An Invitation to System-Wide Algorithmic Fairness

Efrén Cruz Cortés*
efren.cruzcortes@gmail.com
Pennsylvania State University
State College, Pennsylvania

Debashis Ghosh
debashis.ghosh@cuanschutz.edu
University of Colorado, Anschutz Medical Campus
Aurora, Colorado

ABSTRACT

We propose a framework for analyzing and evaluating system-wide algorithmic fairness. The core idea is to use simulation techniques in order to extend the scope of current fairness assessments by incorporating context and feedback to a phenomenon of interest. By doing so, we expect to better understand the interaction among the social behavior giving rise to discrimination, automated decision making tools, and fairness-inspired statistical constraints. In particular, we invite the community to use agent based models as an explanatory tool for causal mechanisms of population level properties. We also propose embedding these into a reinforcement learning algorithm to find optimal actions for meaningful change. As an incentive for taking a system-wide approach , we show through a simple model of predictive policing and trials that if we limit our attention to one portion of the system, we may determine some blatantly unfair practices as fair, and be blind to overall unfairness.

CCS CONCEPTS

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

KEYWORDS

fairness, ethical ai, agent based modeling, recidivism

ACM Reference Format:
Efrén Cruz Cortés and Debashis Ghosh. 2020. An Invitation to System-Wide Algorithmic Fairness. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375860

1 INTRODUCTION

Machine learning is proving itself beneficial to society through advances in medicine, public health, climate change and poverty research, as well as many other disciplines in the natural and social sciences. However, many governmental institutions and private companies have bought into the promise that machine learning can solve all problems with enough data. As such, there is a trend towards automation of important decision-making procedures that affect the lives of individuals and communities. While the desire for automation may be driven out of pure profit interests, as is the case with insurance and advertising companies, governmental institutions and the criminal justice system adopt the use of algorithms in their wish to improve societal conditions. Paradoxically, implementation of such algorithms has proven to be detrimental in many cases. Particularly, the strong dependence on historical data is a major cause for the reproduction and reinforcement of undesirable patterns. The data used for training is already biased due to historic discrimination and other structural deficiencies, which leads to biased machine learning algorithms. See [29], [12] for further discussion of this point.

Sadly, instances of discrimination occurring due to the applications of algorithmic tools by public and private institutions, have mostly gone unchecked until recently. However, in the last few years much preliminary research has been done on determining statistical properties of such algorithms and the data they are trained on. A major area of focus has been on defining measures of fairness and designing algorithms to satisfy such measures. Such research has allowed many notions of fairness to be quantified into statistical properties, see for example, [4] and [28]. In general, not all statistical quantifications of fairness are compatible (e.g. [6],[26]), and for now the best practice is to choose the most suitable one to the problem at hand knowing possible drawbacks in other directions.

Since much of this research has been developed to counteract the harmful effects of industry's use of algorithmic decision making, it disproportionately focuses on problem where products and not processes are the core of analysis, leading to 'tail-end' solutions, a posteriori patches to what is at best a dubious practice. In a sense, industry's practices have deviated research into curing the symptom as opposed to the disease.

We aim to complement that approach by studying the social dynamics in which these algorithms are implemented. Given the large-scale transformation these new technologies elicit, a joint effort of social sciences and machine learning researchers is necessary. Our approach is to implement a system-view paradigm that can account for context, feedback effects, and structural deficiencies in the society using AI tools. We start in a modest manner by taking a simulation strategy and exemplifying the amenability of its analysis through a simple case. In particular, we develop a very simple model of an arrest-recidivism system. We will see that even if the algorithmic tool itself satisfies certain notions of fairness, the system may fail to do so. The use of agent based models (ABMs) helps us realize that phenomenon as well as to find the reason behind it.

We briefly provide the rationale and structure for a system-wide analysis framework in Section 2, introduce a simple example model

*A substantial portion of the paper was written while at University of Colorado, Anschutz Medical Campus

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375860

in Section 3 and analyze it in Section 4. We conclude with some future research directions in Section 5.

2 A SYSTEM-WIDE ANALYSIS FRAMEWORK

While fairness and justice have served as catalysts for the new wave of sound research regarding algorithm accountability, the fundamental structures and dynamics in which these same fairness and justice concepts are reified get lost in the background. In the majority of previous studies, authors have focused on statistical properties of the algorithms being implemented, acknowledging the inherent problem in the data and data collection strategies but not always consolidating such processes as part of the system of analysis. This concern is particularly relevant for policy makers, since they ideally will rely on our expertise to take influential decisions.

Sadly, utilitarian logic has percolated into fairness research in machine learning, tricking us into thinking that given the use of an algorithm by a certain institution, our task is to maximize some type of "utility function", while keeping the procedure marginally fair. The main danger of this mode of thought lies in us believing that algorithms are neutral tools without a social context, and that improving their use and bettering society means keeping their utility maximum. Notwithstanding the efforts to make algorithms fair, many of these algorithms are in themselves harmful weapons. For example, recidivism prediction tools are counterfactually punitive, working on a double illusion: that a particular individual's characteristics are those of the statistical aggregate of their inputted population, and that punishment based on predictions of human actions are in any way philosophically founded. This might be an obvious example, but more furtive algorithms are out there supporting dubious practices, as it is the case, for example, with targeted advertisement of junk food to children, where the problem is not how fair it is but that it exists at all. The same goes for financial and economic research, where relations of exploitation are not questioned by algorithm makers, focusing only on maximizing profit.

It is not controversial anymore to say technology reformulates social dynamics, making our task and responsibility as machine learning researchers even more delicate and demanding. Technologies and infrastructures are intertwined with and often etiological to social relations. The internet and cellular technologies, for example, as enhancements of previous communication and information devices have transformed modern economies, warfare, surveillance, modes of learning and production, as well as human health and conceptions of the self. A more fundamental example is the historical transformations of family and gender relations catalyzed by material conditions ([25], [13]). As already eloquently put by a well-known economist (pardon the use of "man" as a universal):

> Technology reveals the active relation of man to nature, the direct process of the production of his life, and thereby it also lays bare the process of the production of the social relations of his life, and of the mental conceptions that flow from those relations.

[24]. Hence, we believe a thorough examination of social and artificial mechanisms is necessary for a true transformation into fair and ethical societies.

Three principles guide this paper: systemic analysis, causal relations and mechanisms, and optimal interventions. **Systemic analysis** refers to setting the algorithm into a context. This implies analyzing the data generating process, the decision making stage, and its consequences all under the same framework. Most previous research does set its analysis into context and warn of its consequences, however we strive to analyze these portions of the system together with the algorithm itself, subject to the same quantification, processing, and statistical measures. This principle provides a structure to the system and an engine for its overall dynamics. **Causal relations and mechanisms** are imperative in our understanding of how (un)fairness arises in a particular system. Note the difference between causal relations and causal mechanisms. A causal relation indicates what variable has a causal effect on another variable, a causal mechanism explains how this effect is produced. Causal relations are discovered under causal inference frameworks in statistics; in this paper we use ABMs as resources for causal mechanism discovery. Finally, although intervention effects are part of causal analysis, it is not always straightforward to find **optimal interventions** when the system under consideration is too complex or computationally demanding, so we give them their own special stage. In this paper we take advantage of the nature of ABM simulations as counterfactuals and embed our ABM in a reinforcement learning framework in order to find optimal policies, or parameter combinations, through a principled process.

We propose a system-wide framework based on a reinforcement learning paradigm. Under this scenario, the agent will be the policy maker and the environment a conglomerate of social, institutional, and technological mechanisms, as shown in Figure 1. In the model, the three mechanisms inside the dashed stage interact in any general way, and may do so for a while before an output is observed. Notice we have explicitly marked in the schematic a "fairness assessment" stage, as different definitions of fairness may yield different equilibrium points of the system. Indeed, most previous research on "fair machine learning" regards this stage of the process.

Take as an example the social phenomenon of "street crime" and the current institutional actions around it. Two particular mechanisms have been of interest in the literature: police surveillance and bail grants. In the schematic the social mechanism would be all social aspects which relate to crime, the institutional mechanism pertains to police surveillance and arrests, as well as the eventual judicial process, and the AI mechanism would be any tool involved in relevant decision making, for example predictive policing and recidivism assessment algorithms. Notice this particular case is that of "street" crime and not crime in general, since more harmful forms of potentially crime, like environmental, corporate, and war crimes, are policed in different ways, if policed at all.

Previous work has studied some of the relations inside the environment stage. The work in [9] and [10] treats the system of predictive policing as a feedback loop. Notice in our schematic that this inner loop is different from the outer loop, which allows for policy interventions by the global policy maker agent. The article [8] considers fairness when looking at a system composed of many parts, and arrives at a similar conclusion than this paper that parts being fair do not yield a system being so.

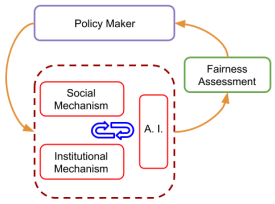


Figure 1: System Loop

2.1 ABM rationale

Collecting data from and intervening on social systems is an expensive and time-consuming process; in many situations it is impossible due to ethical or logistic constraints. Computational models provide a feasible avenue to understand social systems provided they exhibit similar characteristics. ABMs are a successful simulation paradigm in that they can generate a plethora of social dynamics and allow for policy experimentation. ABMs differ from other models, for example differential equation-based models, in that they admit more flexible heterogeneity in their populations (see [32]). Each agent has its own characteristics and often acts according to its local environment and imperfect information. This behavior is more realistic than many classical models of human behavior where perfect knowledge and rationality were assumed, as well as a homogeneity in the population.

ABMs are particularly at identifying causal relations by providing simple generative mechanisms to an observed phenomenon and since each simulation run can be interpreted as a counterfactual ([21], [23]). A general motto for agent-based modeling is "simpler is better", as the distillation of complex dynamics into smaller parts provides better insight into the system. In a sense, these explanatory mechanisms can be thought of as the simplest dynamics necessary to generate our observation. Occam-like induction philosophies assert these simple mechanisms are the most probable actual causes, see for example [36].

A general overview of the usefulness and history of agent-based models in the social sciences can be found in [22]. The book [11] contains numerous examples of ABMs in the social sciences, their design, and study. Finally, [40] is a step by step introduction to the design of ABMs through the popular software Netlogo.

3 AN EXAMPLE MODEL IN POLICING

A notable and now-classic example of particular interest to policy makers is the unchecked implementation of machine learning algorithms in the criminal justice system. Our case example will consider both recidivism prediction and predictive policing.

Regarding recidivism prediction, the 2016 report by ProPublica [3], [19] studied one such algorithm used in Broward County, Florida, which intended to predict recidivism among defendants up for parole. The report found out the algorithm was reproducing systematic bias against people of color and in favor of whites, violating certain statistical notions of fairness. The creators of the algorithm, however, argued their algorithm was fair since it satisfied equal positive predictive values, [6] showed an inherent mathematical trade-off among these fairness measures when considering heterogenous groups. In [7] COMPAS is shown to be as accurate and fair as a simple linear model and an aggregate of people with little criminal justice expertise, rebutting the belief that these prediction tools, if imperfect, are fairer and more accurate than decisions taken by humans.

The work in [20] shows that predictive policing algorithms reinforce historical police activity, resulting in a suboptimal process through which the algorithm is incapable of accurately estimating crime rates. A survey of crime forecasting procedures can be found in [5]. A survey on statistical notions of fairness for algorithmic tools in the criminal system is given in [4]. For a civil rights perspective on the problem, see [16]. The recent book [14] provides a comprehensive overview of tools, interventions, effects, and context of algorithmic use in policing and surveillance. Finally, [15] provides a framework to compare different proposed fairness metrics.

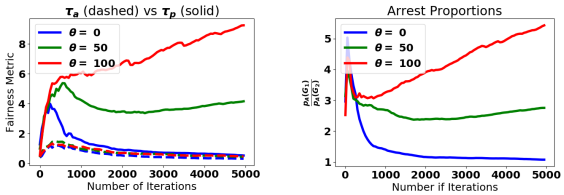
To understand the arrest-sentence system of interest we start as simple as possible. We model each member of the population as an independent agent that randomly moves in a confined region with demographically similar agents. Besides their membership to one of two groups, their crime and recidivism rates, agents do not have other characteristics. We also have a few cops moving, but these cops interact with population agents when a crime is witnessed, producing an arrest.

The model is composed of a grid of cells, **the world**, in which **agents** move and act. The agents are divided into cops and two population groups, G_1 and G_2. The two populations commit crime at a constant rate c_0, and if a cop is present the crime-committing individual from the population will get arrested. Cops, who are unevenly distributed among the populations at the beginning of the simulation, move "following their noses" according to a **stigma field** which places bias in locations with crime history and reinforces repeated surveillance. The probability with which a cop will follow the stigma field is denoted θ and it will have an important role in the subsequent analysis. Once a population agent is arrested they undergo a trial stage, for which the trial decision (i.e., go to jail or not) is based on a random classifier. Notice we have set the crime and recidivism rates constant across groups. We will see that even under such assumption an original placement bias of police can lead to large disparities. For the rationale and justification of these assumptions, see the appendix.

The purpose of the paper is not to present a novel and complicated ABM for predictive policing or recidivism prediction. The goal is to present how we could use these models to address the questions stemming from a system-wide paradigm. We then invite the community to engage in interdisciplinary research and use these tools for a wide variety of applications. We think such approach can enlighten us in our efforts to bring about systemic change.

4 MODEL ANALYSIS

We will use two basic notions of fairness as described by [6]: Predictive Parity and Error Rate Balance. We define them slightly different than in [6], as they were defined using score outcomes and we use a hard 0/1 classifier outcome. These conditions are based on the false positive and false negative rates (FPR, FNR), as well as the positive predictive value (PPV) of the score or classifier, which only



(a) Arrested and population measures τ_A (dashed) and τ_P (solid).

(b) Proportion of arrested members of group G_1 to arrested members of group G_2.

Figure 2: Fairness assessment comparison and arrest proportion, the missing link.

witnesses people who have been arrested. Our analysis considers the entire system and not just the sentencing stage so that we make explicit the condition "arrest". In the following A stands for the arrest variable, R for true recidivism, J is the output of the classifier, and G is the group membership.

Definition 1 (A-conditioned Predictive Parity). *A system implementing classifier J satisfies A-conditioned predictive parity if the probability of true recidivism, given a positive classifier assignment and given A, is the same across groups. That is: $\mathbb{P}(R = 1|J = 1, A = 1, G = G_1) = \mathbb{P}(R = 1|J = 1, A = 1, G = G_2)$. We refer to these probabilities as $PPV_A(g)$.*

Definition 2 (A-conditioned Error Rate Balance). *A system implementing classifier J satisfies A-conditioned error rate balance if the False Positive and False Negative Rates, given A, are the same across groups. That is: $\mathbb{P}(J = 1|R = 0, A = 1, G = G_1) = \mathbb{P}(J = 1|R = 0, A = 1, G = G_2)$ for the FPR, and $\mathbb{P}(J = 0|R = 1, A = 1, G = G_1) = \mathbb{P}(J = 0|R = 1, A = 1, G = G_2)$ for the FNR. We refer to these probabilities as $FPR_A(g)$ and $FNR_A(g)$.*

Notice two main differences from [6]: First, we talk about a system implementing a classifier, as opposed to the classifier itself. Second we condition on the arrest variable A.

Equation (2.6) of [6] describes the relationship among PPV, FPR, FNR, and prevalence $p_A(g) = \mathbb{P}(R = 1|G = g, A = 1)$. With our notation: $FPR_A(g) = \frac{p_A(g)}{1-p_A(g)} \frac{1-PPV_A(g)}{PPV_A(g)}(1 - FNR_A(g))$. As stated in [6], if the prevalence differs across groups, we cannot obtain ERB and PP simultaneously. If prevalence is equal between the groups, however, it is possible to satisfy all of these fairness metrics.

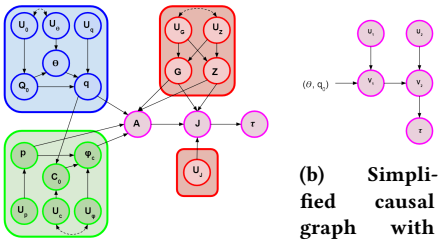
It is easy to show that for our model $FPR_A(g) = r_c$, $FNR_A(g) = 1 - r_c$, and $PPV_A(g) = r_0$. While this fairness outcome is satisfactory for the classifier itself, it doesn't provide information about system-wide fairness. To assess overall fairness in our model we can add the deviations from the ideal case, by defining the quantity $\tau_A := \left|1 - \frac{PPV_A(G_1)}{PPV_A(G_2)}\right| + \left|1 - \frac{FPR_A(G_1)}{FPR_A(G_2)}\right| + \left|1 - \frac{FNR_A(G_1)}{FNR_A(G_2)}\right|$. In the ideal case the numerator and denominator of each component are equal and therefore $\tau_A = 0$.

In Figure 2a we plot the averaged value of τ_A (dashed lines) over 30 simulations for different values of the parameter θ. τ_A approaches zero as the system stabilizes. However, as we will shortly see, there is still an unfair process not captured in this: the proportion of arrests among different groups. In Figure 2b we plot the ratio of arrested members of G_1 to arrested members of G_2, averaged over 30 runs. This time we note the arrest rates are highly disproportionate. In particular low values of θ (the probability with which cops follow the prior stigma of a neighborhood) lead to almost equal arrest rate. Large values of θ, however, can lead to ever-increasing extreme disparities between both arrest rates.

Disproportionate arrest rates are not justified when crime rates are the same and should be considered part of the fairness assessment. The observed difference could be explained by differences in crime rates; however, as already mentioned, the agents in the model commit crime at constant rate independent of group membership. There is another element driving these ratios upward. Recognizing this element motivates us to introduce population fairness metrics. Analogous to the previous definitions, we define **Population Error Rate Balance** as a system satisfying $\mathbb{P}(J = 1|R = 0, G = G_1) = \mathbb{P}(J = 1|R = 0, G = G_2)$ for FPR and similarly for FNR. We also define **Population Predictive Parity** as a system for which $\mathbb{P}(R = 1|J = 1, G = G_1) = \mathbb{P}(R = 1|J = 1, G = G_2)$.

It is straightforward to show that Population PPV equals A-conditioned PPV and hence population and arrest PP are equal conditions in this case. We now define a new overall measure of fairness, this time for the population: $\tau_P := \left|1 - \frac{PPV_P(G_1)}{PPV_P(G_2)}\right| + \left|1 - \frac{FPR_P(G_1)}{FPR_P(G_2)}\right| + \left|1 - \frac{FNR_P(G_1)}{FNR_P(G_2)}\right|$. Figure 2a shows τ_P averaged over 30 runs for several values of θ. This time we can easily see the discrepancy of results. If we consider how the implementation of the algorithms affects the whole population, taking into account the context under which data is collected, bias surfaces to light.

By adapting Chouldechova's trade-off to the new case, we obtain: $FPR_P(g) = \frac{p_P(g)}{1-p_P(g)} \frac{1-PPV_P(g)}{PPV_P(g)}(1 - FNR_P(g))$, where the population prevalence $p_P(g) = \mathbb{P}(R = 1|G = g)$. Again, ERB cannot be achieved because the prevalence is different, this time we get $p_P(g) = \mathbb{P}(R = 1|G = g, A = 1)\mathbb{P}(A = 1|G = g)$ which reveals the new culprit $\mathbb{P}(A = 1|G = g)$, the probability of arrest. Since this probability varies by group G, prevalence also depends on G. Remember, however, that cops in our model did not discriminate when they observed crime, and both groups had the same crime rate. What then leads to the different arrest probability? Well, although, unfortunately, there is well-documented evidence of disparities in arrest rates among different demographic groups as well as arrests and convictions of innocent people (see for example [18], [39] and the many articles in [38]), we do not treat such case here, and we assume the cops do not arrest when there is no crime or at disproportionate rates. Then we obtain $\mathbb{P}(A = 1|G) = \mathbb{P}(A = 1|C = 1, G)\mathbb{P}(C = 1|G)$. There are two reasons for which arrest probability would be different, crime rate and arrest probability given a crime is committed. We can similarly further break up $\mathbb{P}(A = 1|C = 1, G)$ conditioning on the cop being present, and it is then that the surveillance rate is revealed. We conclude that even if anything else in the system is "fair", dissimilar surveillance rates propitiate unfair outcomes. As our model shows, there is no need for cops to discriminate themselves, nor to surveil differently across groups. The only thing in our model enforcing systemic discrimination is that cops follow the historic stigma of a region, and that in the initial conditions there is a higher distribution of cops in the first group's region.



(a) Full causal graph

(b) Simplified causal graph with intervention

Figure 3: Causal graph and simplified intervention model.

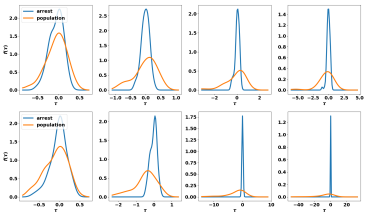
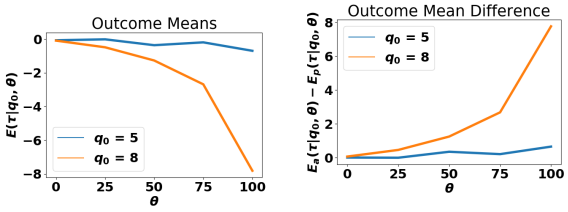


Figure 4: distributions of outcomes for $q_0 = .5$ (first row) and $q_0 = .8$ (second row). Each column is a value of theta, from low to high, in {0, .25, .75, 1}.



(a) Outcome means for population

(b) Difference between population and arrested outcome means

Figure 5: Comparison between outcome means for arrested and overall populations.

4.1 Causal Effect of Policing Parameters

A guiding principle for the paper is to provide possible causal mechanisms in addition to causal relations. Since the structure of ABMs is mechanistic in nature, ABMs are good candidates to fulfill this principle. The affinity to ascribe causal mechanisms for ABMs is set by any causal framework formulated around counterfactual outcomes (see [17]). Each simulation run (given a specified set of initial conditions) can be considered a counterfactual (see [23]), and it is under this framework that we explore the outcomes of our model.

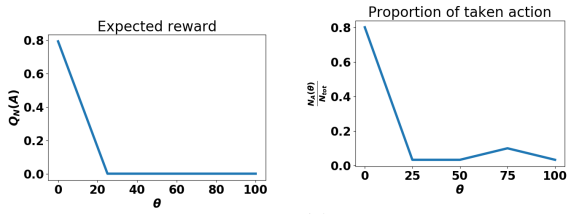
A full causal graph would at least include the relationships shown in figure 3a. The **U** variables represent endogenous variables, $(\mathbf{p}, \mathbf{q})$ are the coordinates of civilians and cops, respectively, $\mathbf{c_0}$ is a variable crime rate while $\psi_\mathbf{c}$ is a binary crime indicator, **Z** are personal characteristics and Q_0 is the bias with which cops are placed on the first zone at the beginning. We have made some assumptions in our ABM model that simplify the full graph. For example, the crime rate is held constant, while arrests by cops depend only on having committed a crime, and the probability with which cops will surveill with stigma is independent of their current location. Since our outcome of interest is the variable τ, after intervening on Q_0 and $\boldsymbol{\theta}$, we can group all other variables into variables **U** and **V** and hence further simplify the graph. We are now ready to consider different intervention outcomes. We do this by setting particular values of $\boldsymbol{\theta}$ and Q_0. The resulting model is shown in figure 3b, which provides us with the distribution of interest: $\mathbb{P}(\tau|do(\boldsymbol{\theta} = \theta), do(Q_0 = q_0))$, see [31], and [30]. Therefore, we can estimate causal effects from the outcome distributions.

For simplicity we focus on $\tau_A^1 := 1 - \frac{FPR_A(G_1)}{FPR_A(G_2)}$, and τ_P^1, defined similarly. To estimate these outcomes for different values of θ and q_0 we ran a simulation of the model for 5000 steps and computed the outcomes at the last step. We then averaged over 60 of these runs. We chose $q_0 \in \{.5, .8\}$ and $\theta \in \{0, .25, .5, .75, 1\}$. Figure 4 shows the kernel density estimates of the outcomes for the different combinations of (q_0, θ). The first row shows the results for $q_0 = .5$ and the second those of $q_0 = .8$. We notice that independent of the value of (q_0, θ), τ_A^1 concentrates much more around its means than τ_P^1. Furthermore, this mean is close to zero (desired outcome), as expected. This is not surprising in light of the results of section 4. What is surprising is that such stark difference prevails even for the "fair" initial allocation of cops dictated by $q_0 = .5$. The means of τ_P^1 given $q_0 = .5$ are indeed close to zero (see figure 5a), however, there is a nonnegligible amount of mass away from zero. This is because with an even distribution of cops on the two group neighborhoods, converging into a state of high stigma differences is equally likely across groups, but it remains a likely situation. That is, the mere dynamics of stigmatic surveillance make unfair population outcomes likely. The case $q_0 = .8$ is more drastic. We again have a great amount of mass away from zero, but also the means have shifted away from zero, showing a negative bias towards the non-privileged group. In Figure 5a we can see how rapidly the means move away from zero whenever $\theta > 0$. A comparison of the outcomes means between τ_P^1 and τ_A^1 is shown in figure 5b.

4.2 Finding Optimal Policies

The last guiding principle revolves around optimal interventions. To explore how to satisfy such principle we setup a simple multi-armed bandit problem [37]. We focus on the case $q_0 = .5$, and τ_P^1. The set of actions to be taken is the set of possible values for θ: $\{0, .25, .5, .75, 1\}$. The rewards are binary, based on the outcomes from section 4.1, in which a 1 is assigned if τ_P^1 does not exceed a tolerance value $\epsilon_{tol} = .5$. These are measured after letting the ABM run with the chosen value of θ for 3000 steps. The goal is to explore values of θ and design a policy to choose the θ that maximizes expected reward. While for the simple model it was easy to determine an optimal value of $\theta = 0$, this stage will be particularly beneficial in any future model with many more parameters and not explicitly tractable behavior.

We used an ϵ-greedy algorithm with $\epsilon = .1$ and with value function of an action equal to the average reward obtained when that action was chosen. Figure 6a shows the expected reward for



(a) Expected reward over 30 runs of the simulation.

(b) Proportion of an action taken by ϵ-greedy algorithm.

Figure 6: Expected reward and proportion of actions taken according to these rewards.

different actions over 30 runs of 3000 steps each. Figure 6b shows the proportion of times a specific action was taken. It does not take many runs for the algorithm to realize the optimal action is $\theta = 0$, that is, completely unbiased surveillance. Similar results hold for different values of q_0.

This simple example shows a reinforcement learning scenario is able to produce an optimal policy choice when dealing with this particular ABM model. Although simple, the purpose of this example is to be a precursor for more complicated settings. In reality, ABMs will have more than one parameter for which we'll want to create a policy. For example, we could introduce a "social worker" agent, a policy would then involve both θ as well as the distribution and behavior of these new social workers, a resource trade-off between social workers and cops, etc. The more we refine our basic AB model the more the need to have an automated way to learn an optimal policy. We believe reinforcement learning is a good candidate for such an enterprise.

5 CONCLUSION AND FUTURE WORK

We have presented a simple agent-based model with the aim of understanding some elements that give rise to inequality in an arrest-sentence system. With the aid of the model we discovered discrimination can occur at the population level, even if this discrimination is not apparent when studying the algorithmic tool in isolation. Through population level fairness metrics the model indeed shows highly disparate results. An important takeaway is that under very simple assumptions about agent and cop behavior, including constant crime and recidivism rate, as well as unbiased arrest in the presence of crime, discrimination can still arise as a consequence of prior history. The best way be de-bias the system we study is for the cops to ignore the stigma of neighborhoods, and follow a random path.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their comments and the University of Colorado Data Science to Patient Value (D2V) Initiative and Grohne-Stapp Endowment from the University of Colorado Cancer Center

REFERENCES

[1] Michelle Alexander. 2012. *The new Jim Crow: Mass incarceration in the age of colorblindness.* The New Press.
[2] James M Anderson and Paul Heaton. 2012. How much difference does the lawyer make: The effect of defense counsel on murder case outcomes. *Yale Law J.* 122 (2012), 154.
[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica* (2016).
[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The State of the Art. *Sociological Methods & Research* (2018).
[5] Richard A Berk and Justin Bleich. 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy* 12, 3 (2013), 513–544.
[6] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[7] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018).
[8] Cynthia Dwork and Christina Ilvento. 2018. Fairness under composition. *arXiv preprint arXiv:1806.06122* (2018).
[9] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway feedback loops in predictive policing. In *Proceedings of Machine Learning Research*, Vol. 81.
[10] Danielle Ensign, Frielder Sorelle, Neville Scott, Scheidegger Carlos, and Venkatasubramanian Suresh. 2018. Decision making with limited feedback. In *Algorithmic Learning Theory*. 359–367.
[11] Joshua M Epstein. 2006. *Generative social science: Studies in agent-based computational modeling.* Princeton University Press.
[12] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.
[13] Silvia Federici. 2004. *Caliban and the Witch.* Autonomedia.
[14] Andrew Guthrie Ferguson. 2017. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement.* NYU Press.
[15] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2018. A comparative study of fairness-enhancing interventions in machine learning. *arXiv preprint arXiv:1802.04422* (2018).
[16] Rachel Goodman. Winter 2018. Algorithms and civil rights: understanding the issues. *Civil Rights Insider* (Winter 2018), 3–4.
[17] Peter Hedström and Petri Ylikoski. 2010. Causal mechanisms in the social sciences. *Annual review of sociology* 36 (2010), 49–67.
[18] Innocence Project. 2019. https://www.innocenceproject.org/. Accessed 21/MAR/2019.
[19] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica* (2016).
[20] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.
[21] Michael Macy and Andreas Flache. 2009. Social dynamics from the bottom up: Agent-based models of social interaction. *The Oxford handbook of analytical sociology* (2009), 245–268.
[22] Michael W Macy and Robert Willer. 2002. From factors to actors: computational sociology and agent-based modeling. *Annual review of sociology* 28, 1 (2002), 143–166.
[23] Brandon DL Marshall and Sandro Galea. 2014. Formalizing the role of agent-based modeling in causal inference and epidemiology. *American journal of epidemiology* 181, 2 (2014), 92–99.
[24] Karl Marx. 1867. *Capital: Volume 1: A Critique of Political Economy.* Penguin Classics. Reprint 1990.
[25] Mary Jo Maynes and Ann Waltner. 2012. *The Family: A World History.* Oxford University Press.
[26] Andrew Morgan and Rafael Pass. 2019. Paradoxes in Fair Computer-Aided Decision Making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* ACM, 85–90.
[27] Daniel S Nagin and G Matthew Snodgrass. 2013. The effect of incarceration on re-offending: Evidence from a natural experiment in Pennsylvania. *Journal of Quantitative Criminology* 29, 4 (2013), 601–642.
[28] Arvind Narayanan. 2018. FAT* 2018 Translation Tutorial: 21 Definitions of Fairness and Their Politics. https://www.youtube.com/watch?v=wqamrPkF5kk. Accessed 23/AUG/2018.
[29] Cathy O'Neill. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Penguin.
[30] Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
[31] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms.* MIT press.
[32] Hazhir Rahmandad and John Sterman. 2008. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management Science* 54, 5 (2008), 998–1014.
[33] Greg Ridgeway. 2019. Experiments in Criminology: Improving Our Understanding of Crime and the Criminal Justice System. *Annual Review of Statistics and Its*

Application 0 (2019).
[34] Thomas C Schelling. 1969. Models of segregation. *The American Economic Review* 59, 2 (1969), 488–493.
[35] Thomas C Schelling. 1971. Dynamic models of segregation. *Journal of mathematical sociology* 1, 2 (1971), 143–186.
[36] Ray J Solomonoff. 1964. A formal theory of inductive inference. Part I. *Information and control* 7, 1 (1964), 1–22.
[37] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
[38] The New York Times. 2019. False Arrests, Convictions and Imprisonments. https://www.nytimes.com/topic/subject/false-arrests-convictions-and-imprisonments. Accessed 21/MAR/2019.
[39] The Sentencing Project. 2019. Report to the United Nations on Racial Disparities in the U.S. Criminal Justice System. https://www.sentencingproject.org/publications/un-report-on-racial-disparities/. Accessed 21/MAR/2019.
[40] Uri Wilensky and William Rand. 2015. *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. MIT Press.

A MODEL DETAILS

A.1 The World.

A grid in which agents can move and act, divided in two regions accounting for the neighborhoods in which our two types of agents (besides cops) interact. Cops will be able to move freely between regions, while the two population types are assumed restricted to one region, with no interaction among them. Such constraint reflects the way most neighborhoods have statistically skewed demographics. Indeed, an early example of ABM's aimed to explain racial segregation [34], [35]. There are no further constraints in the original setup of the world.

A.2 The population agents.

Two types: groups G_1 and G_2, where the groups are taken to represent a distinction over one or more sensitive or protected variables. For example, group G_1 could be a minority as defined by socially arbitrary racial categorizations, while group G_2 would be the majority or privileged group (say, white defendants). Besides position $\mathbf{p}_t^{(i)}$ of agent i at time t, their parameters are:

Crime rate. The rate c_0 at which agents commit crime. In this toy model such rate is kept constant across individuals, independent of group membership. Although in many cases true crime rates are not known, there is reason to believe some actions considered crime are constant across demographics groups, for example, drug use [20].

Recidivism rate. The rate r_0 at which individuals recidivate. Although this parameter is hard to estimate from real data, numerous studies provide evidence that prison sentences do not affect likelihood of rearrest, allowing us to make the simplified assumption that the classifier's decision and the iteration time won't affect r_0, see [27], [33].

A.3 Cop agents.

Cops are allowed to move freely among group neighborhoods. They are initially unevenly distributed among the populations, with a ratio of q_0 in the discriminated group. We encode the cops positions in the vector $\mathbf{q}$. Cops also have the following parameters:

Arrest rate given crime observed. In this simple model we also simplify the behavior of cops by assuming a constant arrest rate given crime observed r_a. Notice that this simplification is in general unrealistic as racial bias is well-documented among police ([20], [1]). However, there are instances in which race, at least directly, is not an influencing factor for arrest, as is the case of traffic violation and the so called "veil of darkness", [33]. Note, however, that a backdoor path is possible through, for example, car make, year, and well-kept status. In the simplest of cases we have set $r_a = 1$.

Surveillance bias. Cops direct their surveillance efforts by "following their nose". As explained below, there is a stigma field in the neighborhoods which the cops use to choose where to patrol. We denote by θ the rate at which cops follow a stigmatized route.

A.4 The classifier.

During this toy model we will keep the classifier as simple as possible, so we choose a random classifier with a jail sentencing rate of r_c. As unrealistic as this may seem, courts randomize assignment of defense lawyers and judges to defendants. In many cases the incarceration rates differ dramatically according to such assignment, allowing us to justify r_c in this case as the probability of being assigned a good/bad lawyer and a lenient/strict judge [2].

A.5 Stigma Field.

There is a variable at each point in space we call the "Stigma Field". It represents the bias towards a region where crime has been recorded. It is initialized to be zero.

A.6 The Setup.

The initial conditions can be described by the following parameters: *Population size.* An original population size N_G for each group. At the beginning we assume it to be the same across groups. *Original configuration.* At the beginning the agent population is randomly distributed across the grid points in their particular regions. *Cop distribution.* The cops are also placed at random but with a bias q_0 are placed in one region, while only $1 - q_0$ are placed in the other. Note this is only their original configuration, they can still move between regions.

A.7 Dynamics

Population agents first move to a neighboring cell at random and then, with probability c_0, commit a crime. Cops first, with probability θ, move to the neighboring cell with the highest value of the Stigma Field. Then face a random direction. With probability ω, move m_c steps, otherwise move one step. Finally they arrest agents in neighboring cells that have committed a crime this iteration. When arrest happens in a cell, we increase the Stigma Field at place of arrest by a given amount, and at neighboring cells by smaller but nonzero amount. An arrested agent is judged by the classifier with hard $0/1$ assignment and, with probability r_0, recidivates.

Regarding the parameters of our particular model, we chose the crime rate to be $c_0 = .01$. The recidivism rate was $r_0 = .4$; we chose recidivism rate this large not because it reflects truth but because it helped the model stabilize faster, smaller values also work. We also picked an initial population size of 100 and a cop probability of moving away from its position $\omega = .1$ and $m_c = 3$.