Locality of Technical Objects and the Role of Structural Interventions for Systemic Change

Efrén Cruz Cortés [*]
Michigan Institute for Data Science
University of Michigan
Ann Arbor, MI, United States
encc@umich.edu

Sarah Rajtmajer
College of Information Sciences and Technology
Rock Ethics Institute
Pennsylvania State University
State College, PA, United States
smr48@psu.edu

Debashis Ghosh
Department of Biostatistics and Informatics
University of Colorado School of Public Health
Aurora, CO, United States
debashis.ghosh@cuanschutz.edu

[*] Corresponding author. https://efrencc.github.io/

Abstract

Technical objects, like algorithms, exhibit causal capacities both in terms of their internal makeup and the position they occupy in relation to other objects and processes within a system. At the same time, systems encompassing technical objects interact with other systems themselves, producing a multi-scale structural composition. In the framework of fair artificial intelligence, typical causal inference interventions focus on the internal workings of technical objects (fairness constraints), and often forsake structural properties of the system. However, these interventions are often not sufficient to capture forms of discrimination and harm at a systemic level. To complement this approach we introduce the notion of locality and define structural interventions. We compare the effect of structural interventions on a system compared to local, structure-preserving interventions on technical objects. We focus on comparing interventions on generating mechanisms (representing social dynamics giving rise to discrimination) with constraining algorithms to satisfy some measure of fairness. This framework allows us to identify bias outside the algorithmic stage and propose joint interventions on social dynamics and algorithm design. We show how, for a model of financial lending, structural interventions can drive the system towards equality even when algorithmic interventions are unable to do so. This suggests that the responsibility of decision makers extends beyond ensuring that local fairness metrics are satisfied to an ecosystem that fosters equity for all.

1 Introduction

The widespread use of machine learning algorithms and the practice of data science to support everyday life has brought about unintended consequences. Some of these consequences such as environmental impact [67, 23, 44, 36], extractivism [17, 13], labor exploitation [40, 29, 30, 9, 61, 63, 26], and privacy intrusion [22, 11, 66, 56, 70, 42] are general to the practice and deployment of pervasive computing. A particular problem, salient in the last few years, is the form of discrimination exhibited towards certain population subgroups by data-informed algorithmic decision making systems. Algorithms have been shown to make biased predictions based on race [32, 57, 47, 58, 5, 6, 62], gender [46, 8, 14, 50] and class [27, 69, 31], among other attributes.

Algorithmic discrimination is for the most part unintentional, and much research is being done to correct the problem and identify its nature and origin. Broadly speaking, researchers have recognized two potential sources of bias–bias in algorithms and bias in data [51]. Bad algorithm design can disproportionately impact a single subset of the population (see [58] for a prominent example). Most modern algorithms, however, are based on data, so even well-designed algorithms can reproduce discrimination encoded inherently in the data itself [33, 2]. Data, in turn, is the result of mechanisms contingent on institutional norms, laws, ideologies, attitudes, protocols, and overall social dynamics. Likewise, a well-designed algorithm with unbiased data can still be discriminatory by virtue of its use within a biased system of decisions. Discrimination, indeed, is a social problem predating algorithmic decision making. While algorithms may exacerbate the situation and should be designed carefully, it is likely that algorithm design alone cannot fully solve the problem.

There are different perspectives regarding the extent to which computing, as an active part of the problem, can contribute to the solution. As pointed out in [1], fair machine learning research works well as a diagnostic tool to detect problems of discrimination and bias in algorithms [48, 4, 2, 41, 58]. Fairness metrics, however, may fail to properly diagnose discrimination if social context is not taken into account [19]. Although major sources of bias are due indeed to the environment in which algorithms are embedded [64], technologists direct their efforts towards better algorithmic design as the primary point of inquiry [52, 16, 37]. Critics of technology, on the other hand, argue that harmful bias can stem from a variety of situations and optimal design is not sufficient to counteract algorithmic discrimination, which merits different solution approaches [64, 21, 68].

The main criticism of technological solutions is that they don't account for the discriminatory structure inherent to society itself, algorithmic deployment and infrastructure, and differentiated impact. These are criticisms of scope rather than method. By de-centering algorithms as the objects of study, and taking a systemic approach, machine learning research itself can expand its contributions. Some authors propose, for example, considering implementation pipelines to assess discrimination over a linear system of decisions[25, 39]. In this paper, we propose a methodology to account for elements of algorithmic discrimination with social origin. That is, instead of considering the algorithm in isolation, we place it in the social context in which it is deployed. We thereby shift focus from the statistical properties of an algorithm to the properties of the system in which the algorithm resides. This perspective is complementary to fine-scale analysis of algorithmic fairness metrics.

Having placed an algorithm in context, we then explore the potential of different system-level interventions to reduce unequal outcomes. We compare the effect of intervening solely on the algorithm, by constraining it to satisfy fairness metrics, to the effect of interventions on the data generating mechanism and other social aspects of the system. As an example we use the bank loan problem described in [49], and show that social interventions are more effective than algorithmic interventions. Our goal is to show that systemic analysis falls within the scope of machine learning research, and should be carried forward as such.Most fair machine learning research focuses on algorithm-level solutions because the tools at our disposal are already calibrated to this level of analysis. Our intention in this work is to suggest that, with small modifications, we can use those same tools for systemic analysis.

In Section 2, we outline the use of machine learning and causal inference tools in a systemic approach to the problem of algorithmic discrimination. In Section 3, we build on the work of [18] to present the lending example in [49] within this framework, and present experimental results supporting our thesis that structural interventions are transformative when algorithmic interventions fall short. Section 4 presents suggestions for how to implement these principles in practice. We conclude in Section 5.

2 Systems Framework

The conversation about algorithm harms in social-technological systems often invokes the need for systemic change. However, it is not always clear what constitutes a system, what systemic change means, or how it is to be achieved. To begin answering these questions, we focus on intervention practices, and the difference between local and non-local interventions. We want to know when local interventions have a strong enough effect to change the system. For example, does enforcing a fairness metric in an algorithm change discriminatory harms in a population, or only conceals the role certain decisions take on keeping a structure of power as it is?

Our working definition of a system will be a set of objects, and a corresponding set of relationships and constraints these objects have. Let $S = (O, R)$ be the tuple of objects and relationships. For socio-technical systems we can divide objects into social agents and technical objects. Examples of technical objects include datasets, classifications algorithms, digital infrastructures, and even more abstract automated protocols. Each of these objects often exhibits internal structure, and can be related in multiple ways to other objects. Relationships themselves also exhibit structural patterns. A crucial one is that of modularity, through which we can compartmentalize objects and relationships often in the shape of separated mechanisms. Given a modular system, we can separate relationships into local relationships, which connect objects within a module, and global relationships, which connect modules themselves. Another important relationship is that of feedback, in which once we have identified some sort of temporal passage, the output of a later mechanism in the system has a relationship with the input of an earlier mechanism. Finally, we have emergence, which is a property of a system not exhibited by its parts.

Let's provide a concrete example in machine learning. Many decision making systems include as a crucial step a binary decision that members of a population are subjected to, for example whether or not to provide a financial loan, a medical treatment, a job offer, or a prison sentence. Automation is possible only because the individual for whom the decision is being taken is represented as a set of features. These representations are mapped onto the decision contingent on a particular attribute that, purportedly, divides the population into appropriate subgroups (race, sex, class, etc.). That is, for an individual who has been represented by features X and labeled as sharing the characteristics of subgroup A, the decision is taken through a mapping $f(X, A) = Y$ (A need not be used by f explicitly). The system itself has many components, f being just one of them. For analytical purposes, we can create subdivisions of these components.

Individuals develop the attributes by which they are represented through a complex mechanism of relationships, actions and social and institutional constraints, e.g., educational opportunities and cultural inheritance. We call this process the *data generating mechanism*. We then have the *sampling mechanism*, through which individuals are datafied as a set of features and these features are made available to the decision function, e.g., loan application, doctor visit, criminal sentencing. Then, through the *design mechanism* an algorithm is derived according to a scientific paradigm. Once designed, the *deployment mechanism* makes decisions over individuals. Finally, decisions made for individuals have effects on those individuals and their social groups. The aggregate of these effects is significant enough to elicit a change in the data generating mechanism. We call this the *feedback mechanism*. While for most systems these subdivisions are permeable, we will keep them

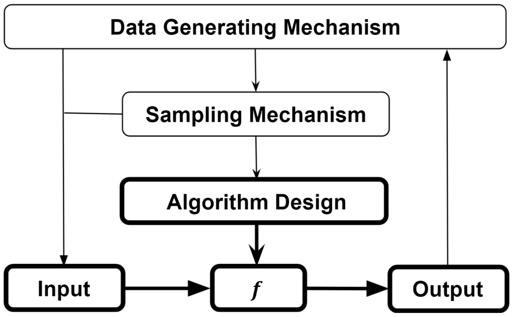

Fig. 1: Diagram of relationships among different objects of an automated decision making system. The bolded region indicates where algorithmic local analysis concentrates, leaving out generating mechanisms and feedback effects.

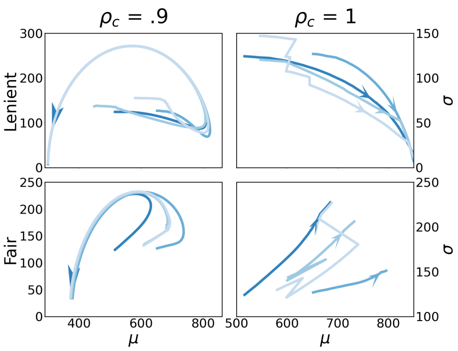

Fig. 2: Convergence behavior of a lending system with feedback. When the probability of repayment is low, the distribution concentrates on low scores after a while, independent on the classifier being fair or not. If the repayment probability is high, the opposite happens.

for practical use. Figure 1 provides a schematic of these mechanisms and their relationships in a decision making system. We note our assumption that each of these has internal structure.

We can now ask what the goal of ethical A.I. and fair machine learning research is. As stated in the introduction, these systems are often harmful, for example by reproducing discrimination. However, the harm arises from the many factors of the system, algorithm design being one of them. To put it bluntly, we'd want to know in particular if the effect of imposing fairness metrics on an algorithm is indeed transformative or unsuspectingly redeem unchanged oppressive mechanisms. Algorithms as tools of control attain a role–for example, the role of divination, or the role of punitive classification–by virtue of their position in a system and not necessarily by their detailed inner workings. Sharpening a tool does not imply bettering the system.

Let's think of it in terms of locality. We define a local intervention as that in which an object $\mathcal{O}$ is changed by another object $\mathcal{O}'$. A strictly local effect with respect to some global property P will then be one for which such an intervention does not affect that property, that is

$$P(S(\mathcal{O})) = P(S(\mathcal{O}')),$$

where we have omitted the relationships and rest of objects. If the property changes, then we can say systemic change has been achieved with respect to P. In contrast, we can define structural interventions as those which rearrange the relationships in the system, and as such are more likely to have global effects.

Much of fair machine learning research focuses on the design mechanism (bold region of figure 1), but properties like fairness, discrimination, and inequality are properties of the system as a whole. It is possible that achieving a certain property at the algorithmic level is not sufficient to achieve it at the system level, and may instead mask a lack for the overall population [19, 43]. In these cases, algorithmic interventions have strictly local effects, and therefore limited benefits.

As a particular example consider the design of a classifier we wish not to discriminate based on race. In the design process we have access to data D, a collection of mathematical objects representing individuals. Assume the race attribute is represented as a single categorical variable A. We can then set constraints $\mathcal{C}$ based on this A, and design the classifier f according to a minimization problem that looks like this:

$$\begin{aligned} f = \min_{h \in \mathcal{H}} \mathcal{E}(D, h) \\ s.t. \quad \mathcal{C} \end{aligned} \tag{1}$$

where $\mathcal{H}$ is an appropriate space of candidate classifiers and $\mathcal{E}$ some error. This system is indeed the bolded subdiagram in figure 1. Notice that while we mention classifiers, f could be something else, like a regression function or clustering assignment, depending on the problem. However, most approaches in fair machine learning have a similar structure as above.

There are two important aspects to the fairness-inducing constraint $\mathcal{C}$. First, that it can guarantee the satisfaction of fairness metrics only with respect to the scope of subsystem 1. Second, that it is based on the variable A which, we recall, represents attributes we wish to protect. Both of thee aspects have the limitations of locality, let's explore these further.

2.1 Local interventions and systemic transformation - scope of algorithmic subsystem

The constraint $\mathcal{C}$ can only be guarantee with respect to subsystem (1). Considering input and output (x, y) during deployment, the objects of this system are $(D, \mathcal{H}, \mathcal{C}, f, x, y) \subset O$. Any property we measure with respect to this system, like fairness, justice, equality, that we may want to optimize, can only be done with respect to these objects. The advantage of circumscribing our focus around f is that we can fine-tune the local behavior of our classifier. However, it has restricted knowledge of the data generating mechanism and the dynamics of a changing population, including feedback effects.

Changing the scale of analysis, we can view algorithms as objects in the system, and study the relationships it has with other objects. In so doing, we can identify if undesired characteristics of the system, for example discriminatory behavior, which are derived from the tuple (O, R), are mostly due to the implementation of a particular algorithm or to the structure of the system as a whole. Identification of such emerging patterns allows us in turn to judge what type of interventions we can do on algorithmic design, with what consequences, and what type of interventions must be done on other objects and relations of the system.

As a case example consider the bank loans scenario posed in [49]. In summary, members of a population (split into racialized groups) apply for loans at a bank and are either given the loan or denied. If they are given the loan they either repay or default, and their credit score changes according this outcome. The sub-populations exhibit some initial distribution of credit scores π which is updated at each decision stage. The property we are interested in is the shape of the distribution π, if it leans heavily to the lower scores, it represents a detrimental financial situation, if it leans heavily to higher scores, it represents stability and some amount of wealth[1].

The evolution of the system over time depends, among other parameters, in the decision policy and the probability of repayment. The decision policy is reflected in the classification algorithm itself, which will decide to give a loan or not depending on if the credit score is larger than a given threshold. In Figure 2 we plot the dynamics of a simplified version of this system [2]. We

[1] This is of course an idealization of the problem

[2] The details of the model are found in the appendix. We will revisit this model in section 3 but in a more complicated manner, so the dynamics will change a little. The perspective shown here provides us with initial intuition.

observe the distribution, represented by its mean and standard deviation, over time for different starting points, including the empirical distributions of each sub-population. In the first situation (top row) the lending policy is lenient, in that it provides a loan to everyone independent of the credit score. The left side shows the dynamics for the case in which the repayment probability is $\rho = .9$ over all scores, while in the right we have the case of perfect repayment.

As the figure shows, when the repayment probability is low, the system converges to a distribution concentrated on low score values, on the other hand, if the repayment probability is high, the distribution concentrates on high values. In the second situation, the lending policy has been chosen to satisfy a fairness metric (equalized odds). However, the convergence behavior remains the same, that is, even though the classifier is fair, the system attains the same final states depending on the value of ρ (similar results hold for other fairness metrics). What we see is that a local intervention (changing a technical object for its “fair” version) did not have a global effect on the system. A significant change in behavior is brought about instead when ρ changes. Now, ρ was represented as a single number for simplicity, however, it represents a complex combination of social factors evolving in the data generating mechanism. To change ρ, we would need to change the structure of interaction in this previous mechanism. This is the difference between local and structural interventions.

There are other considerations to take, for example, the factors driving the distribution of credit scores outside of pure feedback from the classifier. For example, FICO scores are calculated by looking at payment history, outstanding debt, length of credit history, number of new credit inquiries, and credit mix (number of bankcard trade lines)[54, 24]. Each of these factors, in turn, depends on other factors like person’s original wealth, family and communal relationships, educational level, socioeconomic background, among others. As such, there will be forces outside of an algorithm’s potential driving the system to a particular state.

To illustrate this scenario, consider a simplified generating mechanism which depends on a number of factors, and which produces a baseline distribution g_0 as in Figure 3. Once a distribution of scores has been updated after some decision process, it is mixed with the baseline g_0 and new observations are obtained from the mixture. Say the mixture has the form

$$\alpha g_0 + (1 - \alpha) g_t.$$

In this case, the parameter α represents the relative influence that the outcomes of decision process have with respect to the generating process. If the data generating process is robust, even the fairest of algorithms cannot counteract its constancy. In our lending scenario, if social factors of economic inequality are always present and constantly driving certain populations to disparate financial states, local interventions are severely limited.

It is important to emphasize we are not trying to diminish the importance of local effects. If a fairer algorithm helps even one person, then it is more than worth it. Instead we think machine learning research can expand its scope of analysis and provide a framework for which we can study systemic effects and interventions in a quantifiable way. By doing so, we complement standard algorithmic techniques with the critical plea of social science to look at systemic harms and solutions.

We will come back to this example in detail later in the paper. There we will use a particular modeling paradigm for (O, R), structural causal modeling. In that case, the transition to the larger system is eased through the concepts of endogenous and exogenous variables. Endogenous variables are those inside the system, these are the ones we assume we can intervene on and for which we can explain causal connections. Exogenous variables are excluded from the modeling in the system but accepted as influences on endogenous variables. Looking at Figure 1, the subsystem of analysis will be limited by the endogenous variables you have access to, often the variables you can observe. Exogenous variables are usually thought of as complicated background conditions that can't be interfered with. Our advocacy for systemic expansion implies co-opting the data generating mechanism–or at least portions of it–into our modeling scope. In a sense, this is equivalent to converting some of those exogenous variables into endogenous variables through modeling assumptions and knowledge from social sciences and related fields that indicate what type of relationships are involved. For more details on causal models, diagrams, exogeneity and counterfactuals see [59, 60].

Consider a property of the system like inequality, our goal is to determine which interventions change that property and to what degree. While the division is porous, local interventions in this case involve acting on a specific variable, while structural interventions involve acting on a collection of variables *and* the relationships these have on each other. Local interventions are more likely to be adopted since in general they are easier to implement, for example by enforcing a fairness metric in the design of a classifier, or preprocessing and post-processing data. Structural interventions are generally

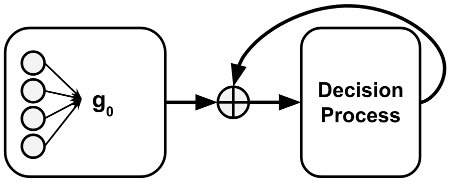

Fig. 3: Robust factors inside generating mechanism may hinder the effect of local interventions. In this case, the persistent distribution g_0 mixes with the results of the decision process. The strength of each in the mixture determines the dynamics of the system.

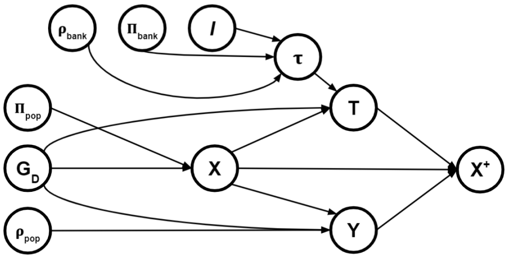

Fig. 4: Structural causal model for the lending system studied in Section 3. The inclusion of population characteristics (Π, ρ) and separation between datafied and structural categories permits analysis of extra-algorithmic interventions.

more expensive and complex. A type of structural intervention would include policies that constrain the behavior of institutions, such as economic, political, and educational reforms. Another type would constrain the behavior of individuals, for example by restructuring strategies for police surveillance and medical diagnosis. The complexity of structural interventions, however, does not preclude us from studying their effect on a system property. If, for a particular scenario, all information is filtered by an observed variable we do not necessarily have control over while designing an algorithm, we can still study interventions on that variable as proxies for having implemented a social policy that led to such changes. We will explore this approach for the case of lending in section 3.

2.2 Datafied vs Structural Categories

Another aspect of locality related to $\mathcal{C}$ is how we think of the variable A. $\mathcal{C}$ is expected to induce fairness of a similar property with respect to A, for example with the exigency that decisions be made independent to A. While we can statistically satisfy independence with respect to A, we cannot satisfy independence with respect to race, or gender, or class, since these are structural relationships in which individuals participate and cannot be detached from the individual. That is, it is imperative not to confuse the categories themselves with A, which is how we operationalize these categories.

Categorization of individuals is for the most part arbitrary, yet has significant consequences as those categories are used for the provisioning of resources and decisions [10, 53]. In demographic and statistical settings, these categories are often represented by a categorical variable, included in the vector of features representing an individual (race, gender, etc.). A basic goal of fair machine learning is to ensure there is no damaging discrimination due to membership to protected categories. Therefore, several fairness metrics are designed to address some type of independence between the variable representing a protected attribute and the algorithm's output. This approach, however, can be deceiving. A category like race, for example, cannot really be encompassed in a single variable. Instead, it is a set of variables, interactions, social biases and experiences, that the individual is subject to throughout their lives, and that influence all the observed variables inscribed in a particular feature vector. Hence, while statistical fairness metrics could enforce independence to a variable indicating race, which is possibly the result of an institutional agent checking a box in a survey or official form, they cannot enforce independence to "race" conceptualized as the structural position in which an individual finds themselves within a social context. This differentiation is seldom discussed in the fairness technical literature. To differentiate between the "check mark" categorical attribute and the real experience of that category, we call the former **datafied** and the latter **structural**. For example, datafied race would be the variable indicating one of the many institutionally defined "races" (white, black, asian, etc.). Structural race on the other hand would be a function of the system structure as a whole, not a variable within the system. Causal interventions and counterfactual analysis that rely on representing A as a single variable ontologically independent of other objects and mechanisms, necessarily ascribe to this modeling paradigm, and therefore attain a local scope through datafication.

In previous literature [45, 55] fairness has been proposed in terms of blocking the effect of race on the classifier. However, in those instances they refer to *datafied* race, a single variable. While the causal effect of this variable can be blocked by blocking specific paths, structural race cannot, as it would imply blocking most characteristics of the individual with respect to its society, which is impossible. For example, if an individual is represented by their credit score and datafied race (x, a), and we get rid of a, we might be tempted to equalize treatment on all individuals with same x level, however, level x means very different things for different subgroups. Just think about how much harder it is for a person born in poverty to achieve and maintain a high credit score compared to a person born into a wealthy family. To clarify a misconception, attributes like SES, family history, and neighborhood, are not *proxies* for race, meaning variables that exist by themselves outside of race (if correlated), they are indeed part of structural race. Notice that even a completely random classifier would not avoid discrimination because equal decisions have different impact on different subgroups. Therefore, trying to be "blind" to race or other structural categories is not a useful enterprise, and may even be counterproductive [58, 15, 3].

Causal inference does help us identify the differential effects of interventions on subgroups, and we will use a causal model ourselves. However, we will not model structural categories directly. Even if we complexify the concept and define a structural attribute as a set of variables and their connections, the particularities of those connections would be different for each person and context. To acknowledge structural categories but still have a working model, we will simplify them as a set of prior exogenous variables we don't have control over but that influence all observed variables. Other excellent approaches to tackle the problem of datafied versus structural race in machine learning and causal inference can be found in [65, 38, 34, 7].

3 Exemplary application to a lending system

We apply the systemic framework to a particular scenario, the lending system described in [49]. In this scenario, people, separated into subgroups according to their datafied race G_D, apply for loans to a bank. Individuals are judged by the bank's classifier according to their credit score X, and the credit score distribution Π and repayment probability ρ of the group they've been assigned to, notwithstanding personal history. The variable T indicates if they were given the loan or not, Y indicated their actually repaying, and X^+ their updated score. The true characteristics of the population (e.g. Π_{pop}) are not necessarily known to the bank, which may use an outdated estimate of these (Π_{bank}).

In [18] an SCM of the lending system was constructed. We have made some alterations based on the framework presented here, and present the one-step SCM in Figure 4. Notice we have not identified the structural category G_S with any particular node, and we don't want to commit to such a representation, since it is constituted by both endogenous and exogenous (as well as possibly not measurable) conditions [3]. Details of the model, including time dependency, feedback and functional relations, can be found in the supplementary material.

The classifier τ, following the setup in [49], is derived by maximizing the utility:

$$\tau^* := \arg \max_{\tau=(\tau_A,\tau_B)} \sum_{j\in\{A,B\}} g_j \sum_{x\in\mathcal{X}} \tau_j(x)\pi_j(x)u_j(x)$$

where $u_j(x) := u_+\rho_j(x) + u_-(1-\rho_j(x))$, for constants u_+ and u_-, is the bank's expected return for an individual with score x and from group j, g_j is the proportion of applicants from group j and each τ_j is a vector containing the probability of lending for a given score value x.

The optimal τ is a threshold classifier, meaning that if a credit score is greater than a group-specific threshold the loan is given, and denied otherwise. If fairness metrics are enforced this threshold value changes. The metrics we consider are demographic parity and equal opportunity, as used in [49] and defined in the supplementary material. Both fairness constraints involve an intervention on the node τ. Our interest is not to compare different types of fairness criteria, instead, we want to compare different types of interventions: algorithmic interventions on the one hand–that is, enforcing fairness criteria–which are local, and systemic interventions on the other.

Ideally we would use a model of the generating mechanism, with added objects and relations, in which we could intervene. In this example we are not adding such a model but we can still model the effects of structural interventions by appropriate interventions on the original distributions. That is, we have co-opted Π and ρ as endogenous variables, extending the scale of analysis to account for the conditions under which subgroups have different score distributions or financial capabilities for repayment. This variables are often left out of the model as exogenous and taken as essential characteristics of populations. Intervening on these variables reflects an intervention on the data generating mechanism. A full model for their determination would include the family and social support individuals have, wealth, income, financial and educational opportunities, health status, commuting time, neighborhood resources, social capital, etc.[4] As such, Π and ρ represent a network of

[3] For now it is important to separate them analytically so we position claims about fairness and causal paths properly.

[4] If such a model is available, that is, if a working social theory can work as the basis

relationships from which individuals eventual attributes are derived; changing them represents a change in the system as a whole. We will compare systemic interventions, realized through interventions on Π and ρ, to algorithmic interventions, realized through intervention on τ.

Using this model, we initialize the score distributions and probability of repayment for each group with the data provided by [49]. We also borrow and use portions of the code provided by [49] (BSD-3 license) for our implementation of the classifier and fairness metrics constraints. The data contains no personally identifiable information. Specifics of model implementation and parameter choice can be found in the supplementary material. Code can be found online at https://github.com/this url.

The system evolves over time. The initial score distributions for each group are shown in Figure 5 (top). If we enforce any of the fairness metrics described above, or we don't do any intervention, the classifier eventually drives inequality into the system, as shown in Figure 5 (bottom), where the final distributions after $N = 40$ steps are plotted for the case of demographic parity enforcement. Similar distributions result from maximum profit and equal opportunity. The takeaway is that subgroup A does not have the same opportunities as B when it comes to repayment, so most of the population end up with lower credit scores.

In order to assess which interventions are more beneficial we need to define a metric m for what a "good" outcome is. Since we don't include other indicators of wellness in our population (health, education, etc.) we will use their credit score distribution as a diagnostic [5]. An ideal distribution would be one where everyone in the population has high scores. For our lending scenario we will look instead at the relative number of people whose score exceeds the threshold value set by the bank, that is, the percentage of the population with the opportunity to receive a loan in the first place. Therefore $m(\pi) := P_\pi(X > \tau)$. Note that π and τ change over time. We applied three intervention cases.

Intervention 1. First, we only intervene on the classifier by enforcing different fairness metrics, either equal opportunity (EO) or demographic parity

for the data generation, it can be properly merged with the variables one has access to. A promising direction is to use agent based modeling or other simulation paradigms. See [35, 19, 12].

[5] The use of these proxy variables may be problematic in itself. We acknowledge that limitation but remind the reader that the focus is not on obtaining the best model, but to determine how, under established models, different intervention types affect the population

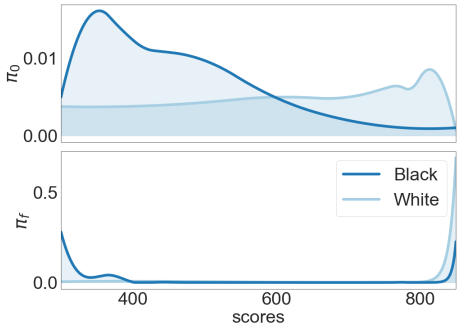

Fig. 5: Initial and final distributions of both populations, π_0 and π_f, after $N = 40$ iterations of enforcing DP, smoothed for visualization. The non-privileged population, starting with lower scores, ends with even lower scores; the privileged group's scores concentrate in high regions. Similar results hold for EO and MP.

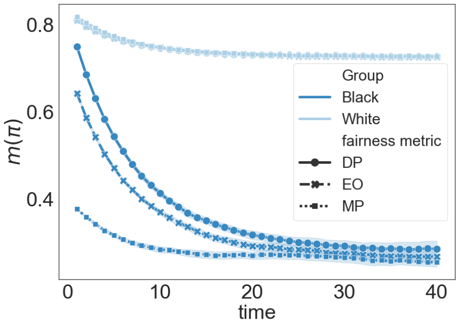

Fig. 6: Convergence behavior of the wellness metric m for both groups under different metric constraints, without social interventions. While fairness constraints are better than unrestricted MP, the improvement is marginal, and an inequality gap is formed after several generations of the system.

(DP). As a baseline we use the case in which no intervention on the classifier is made (utility is maximized without constraints), we call this baseline maximum profit (MP). This intervention is strictly local, as we modify only one technical object (the classifier itself). It's global effect will be seen in the change of $m(\pi)$. Succinctly, the intervention is

$$f_\tau \leftarrow f'_\tau.$$

Intervention 2. Second, we explore the case in which the score distribution of the population is changed, by setting the score distribution of the vulnerable population to that of the privileged population. This intervention models the hypothetical case in which the disadvantaged population exhibits the same score distribution as the privileged one, even if their repayment probabilities remain the same. Since this is a change in initial conditions it can be seen as a "history-correcting" situation, in which everyone started in a level playing field. To achieve that, even in a hypothetical scenario, the historical factors yielding those initial conditions would have to be modified, therefore this is a structural intervention. Succinctly, it is

$$\pi_A \leftarrow \pi_B.$$

Intervention 3. Finally, we explore the case in which the probability of repayment is changed. In this case, because the bank will still set a threshold based on the heavily skewed π values, a significant percentage of population A will not have the chance to get a loan and improve their scores. Hence, to boost initial lending rate on A, we further adjust ρ to match the initial percentage of people in B who get the loan (see supplementary material for further details). This just means we are providing population A with the same repayment power as population B. If we were to implement such an intervention in real life, it would be a structural intervention, since the many factors influencing repayment capability, like education and job opportunities, must be modified. Succinctly, this intervention is

$$\rho_A \leftarrow \rho_B + \delta,$$

where δ ensures equal initial percentage of loan opportunities.

For both the second and third interventions, we also explore their combination with the baseline intervention by jointly enforcing EO or DP. Intervention 1 changes the way a classifier is designed, based on fairness principles. This intervention can be considered *local* to the algorithm. It does not intervene directly on any element of the social system outside algorithmic design. The second and third interventions represent actions taken on the social system to change the properties of the population. The particularities of how these would be carried out are beyond the scope of this paper, but if policies were enacted such that as a result the properties of the populations changed in such a way, then we could assess their causal effect. The last two interventions are structural interventions because they encompass a restructuration of the system outside of the algorithmic locus.

Figure 6 shows the evolution of $m(\pi)$ over time for Intervention 1 and both fairness metrics. We can see that both fairness metrics DP and EO improve over the “unfair” classification scheme MP. However, there remains a substantial gap in outcomes among different groups. That is, even if at the classification stage of the system we have acted fairly, the disparity in well-being of the populations has not really been addressed. Using the language of our paper, this local intervention has a strictly local effect. As discussed before, it provides an opportunity to reassess the research goals. If our

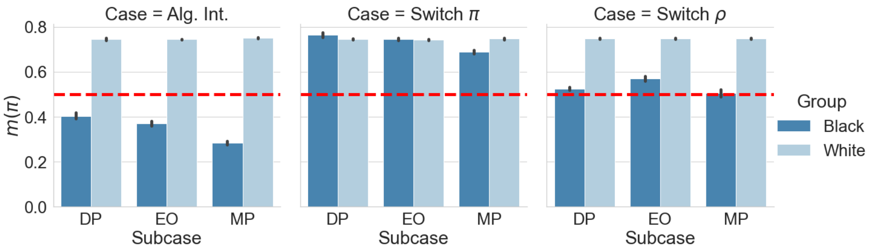

Fig. 7: Probability of credit scores being greater than the lending threshold for both subgroups after $N = 40$ generations of the system. (Left panel) Only algorithmic, local interventions have been applied and inequality prevails. (Middle and right panels) A combination of algorithmic and proxies for systemic interventions is applied, in one case the inequality gap is greatly reduced, in the other it is virtually nullified. The dashed line indicates the 0.5 point, and it may hint towards a phase change.

intentions are to make a technical object locally fair, the mission is complete, if our intentions are to ensure all populations are free of harm induced by this socio-technical system, more needs to get done.

We then compute the effect of structural interventions 2 and 3 in conjunction with local fairness interventions. The end result of these, together with those of intervention 1, are shown in Figure 7 (results shown for 40 steps, at which point convergence has happened). The dashed line at 0.5 indicates that half of the mass is above the desired threshold. Just as in figure 6, not intervening at the system level results in high inequality among both groups. Indeed, less than half of the unprivileged population even has the further opportunity to get a loan (first panel). On the contrary, if we add systemic interventions (second and third panels), inequality is greatly diminished. In the case of intervention 3 some parameters can be changed to achieve better convergence, but we prefer to be conservative. It still manages to provide around half of the disadvantaged population with high credit scores. As a point of clarification, we must point out that our experimental results are limited to the lending model discussed. Like any model, it is bound to miss intricacies of reality. That is OK, since the constant refinement of models, assumptions and hypotheses is inherent part of scientific pursuit.

4 A roadmap

As exemplified in the simulation experiment, interventions applied at the localized algorithm deployment stage are not as impactful in reducing harm as interventions applied to the social mechanism encompassing and sustaining the algorithm itself. This does not imply that algorithm designers have no power to make things better; rather, it suggests they must increase their scope of analysis, e.g., to ensure harms are not reintroduced during algorithm deployment. They might also consider whether the *point of deployment* of the algorithm is displaced. For example, if using machine learning to *predict* recidivism results in harm, we may wish instead to using machine learning to *prevent* recidivism, for example, by reallocating resources to social work and community support [28]. To make concrete the adoption of systemic thinking in algorithmic design, we provide four general principles for the machine learning researcher to incorporate in their practice:

Reformulate. While the discourse surrounding harm in algorithmic systems focuses on fairness, the problem may be rooted in other sources of discrimination and injustice. If that is the case, a proper action to take is to replace the questions driving research. For example, the causal inference question "what causes disparity among populations" can be replaced with "what causes harm to a population". Furthermore, we can ask ourselves if the effects of a particular intervention will change overall properties of the system or will remain local. By doing so, efforts towards fairness are complemented by other regulations that diminish noxious impacts.

Identify. Identify the stakeholders involved. In particular, center those who can possibly be harmed by algorithmic deployment in the analysis as opposed to those who profit from it. Furthermore, identify the relationships (causal or not) among the technical objects and social agents involved. A degree of modeling may need to be done to establish this relationships, and should include domain experts. Finally, identify relations that are potential sources of harm, these include the locally trained algorithm, the mechanism through which data is obtained, the feedback mechanism, and other stages of the system.

Structuralize. If provided with a datafied category, make explicit the distinction between the structural category it is associated with. This way, we complexify a social category, previously modeled as a technical object, into a system structure. If enough information is available to model the dynamics that compose a structural category, add that model to the analysis.

Shift efforts from trying to block causal paths from a structural category to identifying which elements harm populations by virtue of the structure of the system.

Expand. Expand spatially, meaning enrich the model to account for context and external mechanisms, even by co-opting exogenous variables. Expand temporally, meaning that, if available (often from a posited social theory), add a model for the generative mechanism of your data, which comes prior to algorithmic deployment, and add a feedback mechanism by considering the impact of decisions on the population. If not available, explore how changes in the original distribution may reduce harm, as well as long-term effects produced by feedback loops.

These principles make up the easy-to-remember acronym **RISE** (we promise the reader it came up naturally, nice coincidence), we hope these principles help the practitioner reconceptualize the problem of "fairness" in machine learning (or more generally, mitigation of harms) and can be applied in practice to incorporate systems thinking as well as social scientific insights. We know the list is not complete, as more considerations will undoubtedly come up as research and the problem space develops. However, these key principles serve as a basis for a trending shift in perspective from the merely technical to the social systemic.

5 Conclusion

Social systems are composed of many dynamic interrelated parts. To understand the properties emerging from those systems we should complement different analytical frameworks with each other. For systems for which decisions are automated with machine learning technologies we propose expanding the scope of analysis with a structural approach, complementing the standard algorithmic regularization viewpoint. The key advantage of a systemic viewpoint is that we can identify areas of intervention and their consequences inside *and* outside the algorithmic design stage, by considering the concept of locality. Using data from a financial lending system in the USA and building on the model of [49, 18], we showed that difference in impact is much greater when applying structural interventions than when only enforcing local fairness metrics. Indeed, only constraining an algorithm is not sufficient for changing the state of the system.

To clarify, we do not imply that institutions applying machine learning should be free of accountability. To the contrary, many of the complex social mechanisms that are not seen in the algorithmic stage are nevertheless entangled with the institution's position and practices, including labor, environmental, political and economic actions. We do not think the work on fairness from a localized, algorithmic viewpoint is futile. While fairness metric interventions are not sufficient to solve the problem, they are excellent tools for diagnosing it, and to ensure fair treatment among individuals at that particular stage. We should use them for that, but make companies responsible for more transformative practices. These approaches are part of a larger set of strategies which should be understood jointly to find a real solution to discrimination and inequality. Significant improvements can only be brought about through structural change as a collaboration of technology and social policy designers.

References

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

[2] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.

[3] Evan P Apfelbaum, Kristin Pauker, Samuel R Sommers, and Nalini Ambady. In blind pursuit of racial equality? *Psychological science*, 21(11):1587–1592, 2010.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[5] Ruha Benjamin. Assessing risk, automating racism. *Science*, 366(6464):421–422, 2019.

[6] Ruha Benjamin. *Race after technology: Abolitionist tools for the new jim code.* Polity, 2019.

[7] Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the 2019 conference on fairness, accountability, and transparency*, pages 289–298, 2019.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364, 2016.

[9] Albert Borgmann. *Technology and the character of contemporary life: A philosophical inquiry.* University of Chicago Press, 1987.

[10] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: Classification and its consequences.* MIT press, 2000.

[11] David Brin. *The transparent society: Will technology force us to choose between privacy and freedom?* Perseus (for Hbg), 1999.

[12] Elizabeth Bruch and Jon Atwell. Agent-based models in empirical social research. *Sociological methods & research*, 44(2):186–221, 2015.

[13] Tom Burgis. *The looting machine: Warlords, oligarchs, corporations, smugglers, and the theft of Africa's wealth.* PublicAffairs, 2016.

[14] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[15] Victoria Cheng, Vinith M Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 149–160, 2021.

[16] Bo Cowgill and Catherine E Tucker. Algorithmic fairness and economics. *The Journal of Economic Perspectives*, 2020.

[17] Kate Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press, 2021.

[18] Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, pages 2185–2195. PMLR, 2020.

[19] Efrén Cruz Cortés and Debashis Ghosh. An invitation to system-wide algorithmic fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 235–241, 2020.

[20] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

[21] David Danks and Alex John London. Algorithmic bias in autonomous systems. In *IJCAI*, volume 17, pages 4691–4697, 2017.

[22] Judith Wagner DeCew. *In pursuit of privacy: Law, ethics, and the rise of technology.* Cornell University Press, 1997.

[23] Payal Dhar. The carbon impact of artificial intelligence. *Nature Machine Intelligence*, 2:423–5, 2020.

[24] DoughRoller. A rare glimpse inside the fico credit score formula.

[25] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. Individual fairness in pipelines. *arXiv preprint arXiv:2004.05167*, 2020.

[26] Nancy Ettlinger. The governance of crowdsourcing: Rationalities of the new exploitation. *Environment and Planning A: Economy and Space*, 48(11):2162–2180, 2016.

[27] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, 2018.

[28] Andrew Guthrie Ferguson. *The rise of big data policing.* New York University Press, 2017.

[29] Todd C Frankel, Michael Robinson Chavez, and Jorge Ribas. The cobalt pipeline. tracing the path from deadly hand-dug mines in congo to consumers' phones and laptops. *The Washington Post*, 30, 2016.

[30] Christian Fuchs. Labor in informational capitalism and on the internet. *The Information Society*, 26(3):179–196, 2010.

[31] Gemma Galdon Clavell, Mariano Martín Zamorano, Carlos Castillo, Oliver Smith, and Aleksandar Matic. Auditing algorithms: On lessons learned and the risks of data minimization. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 265–271, 2020.

[32] Megan Garcia. Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, 33(4):111–117, 2016.

[33] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126, 2016.

[34] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512, 2020.

[35] Daniel Heard, Gelonia Dent, Tracy Schifeling, and David Banks. Agent-based models and microsimulation. *Annual Review of Statistics and Its Application*, 2:259–272, 2015.

[36] Lorenz M Hilty, Claudia Som, and Andreas Köhler. Assessing the human, social, and environmental risks of pervasive computing. *Human and Ecological Risk Assessment*, 10(5):853–874, 2004.

[37] Sharona Hoffman. Biased ai can be bad for your health – here's how to promote algorithmic fairnes. *The Conversation*, 2021.

[38] Lily Hu and Issa Kohler-Hausmann. What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*, 2020.

[39] Chong Huang, Arash Nourian, and Kevin Griest. Hidden technical debts for fair machine learning in financial services. *arXiv preprint arXiv:2103.10510*, 2021.

[40] Annie Kelly. Apple and google named in us lawsuit over congolese child cobalt mining deaths. *The Guardian*, 16, 2019.

[41] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.

[42] Rabia Khan, Pardeep Kumar, Dushantha Nalin K Jayakody, and Madhusanka Liyanage. A survey on security and privacy of 5g technologies: Potential solutions, recent advancements, and future directions. *IEEE Communications Surveys & Tutorials*, 22(1):196–248, 2019.

[43] Dean Knox, Will Lowe, and Jonathan Mummolo. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637, 2020.

[44] Andreas Köhler and Lorenz Erdmann. Expected environmental impacts of pervasive computing. *Human and Ecological Risk Assessment*, 10(5):831–852, 2004.

[45] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[46] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7):2966–2981, 2019.

[47] Nicol Turner Lee. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 2018.

[48] Jiuyong Li, Jixue Liu, Lin Liu, Thuc Duy Le, Saisai Ma, and Yizhao Han. Discrimination detection by causal effect estimation. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1087–1094. IEEE, 2017.

[49] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[50] Masoud Mansoury, Himan Abdollahpouri, Jessie Smith, Arman Dehpanah, Mykola Pechenizkiy, and Bamshad Mobasher. Investigating potential factors associated with gender discrimination in collaborative recommender systems. In *The Thirty-Third International Flairs Conference*, 2020.

[51] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[52] Sendhil Mullainathan. Biased algorithms are easier to fix than biased people. *The New York Times*, 2019.

[53] Jerry Z Muller. *The tyranny of metrics.* Princeton University Press, 2019.

[54] MyFico. How are fico scores calculated.

[55] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[56] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life.* Stanford University Press, 2009.

[57] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism.* nyu Press, 2018.

[58] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[59] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer.* John Wiley & Sons, 2016.

[60] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms.* The MIT Press, 2017.

[61] Søren Mørk Petersen. Loser generated content: From participation to exploitation. *First Monday*, 2008.

[62] Tage S Rai. Racial bias in health algorithms. *Science*, 366(6464):440–441, 2019.

[63] Siddhartha Sarkar. Use of technology in human trafficking networks and sexual exploitation: A cross-sectional multi-country study. *Transnational Social Review*, 5(1):55–68, 2015.

[64] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

[65] Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19, 2016.

[66] Daniel J Solove, Marc Rotenberg, and Paul M Schwartz. *Privacy, information, and technology.* Aspen Publishers Online, 2006.

[67] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

[68] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[69] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14, 2018.

[70] Heng Xu, Hock-Hai Teo, Bernard CY Tan, and Ritu Agarwal. The role of push-pull technology in privacy calculus: the case of location-based services. *Journal of management information systems*, 26(3):135–174, 2009.

Supplementary Material

This supplementary document is intended to provide additional detail about the lending model and simulations. We first explain the dynamic system used in section 2, then the one-stage model and then the time-dependent model. Implementation code can be found at https://github.com/this url.

Dynamic loan system - simple version. Following [49] we consider a lending system in which different sub-populations have different credit score distributions and repayment probabilities. Let the score distribution π be defined over score values $[300, 850]$. For this section, we update this distribution entirely based on the classifier's decision and probabilistic outcomes. The classifier is a threshold classifier, so for scores above a given threshold is gives a loan and for those below it denies the loan. If a person receives the loan they can either repay or default. If they repay, their credit score increases by c^+, if they default, it decreases by c^-. We don't need to do any sampling under this simple scheme because we are only following the evolution of the distribution itself, so we can do it by constructing a transition matrix. Note this is a simplified version of the system described in the next section. Here, τ and ρ remain unchanged so the transition matrix Q at each step is the same. It's values are:

$$Q(x, x') = \begin{cases} 1 - \rho(x)\tau(x) & if x = x_{min} \\ 1 - (1 - \rho(x))\tau(x) & if x = x_{max} \\ 1 - \tau(x) & if x = x' \notin \{x_{min}, x_{max}\} \\ (1 - \rho(x))\tau(x) & if x + c^- = x' \neq x_{min} \\ \rho(x)\tau(x) & if x + c^+ = x' \neq x_{max} \end{cases}$$

We have chosen $(c^-, c^+) = (-20, 10)$ to slow down convergence, and $\rho(x)$ to be a constant value across scores. However, ρ is in general a function of scores, as seen in the next section. For τ, the lenient threshold is the minimum score 300, the Maximum Profit threshold is 540 for population A (labeled "Black") in the data, 389 for Demographic Parity and 432 for Equalized Odds. For the B population (labeled "White") the values are 514, 520, and 524, respectively. These four different cases are plotted in figures 8, 9, 10, and 11. We have also plotted skewness vs standard deviation to better visualize the distribution shape.

For each combination of (ρ, τ) we run the system for different initial points. The first two curves always indicate the empirical π_A and π_B from the

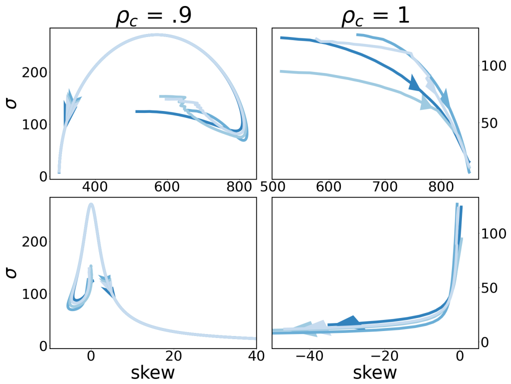

Fig. 8: System Dynamics for the case of lenient classification: threshold is $x_{min} = 300$. (Top) Mean vs standard deviation plot. (Bottom) Skewness vs standard deviation plot.

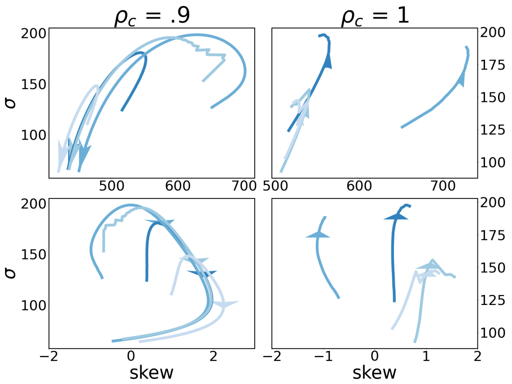

Fig. 9: System Dynamics for the case of maximum profit classification: threshold is 540. (Top) Mean vs standard deviation plot. (Bottom) Skewness vs standard deviation plot.

data, the following curves start at randomly sampled initial distributions. We sample these from a Dirichlet distribution with parameter ω. To construct ω we sample a value p from the uniform and set the first $n/2$ elements of ω to p and the rest to $1 - p$, and then normalize. To construct nicer looking plots we actually mix these sampled πs with the empirical distributions of A (50/50), which keeps initial points closer together. Other initial distributions have similar trajectories.

One-stage model. The original dataset uses Black, Asian, Hispanic, and White as datafied races, we limit our analysis to data from individuals labeled Black and White, and call these groups A and B, respectively, so that $G_D \in \{A, B\}$. The only features used for classification are the population membership G_D and the credit score $X \in \mathcal{X} = [300, 850]$. Each population $g \in \{A, B\}$ has an initial distribution over credit scores π_g and a function $\rho_g(x)$ which is the population-specific repayment probability based on credit score. When the membership g is irrelevant to the discussion, we will use the simpler notation π and ρ. Y_i denotes the actual outcome of repayment for an individual who was given a loan. Note group membership G has been divided into an endogenous variable G_D, datafied race, and the exogenous context G_S representing structural race. It is important to keep a clear distinction among the two, since otherwise statements like "the classifier takes decisions

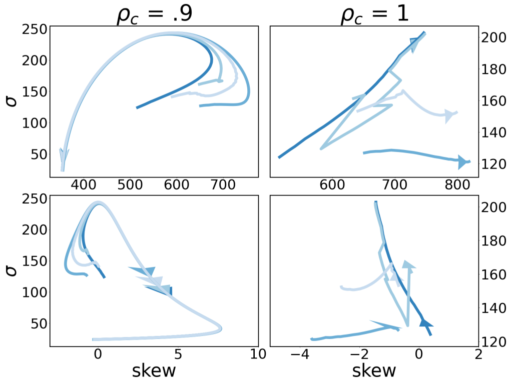

Fig. 10: System Dynamics for the case of demographic parity constrained classification: threshold is 389. (Top) Mean vs standard deviation plot. (Bottom) Skewness vs standard deviation plot.

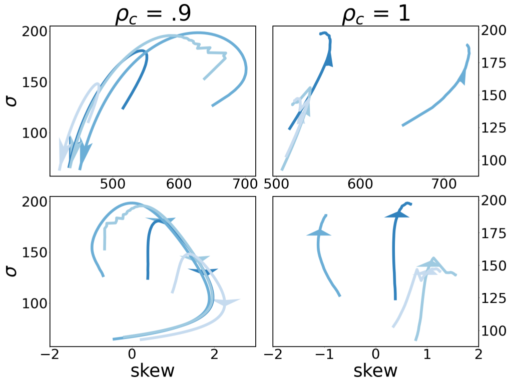

Fig. 11: System Dynamics for the case of equalized odds constrained classification: threshold is 432. (Top) Mean vs standard deviation plot. (Bottom) Skewness vs standard deviation plot.

independent of race" or "we have blocked the causal path from race to the classifier" do not make sense.

Banks judge individuals according to their credit score X. Each population has a random distribution of credit scores Π. If Π takes a particular form for subgroup j, we call that value π_j. The bank has access to a version of this distribution, which we call Π_{bank}, while the population, being dynamic, has a different distribution Π_{pop}. Separating variables this way accommodates for the situation in which the bank has outdated or incomplete data of the population. It also allows us to explore the possibilities in which interventions can be taken on the bank's treatment of the populations but not on the populations themselves, and vice-versa. The probability of repaying a loan, as a function of credit score, is denoted by ρ, and it is separated in the same way as Π. The credit score of an individual is denoted by X, while the actual repayment outcome (not available to the bank before classification) is denoted by Y. If an individual receives a loan and then pays or defaults their credit score is affected, we denote the changed credit score by X^+.

The algorithmic decision making system assigns individuals to one of two conditions, either "receive loan" or "not receive loan", based on a profit maximization paradigm, where profit is a function of π and ρ. We follow the setup in [49] and pose the problem as,

$$\tau^* := \arg\max_{\tau} \ \mathcal{U}(\tau),$$

where

$$\mathcal{U}(\tau) = \sum_{g \in \{A,B\}} r_g \sum_{x \in \mathcal{X}} \tau_g(x)\pi_g(x)\mathbf{u}_g(x).$$

The quantity $\mathbf{u}_g(x) = u_+\rho_g(x) + u_-(1-\rho_g(x))$ is the expected profit per score based on population membership and r_g is the relative ratio of group g, and each τ_g is a vector containing the probability of lending for a given score value x. More details on these parameters are given below.

The optimal τ^* has a threshold form:

$$\tau_g(x) = \begin{cases} 1 & x > c_g \\ \gamma_g & x = c_g \\ 0 & x < c_g \end{cases} \tag{2}$$

for a group specific threshold score c_g and $\gamma \in (0,1]$. T is a binary variable denoting the actual lending decision taken, and is realized probabilistically through $P(T = 1 \mid X = x, G_D = g) = \tau_g(x)$. Once a decision is taken, credit score changes according to the applicant having repaid or defaulted. If the applicant does not receive a loan, their credit score remains the same. The updated credit score is denoted by X^+. The relationship among these variables is illustrated in figure 13.

There are two fairness constraints studied in [49], which we also consider here. First, demographic parity (DP), which enforces equal lending rates, $P(T = 1 \mid G = A) = P(T = 1 \mid G = B)$, and which can be expressed as $\sum_{x\in\mathcal{X}} \tau_A(x)\pi_A(x) = \sum_{x\in\mathcal{X}} \tau_B(x)\pi_B(x)$. Second, equal opportunity (EO), defined as equal true positive rates, $P(T = 1 \mid Y = 1, G_D = A) = P(T = 1 \mid Y = 1, G = B)$, which can also be expressed as

$$\frac{\sum_{x\in\mathcal{X}} \rho_A(x)\tau_A(x)\pi_A(x)}{\sum_{x\in\mathcal{X}} \rho_A(x)\pi_A(x)} = \frac{\sum_{x\in\mathcal{X}} \rho_B(x)\tau_B(x)\pi_B(x)}{\sum_{x\in\mathcal{X}} \rho_B(x)\pi_B(x)}.$$

When the classifier is unconstrained, we call it maximum profit (MP). We denote by $\mathcal{I}$ the variable indicating what type of fairness constraint is set on the classification design. Both of these constraints yield classifiers τ of the same form as Equation 2 but with different thresholds. Enforcing a certain

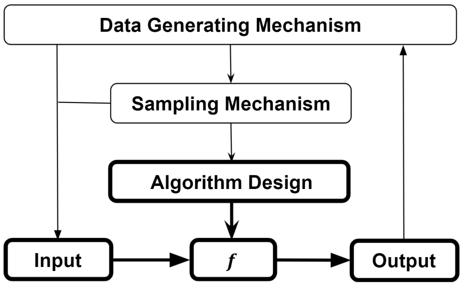

Fig. 12: Diagram of relationships among different objects of an automated decision making system. The bolded region indicates where algorithmic local analysis concentrates, often leaving out the generating mechanisms and feedback effects.

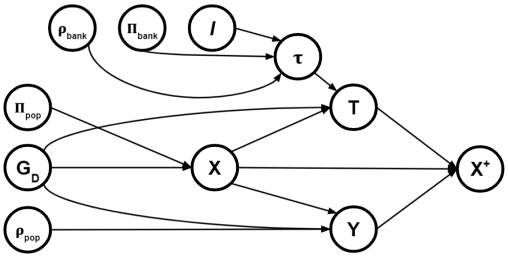

Fig. 13: Structural causal model for the lending system studied in Section 3. The inclusion of population characteristics (Π, ρ) and separation among datafied and structural categories allows us to identify the effect of extra-algorithmic interventions.

fairness constraint is equivalent to implementing a particular intervention on the causal model of figure 13. That intervention would be done specifically on τ. Other fairness constraints would also yield threshold classifiers and act similarly, hence we limit our analysis to DP and EO, as in [49]. Instead of comparing different types of fairness criteria, we are interested in comparing different types of interventions, algorithmic and systemic.

In previous work, [18] propose an SCM for this system, which we modify. Notably, previously π and ρ were treated as immutable characteristics of the population. We have incorporated them as endogenous variables. To understand this rationale let us consider each case carefully. Following our notation in Section 2 of the paper, a system $S = (O, R)$ is a tuple of objects and relations among these objects. When we consider the lending system of [49] in isolation, we don't have intervention access to the objects and relations of the generative mechanism that produced in the first place the score distribution π and repayment probability ρ, that is, the social, historical, and economic dynamics influencing the determination of the specific credit scores of individual members of a population. We note that these include the obligatory relationship among subjects of the state and governmental and financial institutions. We also note that two members of the same population may be meaningfully distinct but have been clumped together in the name of datafied race.

Since we don't have access to those precursor mechanisms, we may think of π and ρ as exogenous to the model. We then have two options; we can consider them as part of the functional relationship of (X, Y) to their corresponding exogenous variables, or we can set them as exogenous variables themselves. As an example, consider π and let F_π be its cumulative distribution function, then by the inverse probability integral transform we have $X = F_\pi^{-1}(U_X)$, where U_X is a uniform random variable on $[0, 1]$. In this way, π determines the relationship between X and the exogenous variable U_X. One can also consider π and ρ as exogenous variables themselves, in which case all analysis in the SCM is conditioned on particular values for π and ρ. In this case, the SCM models the bold region in figure 12. If we want to model the supersystem, we need to specify all its objects and structural relationships among them in the model. To completely do so, we need a model for the mechanisms generating π and ρ in the first place. The construction of such a model is beyond the scope of this paper, and would need to be carried out by social scientists with expertise in this area. Nevertheless, we can incorporate π and ρ into the model by considering them as endogenous variables, that is, by introducing Π and $\boldsymbol{\rho}$, two function-valued variables, for which we have observed the realizations π and ρ. This way, we can ask counterfactual and intervention questions regarding other stages of the system. We can ask, for example, what would the final outcomes be had the score distributions been different.

We have also separated $(\Pi, \boldsymbol{\rho})$ giving the population and the bank each their own variable. This is to allow for the cases in which we can intervene in one and not the other, or for which the information the bank has is outdated and lags the true distribution of a population.

Multi-stage model. In reality, we consider the behavior of this system over time, since discrimination is often encountered as a compound effect in which case feedback is present [20]. In particular, we use the new credit scores X^+ to update π at the next stage. The corresponding SCM diagram for the first two stages is shown in figure 14. The diagram becomes unwieldy so we do not diagram further stages but they are repeated versions of the second one. We did not consider the case in which ρ is updated, although it certainly is a possibility. It is also possible to update respective values for the bank and create a new classifier at each time step, but we did not explore this scenario and leave it to future work.

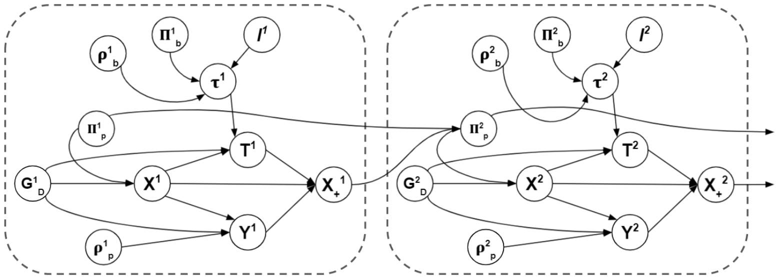

Fig. 14: Diagram for the dynamic model of the lending system. After the credit scores of a batch of applicants at time t has been updated according to the classifier decision and if they defaulted or repaid, the credit score distribution of their respective population is updated according to equation 3. Only the first two time stages are shown since all subsequent stages are exact replicas of the second one.

To update π we use the following equation:

$$\pi^{(t+1)} = r_\pi \pi_+^{(t)} + (1 - r_\pi)\pi^{(t)}, \tag{3}$$

for $r_\pi \in [0, 1]$, and where $\pi_+^{(t)}$ is just the empirical distribution of $X_+^{(t)}$:

$$\pi_+^{(t)}(x) = \frac{1}{N} \sum_{i=1}^{N} 1(X_i^+ = x).$$

We have actually added a small number ϵ and renormalized to avoid zeros. What this means is we mix the distribution of scores at time t with the distribution of the updated scores of the applicants post-classification. The ratio r_π represents, in a sense, the percentage of applicants from the superpopulation.

Note that instead of following the same individual over time, we recalculate the score distribution and sample from the updated π. This is because in reality new people apply for loans, as opposed to the same people reapplying multiple times. This is the case, for example, with student loans, in which new students apply for loans while previous generations strive to pay theirs back.

Running the model. A fairness metric is chosen according to $\mathcal{I}$. In our case, we intervene on this variable and set those values by hand. The classifier is then derived according to [49] and with respect to the chosen fairness constraint.

Population size is denoted by N. G_D is sampled from a categorical distribution $p_G = [r_A, 1 - r_A]$. An estimate of p_G is provided in the data from [49] and it is around $[.12, .88]$. π and ρ can be sampled from Π and $\boldsymbol{\rho}$ and estimates of one such instance are provided in the data from [49]. Here, we only consider situations where we specifically set the values for π and ρ as described in the interventions of section 4 in the main paper. We sample X and Y from $\pi_{pop,g}$ and $\rho_{pop,g}$, respectively.

For classification, T is sampled according to $\tau(x)$. If an individual doesn't get a loan we set $X^+ = X$, if they get the loan and repay then $X^+ = X + c_-$, while if they default $X^+ = X + c_+$. Finally, π is updated as above and the process repeated.

We have chosen $N = 2000$ applicants, with a ratio of applicants to the overall population $r_\pi = .2$. This ratio is conservative as there are situations (for example, student loans) where it is much higher. Following [49] we have chosen a score penalty of $c_- = -150$ and a score reward of $c_+ = 75$, as well as a profit loss in the case of defaulting of $u_- = -4$ and a profit gain in the case of repayment of $u_+ = 10$.

Finally, we recall the types of interventions:

Intervention 1: $f_\tau \leftarrow f'_\tau$.

The first intervention explores modifying only the classifier through fairness constraints.

Intervention 2: $\pi_A \leftarrow \pi_B$.

The second intervention explores the case in which credit score distribution changes. We combine this intervention with different instances of Intervention 1 (MP, DP, EO).

Intervention 3: $\rho_A \leftarrow \rho_B + \delta$, where $\delta := b - \rho_A(c^*)$, c^* is such that $F_{\pi_A}(c^*) = F_{\pi_B}(c_B)$ for F_π the cumulative distribution function, c_B was defined in equation 2, and $b = u_-/(u_- - u_+)$ is what [49] call a "break-even" quantity. We will also combine this intervention with fairness metrics.