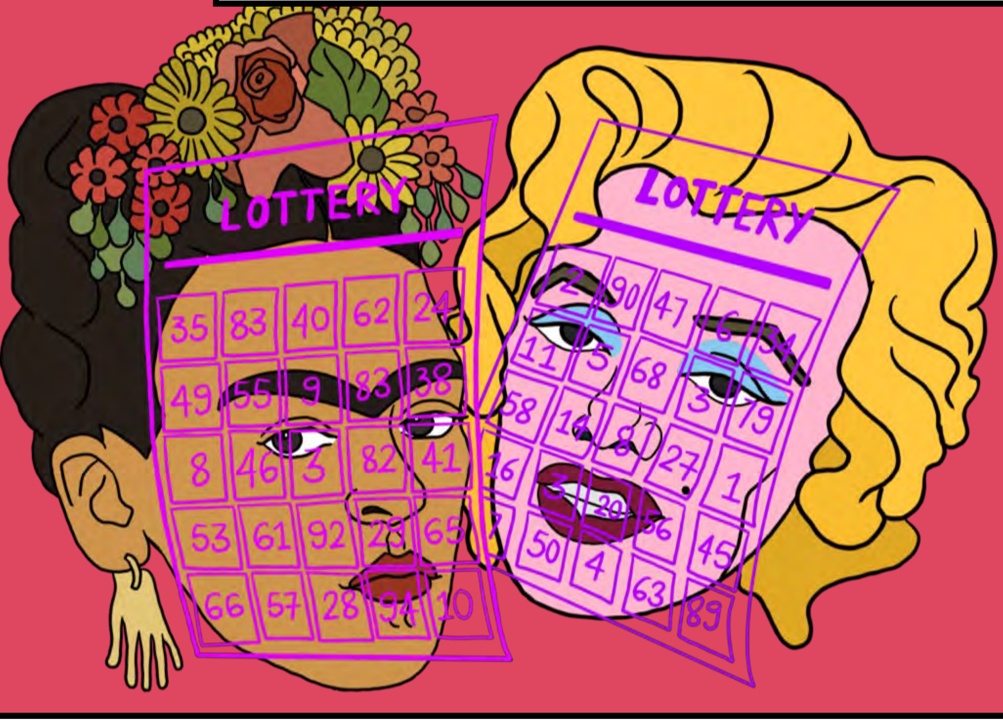
- IF CITIZENS DO NOT KNOW THEIR RACE, CLASS, SEX, TALENTS, SOCIAL POSITION (OR ANY OTHER CHARACTERISTICS THAT MIGHT CAUSE THEM TO FAVOR PEOPLE LIKE THEMSELVES), THEY WILL ADVOCATE FOR ALL SOCIAL POSITIONS AND THEIR ATTACHED PRIVILEGES TO BE DISTRIBUTED 'FAIRLY'.

BUT THEY DO KNOW THAT PEOPLE ARE FREE AND EQUAL AND THAT THEY HAVE THE ABILITY TO CHOOSE A CONCEPTION OF THE GOOD LIFE AND THE ABILITY TO ABIDE BY RULES OF JUSTICE.

AND SO, RAWLS POSITS THAT THE PRINCIPLES OF SOCIAL COOPERATION THAT PEOPLE ARRIVE AT THROUGH SUCH A NEGOTIATION WILL BE APPROPRIATE FOR A FREE AND DEMOCRATIC SOCIETY.

RAWLS USES THE NOTION OF THE **"NATURAL LOTTERY"** TO DESCRIBE THE MORALLY ARBITRARY DISTRIBUTION OF TALENTS, FAMILY CIRCUMSTANCES, AND OTHER AT-BIRTH FORTUNE AND MISFORTUNE TO PEOPLE.

FROM THE ARBITRARINESS OF THE NATURAL LOTTERY, RAWLS CONCLUDES THAT WE DON'T DESERVE OUR STARTING POINTS IN LIFE,



...AND ARRIVES AT THE **DIFFERENCE PRINCIPLE** - WHICH HARNESSES THE ARBITRARY DISTRIBUTION OF TALENTS TO GENERATE A SOCIAL SYSTEM THAT SERVES EVERYONE.

RAWLS' THEORY OF JUSTICE

POSITS THE FOLLOWING HIERARCHICAL PRINCIPLES: [13]

1. **[RIGHTS AND LIBERTIES]** EVERYONE HAS THE SAME INALIENABLE RIGHT TO EQUAL BASIC LIBERTIES

2. (a) **[RAWLSIAN FAIR EOP]** ALL OFFICES AND POSITIONS MUST BE OPEN TO ALL UNDER CONDITIONS OF FAIR EQUALITY OF OPPORTUNITY.

2. (b) **[DIFFERENCE PRINCIPLE]** SOCIAL AND ECONOMIC INEQUALITIES MUST BE OF THE GREATEST BENEFIT TO THE LEAST ADVANTAGED

IN THE RAWLSIAN SYSTEM, THESE PRINCIPLES ARE HIERARCHICALLY ORDERED –

FAIR EOP CAN'T BE SATISFIED AT THE EXPENSE OF CITIZENS' EQUAL BASIC RIGHTS AND LIBERTIES,

AND THE DIFFERENCE PRINCIPLE CAN'T BE SATISFIED AT THE EXPENSE OF EOP

- AKIN TO HOW INCREDIBLY COUNTERINTUITIVE IT WOULD BE TO PUT ON A BLAZER, WITHOUT WEARING A SHIRT FIRST!

FOR EXAMPLE, TAKE THE CHILDREN OF RICH PARENTS -

IN TRYING TO GIVE PEOPLE ACCESS TO EQUAL DEVELOPMENTAL OPPORTUNITIES, ONE MIGHT END UP PREVENTING PARENTS FROM RAISING KIDS ACCORDING TO THEIR VALUES,

BECAUSE THIS WOULD MEAN THAT SOME KIDS GET BETTER DEVELOPMENTAL OPPORTUNITIES THAT OTHERS.

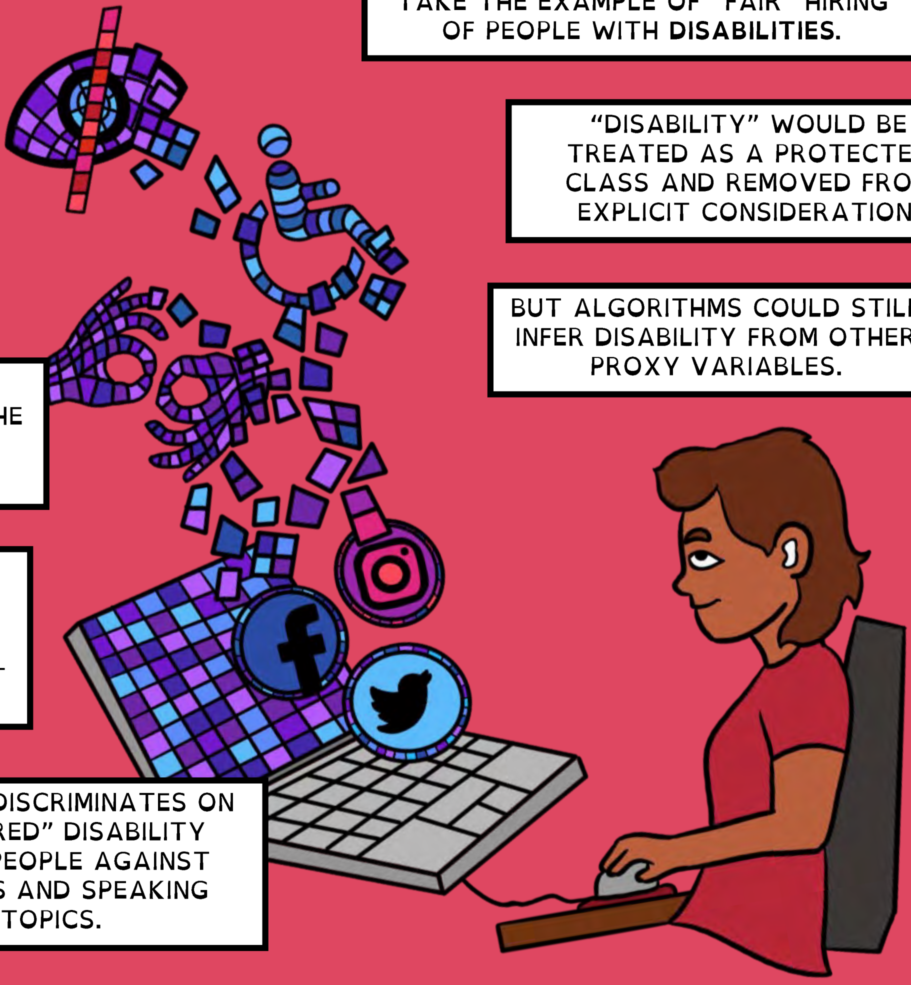
IN TRYING TO SATISFY RAWLS'S FAIR EOP, WE MIGHT END UP INFRINGING ON RICH PARENTS' BASIC LIBERTIES.

IN THE CONTEXT OF ALGORITHMS, THIS BROADER PERSPECTIVE IS HELPFUL TO SEE HOW AN ADS THAT IS (STATISTICALLY) 'FAIR' CAN GO ON TO INFRINGE ON BASIC RIGHTS AND LIBERTIES AND, IN EFFECT, BE UNJUST.

TAKE THE EXAMPLE OF "FAIR" HIRING OF PEOPLE WITH **DISABILITIES.**

"DISABILITY" WOULD BE TREATED AS A PROTECTED CLASS AND REMOVED FROM EXPLICIT CONSIDERATION,

BUT ALGORITHMS COULD STILL INFER DISABILITY FROM OTHER PROXY VARIABLES.



IF SOCIAL MEDIA INFORMATION IS USED, THE ADS COULD INFER DISABILITY STATUS—

FOR EXAMPLE, BASED ON MEMBERSHIP IN CERTAIN SOCIAL GROUPS OR ON POSTING ABOUT DISABILITY-RELATED ISSUES—

THEN A SCHEME THAT DISCRIMINATES ON THE BASIS OF "INFERRED" DISABILITY WOULD INCENTIVIZE PEOPLE AGAINST JOINING SUCH GROUPS AND SPEAKING ABOUT SUCH TOPICS.

SUCH AN ADS COULD SATISFY SOME CONCEPTION OF 'FAIRNESS' AS EOP AND YET BE FUNDAMENTALLY UNJUST: IT WOULD VIOLATE A CANDIDATE'S FREEDOM OF SPEECH AND FREEDOM OF ASSOCIATION.

THERE ARE LIMITATIONS TO WHAT ANSWERS WE CAN GET FROM EOP DOCTRINES,

AND OVERLOOKING THESE CAN EMBOLDEN THEIR APPLICATION IN SPHERES IN WHICH THEORY PROVIDES LITTLE TO NO GUIDANCE...

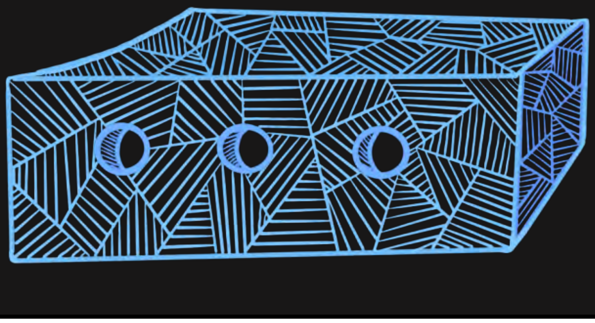
THESE DOCTRINES DO NOT GIVE US ANY DIRECTION ABOUT *WHERE* TO APPLY 'FAIRNESS' - IN THE PROCEDURE OR AT THE OUTCOME.

THE GUIDANCE IS ONLY ABOUT *HOW* A 'FAIR' TEST SHOULD BEHAVE.

WHEN APPLYING THIS TEST TO BLACK BOX ADS, WE RUN INTO ISSUES OF INTERPRETABILITY

AND CAN ONLY INFER DETAILS ABOUT HOW THE TEST IS BEHAVING BY LOOKING AT WHICH INPUTS HAVE BEEN FED INTO THE ALGORITHM,

OR BY SYSTEMATICALLY STUDYING THE OUTCOMES FOR A VARIETY OF CANDIDATES.



THE FAIRNESS YOU ASKED FOR IS INSIDE THIS BOX!

SUBSTANTIVE EOP SEEKS TO PROVIDE ALL INDIVIDUALS WITH REALISTIC OPPORTUNITIES TO DEVELOP QUALIFICATIONS AND HENCE A REALISTIC SHOT AT COMPETING FOR A BROAD RANGE OF POSITIONS.

IF WE DECIDE THAT THE ONLY WAY THAT WE CAN OPERATIONALIZE THE SUBSTANTIVE VIEW IS TO SEPARATE QUALIFICATIONS INTO MATTERS OF CIRCUMSTANCE (TO BE CONTROLLED FOR) AND EFFORT (THAT THE INDIVIDUAL CAN BE HELD ACCOUNTABLE FOR), THEN WE MUST DECIDE HOW TO MAKE THIS SEPARATION!

WHICH OUTCOMES CAN WE HOLD ONE ACCOUNTABLE FOR?

AND WHICH MATTERS ARE COMPLETELY OUT OF THEIR CONTROL?

SOUNDS LIKE WHAT WE NEED IS A SORTING HAT!

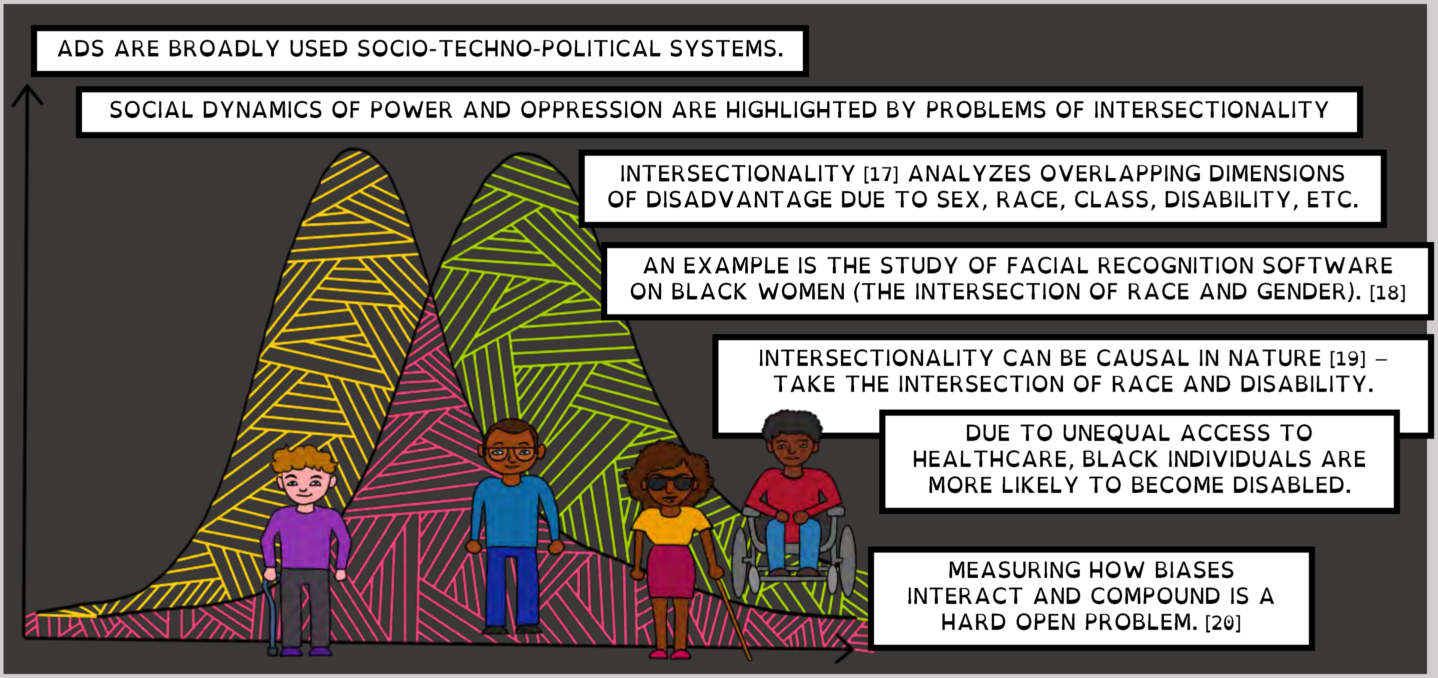
FROM A PRACTICAL PERSPECTIVE, IT IS OBVIOUS THAT WE CANNOT SEPARATE A PERSON'S NATIVE TALENTS FROM THEIR CIRCUMSTANCES, OR THEIR RESPONSIBLE CHOICES FROM THEIR BRUTE LUCK.

AND YET, WE HAVE TESTS LIKE THE SAT AND GRE, WHICH SUPPOSEDLY GAUGE INTELLIGENCE AND ACADEMIC EXCELLENCE, AND ARE USED TO MAKE UNIVERSITY ADMISSIONS DECISIONS.

JUST REGISTERING FOR SUCH A TEST - FORGET ABOUT GETTING ACCESS TO STUDY MATERIALS - IS PROHIBITIVELY EXPENSIVE.

SUCH STANDARDIZED TESTS DO NOT EVALUATE NATIVE TALENT, BUT INSTEAD DISCRIMINATE ON THE BASIS OF SOCIAL ADVANTAGES AND DISADVANTAGES.

ALWAYS.



ADS ARE BROADLY USED SOCIO-TECHNO-POLITICAL SYSTEMS.

SOCIAL DYNAMICS OF POWER AND OPPRESSION ARE HIGHLIGHTED BY PROBLEMS OF INTERSECTIONALITY

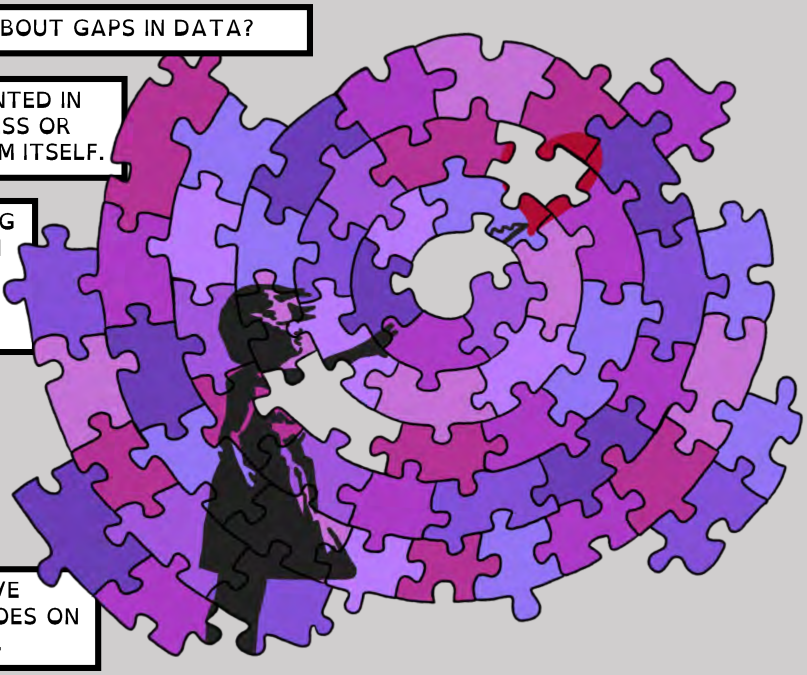
INTERSECTIONALITY [17] ANALYZES OVERLAPPING DIMENSIONS OF DISADVANTAGE DUE TO SEX, RACE, CLASS, DISABILITY, ETC.

AN EXAMPLE IS THE STUDY OF FACIAL RECOGNITION SOFTWARE ON BLACK WOMEN (THE INTERSECTION OF RACE AND GENDER). [18]

INTERSECTIONALITY CAN BE CAUSAL IN NATURE [19] – TAKE THE INTERSECTION OF RACE AND DISABILITY.

DUE TO UNEQUAL ACCESS TO HEALTHCARE, BLACK INDIVIDUALS ARE MORE LIKELY TO BECOME DISABLED.

MEASURING HOW BIASES INTERACT AND COMPOUND IS A HARD OPEN PROBLEM. [20]



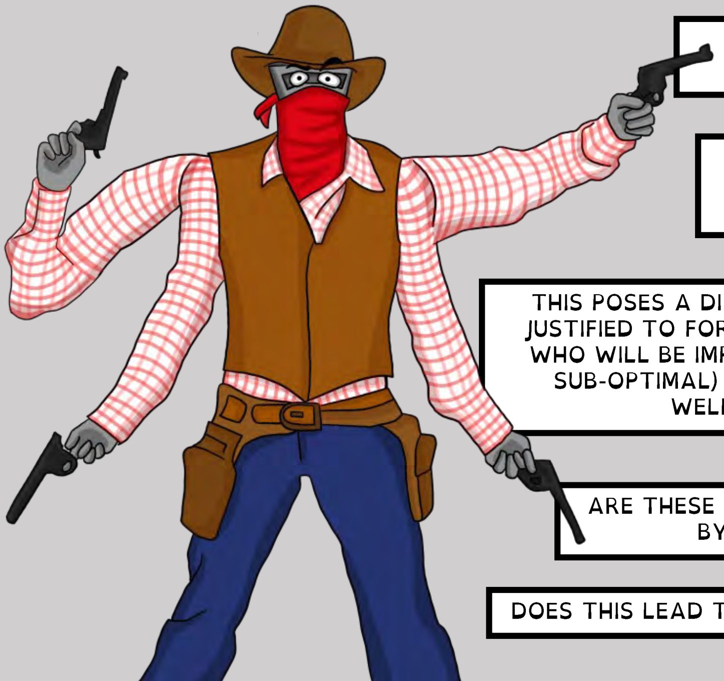
WHAT DO WE DO ABOUT GAPS IN DATA?

MANY DEMOGRAPHICS ARE POORLY REPRESENTED IN DATA, DUE TO ISSUES OF INEQUITABLE ACCESS OR DISTRUST IN THE DATA COLLECTION MECHANISM ITSELF.

DATA GAPS MAKE IT HARD TO VIEW THE BIG PICTURE AND MANIFEST AS A DISPARITY IN MODEL PERFORMANCE (ERROR RATES, FALSE POSITIVES, FALSE NEGATIVES) FOR UNDER-REPRESENTED DEMOGRAPHICS.

THEN THERE'S THE PROBLEM OF OBSERVABILITY [20]. IN MOST 'FAIRNESS' RELATED TASKS WE ARE MODELLING FOR 'RISKS' - 'RISK OF LOAN DEFAULT', 'RISK OF RECIDIVISM', 'RISK OF COLLEGE DROPOUT'.

SELDOM DO WE GET TO OBSERVE WHETHER THE PERSON ACTUALLY GOES ON TO DO ANY OF THOSE THINGS.



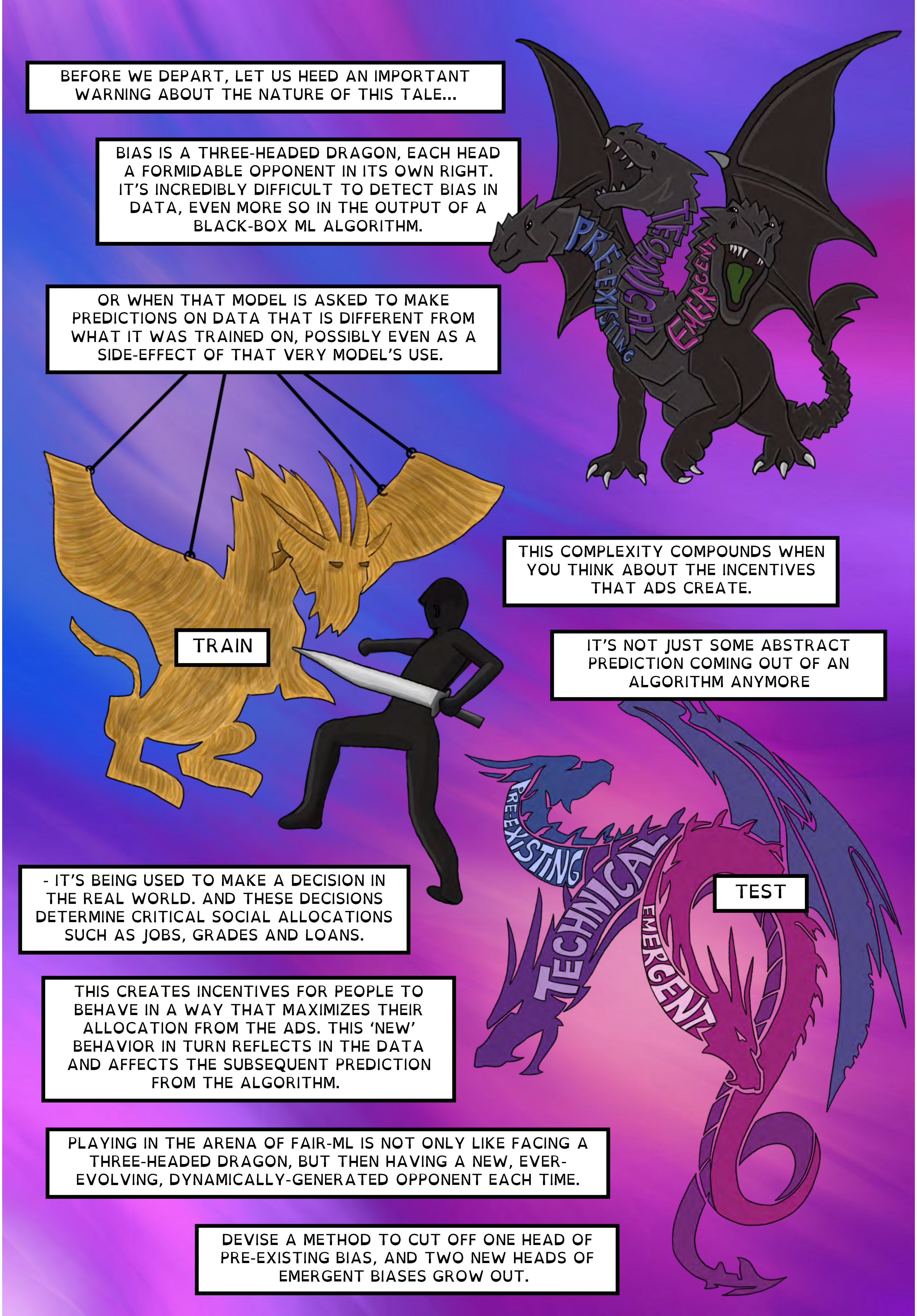
THIS LEADS US TO THE TRADEOFF BETWEEN EXPLORATION AND EXPLOITATION.

IN ORDER TO TEST THE EFFICACY OF AN ADS WE MIGHT NEED TO PUT IT INTO THE REAL WORLD TO GATHER MORE DATA.

THIS POSES A DIFFICULT ETHICAL CONUNDRUM- IS IT JUSTIFIED TO FORGO THE WELLBEING OF INDIVIDUALS WHO WILL BE IMPACTED BY THE CURRENT (PERHAPS SUB-OPTIMAL) ADS FOR THE POTENTIAL FUTURE WELLBEING OF INDIVIDUALS?

ARE THESE COSTS BORNE DISPROPORTIONATELY BY A CERTAIN DEMOGRAPHIC?

DOES THIS LEAD TO NEW FORMS OF 'UNFAIRNESS'? [21]



BEFORE WE DEPART, LET US HEED AN IMPORTANT WARNING ABOUT THE NATURE OF THIS TALE...

BIAS IS A THREE-HEADED DRAGON, EACH HEAD A FORMIDABLE OPPONENT IN ITS OWN RIGHT. IT'S INCREDIBLY DIFFICULT TO DETECT BIAS IN DATA, EVEN MORE SO IN THE OUTPUT OF A BLACK-BOX ML ALGORITHM.

OR WHEN THAT MODEL IS ASKED TO MAKE PREDICTIONS ON DATA THAT IS DIFFERENT FROM WHAT IT WAS TRAINED ON, POSSIBLY EVEN AS A SIDE-EFFECT OF THAT VERY MODEL'S USE.

THIS COMPLEXITY COMPOUNDS WHEN YOU THINK ABOUT THE INCENTIVES THAT ADS CREATE.

IT'S NOT JUST SOME ABSTRACT PREDICTION COMING OUT OF AN ALGORITHM ANYMORE

- IT'S BEING USED TO MAKE A DECISION IN THE REAL WORLD. AND THESE DECISIONS DETERMINE CRITICAL SOCIAL ALLOCATIONS SUCH AS JOBS, GRADES AND LOANS.

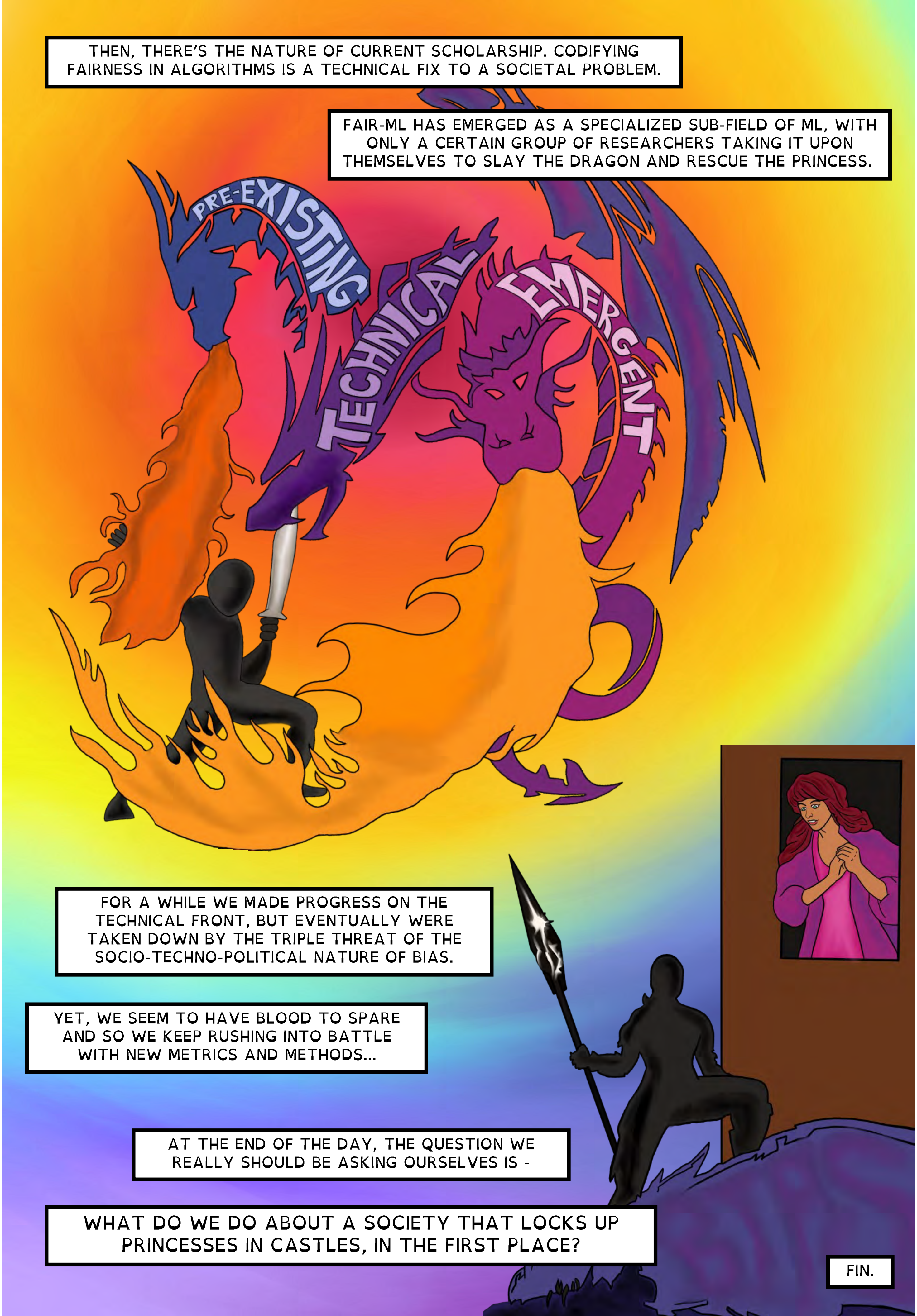
THIS CREATES INCENTIVES FOR PEOPLE TO BEHAVE IN A WAY THAT MAXIMIZES THEIR ALLOCATION FROM THE ADS. THIS 'NEW' BEHAVIOR IN TURN REFLECTS IN THE DATA AND AFFECTS THE SUBSEQUENT PREDICTION FROM THE ALGORITHM.

PLAYING IN THE ARENA OF FAIR-ML IS NOT ONLY LIKE FACING A THREE-HEADED DRAGON, BUT THEN HAVING A NEW, EVER-EVOLVING, DYNAMICALLY-GENERATED OPPONENT EACH TIME.

DEVISE A METHOD TO CUT OFF ONE HEAD OF PRE-EXISTING BIAS, AND TWO NEW HEADS OF EMERGENT BIASES GROW OUT.

THEN, THERE'S THE NATURE OF CURRENT SCHOLARSHIP. CODIFYING FAIRNESS IN ALGORITHMS IS A TECHNICAL FIX TO A SOCIETAL PROBLEM.

FAIR-ML HAS EMERGED AS A SPECIALIZED SUB-FIELD OF ML, WITH ONLY A CERTAIN GROUP OF RESEARCHERS TAKING IT UPON THEMSELVES TO SLAY THE DRAGON AND RESCUE THE PRINCESS.



FOR A WHILE WE MADE PROGRESS ON THE TECHNICAL FRONT, BUT EVENTUALLY WERE TAKEN DOWN BY THE TRIPLE THREAT OF THE SOCIO-TECHNO-POLITICAL NATURE OF BIAS.

YET, WE SEEM TO HAVE BLOOD TO SPARE AND SO WE KEEP RUSHING INTO BATTLE WITH NEW METRICS AND METHODS...

AT THE END OF THE DAY, THE QUESTION WE REALLY SHOULD BE ASKING OURSELVES IS -

WHAT DO WE DO ABOUT A SOCIETY THAT LOCKS UP PRINCESSES IN CASTLES, IN THE FIRST PLACE?

FIN.

ABOUT

A computer scientist, artist and philosopher join a zoom room. This happens!

'Fairness and Friends' is the second volume of the Data, Responsibly Comic series. We hope that it will serve as the computer scientist's guide to political philosophy!

Falaah Arif Khan,
Co-Creator, Author, Artist

Falaah is a scientist/engineer by training and an artist by nature, and the creator of MachineLearnist Comics - a collection of webcomics about the current AI landscape.

Eleni Manis,
Author

Eleni is the Research Director at the Surveillance Technology Oversight Project. She began her career as an Assistant Professor of Philosophy at Franklin & Marshall College, focused on justice in democracies, and now works at the intersection of her expertise in ethics, democratic justice, and technology policy.

Julia Stoyanovich,
Co-Creator, Author

Julia is an Assistant Professor of Computer Science and Engineering and of Data Science and the founding Director of the Center for Responsible AI at New York University. She leads the 'Data, Responsibly' project, the latest offering of which is the inimitable interdisciplinary course on Responsible Data Science.

REFERENCES

1. https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/
2. https://www1.nyc.gov/assets/buildings/local_laws/ll49of2019.pdf
3. New York City Automated Decision Systems Task Force Report.
4. Julia Stoyanovich, Bill Howe, and H. V. Jagadish. (2020). Responsible data management. Proc. VLDB Endow. 13, 12 (August 2020)
5. Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14, 3 (July 1996)
6. Falaah Arif Khan and Julia Stoyanovich. "Mirror, Mirror". Data, Responsibly Comics, Volume 1 (2020)
7. Sorelle A. Friedler and Carlos Scheidegger and Suresh Venkatasubramanian. (2016). On the (im)possibility of fairness
8. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. (2012). Fairness through awareness. ITCS 2012
9. Reuben Binns. (2020). On the apparent conflict between individual and group fairness. FAT* 2020
10. Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, (2017).
11. Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. (2016)
12. Fishkin, Joseph. Bottlenecks: A New Theory of Equal Opportunity. Oxford: Oxford University Press, 2014.
13. John Rawls. A theory of justice. Harvard University Press, (1971)
14. Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. NeurIPS 2016.
15. John. E. Roemer. Equality of opportunity: a progress report. Social Choice and Welfare. (2002)
16. Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. FAT* (2019)
17. Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum (1989).
18. Joy Buolamwini and Timnit Gebru. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT* 2018
19. Ke Yang, Joshua Loftus & Julia Stoyanovich (2020). Causal intersectionality for fair ranking.
20. Alexandra Chouldechova and Aaron Roth (2020) A snapshot of the frontiers of fairness in machine learning. CACM 63, 5 (May 2020)
21. Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz and Hanna Wallach. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In Proceedings of Workshop on Fairness, Accountability, and Transparency in Machine Learning. ACM, 2016.