Models for understanding and quantifying feedback in societal systems

ANONYMOUS

When it comes to long-term fairness in decision-making settings, many studies have focused on closed systems with a specific appointed decision-maker and certain engagement rules in place. However, if the objective is to achieve equity in a broader societal system, studying the system in isolation is insufficient. In a societal system, neither a singular decision maker nor defined agent behavior rules exist. Additionally, analysis of societal systems can be complicated by the presence of feedback, in which historical and current inequities influence future inequity. In this paper, we present a model to quantify feedback in social systems so that the long-term effects of a policy or decision process may be investigated, even when the feedback mechanisms are not individually characterized.

We explore the dynamics of real social systems and find that many examples of feedback are qualitatively similar in their temporal characteristics. Using a key idea in feedback systems theory, namely *proportional-integral-derivative* (PID) feedback and control, we propose a model to quantify three types of feedback. We illustrate how different components of the PID capture analogous aspects of societal dynamics such as the persistence of current inequity, the cumulative effects of long-term inequity, and the response to the speed at which society is changing. Our model does not attempt to describe underlying systems or capture individual actions. It is a system-based approach to study inequity in feedback loops, and as a result unlocks a direction to study social systems that would otherwise be almost impossible to model and can only be observed. Our framework helps elucidate the ability of fair policies to produce and sustain equity in the long-term.

Additional Key Words and Phrases: feedback, inequity, societal systems

ACM Reference Format:

Anonymous. 2022. Models for understanding and quantifying feedback in societal systems. In *FAccT '22: ACM Conference on Fairness, Accountability, and Transparency, June 21–24, 2022, Seoul, South Korea.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Why have inequities persisted for so long, despite years of activism, education, policy changes, and society's stated values of equity and non-discrimination? For example, segregation and inequities in housing and employment have persisted in the U.S. despite decades under the Fair Housing Act and Equal Employment Opportunity laws [7]. The answer, in our view, is rooted in the phenomenon of *feedback*. In every system in which inequity persists over time, there are feedback mechanisms which enable it to survive – as *1984* posits, "The object of power is power" [49]. Conversely, activism, public pressure, and equitable policies are used to push toward equity – Frederick Douglass said "If there is no struggle, there is no progress" [10] – and these can be seen as reactions to historical and present inequality, and thus are also a type of feedback.

This paper argues that feedback modeling tools from systems theory are helpful in quantitatively modeling mechanisms of feedback that help perpetuate and combat inequity. Good models help us gain more understanding into the

Manuscript submitted to ACM

mechanisms maintaining the status quo, predict how inequity will change over time, and help design new policies and algorithms which "produce and sustain equity" [30] when deployed in the real world. However, the economy and dynamics of power are complex, and we do not intend to model the feedback mechanisms individually and in their full complexity (as is attempted in system dynamics [24]). Instead, we focus on inequity at a systems level, essentially from the outside of a black box, both maintained and diminished over time by feedback mechanisms which quantify how much it will change or stay stationary. What is the benefit of such a model? For one, we can use system identification tools to find quantitative estimates for each type of feedback, and compare the amount of feedback by type in different systems. Further, we can use the model to predict future inequity. Finally, new policies and algorithms which influence future inequity can be modeled as having feedback mechanisms which operate in parallel with those in society, and their impact on equity estimated, and compared against other possible policies and algorithms.

One of the most extensively studied feedback systems in systems theory is the PID framework [31], consisting of proportional, integral and derivative forms of feedback. We argue in this paper with extensive examples and a formal analysis that this simple framework for feedback captures a wide variety of societal feedback mechanisms. The framework itself is simple – requiring only a few parameters – and this simplicity is both a value in our ability to interpret what model parameters are telling us about the system, and also situates in a favorable place in the tradeoff between model complexity and the need for large amounts of training data. Since society changes over time, it is useful to have a model whose parameters can be accurately estimated with as little history as possible.

Feedback is not merely a systemic response to how a system evolves over time. In societal settings, policies and interventions form another form of feedback. When new algorithms operate in parallel with societal mechanisms, as depicted in Figure 1, we can model their combined impact as well and predict how they impact our trajectory towards equity. We can also address questions of how different algorithms compare in terms of equity production.

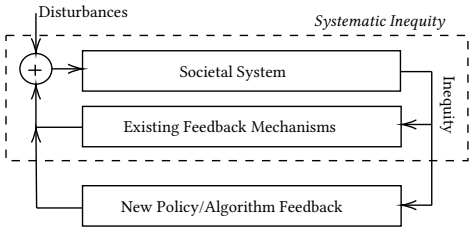

Fig. 1. Current societal systems have feedback mechanisms which make group inequities persist over time. New policies & algorithms can have feedback mechanisms that act in parallel, altering the inequity over time.

An Example: Gender Pay Gaps. The pay gap between men and women is the result of a complex system, impacted by cultural gender roles, biases in the workplace, occupational segregation, and more [2]. While the gender pay gap in the U.S. has reduced over the past 50 years, progress has slowed in recent decades [13]. We note three aspects of the pay gap. First, the current state of inequity is reported on and publicised every year as the pay that an average full-time working woman earns per dollar compared to the average full-time working man[1]. Next, there is a long-term historical inequity, proportional to the sum of this ratio over time, which is the gap in earnings over the lifetime of an average woman retiring today. Finally, there is the short-term change, the change-over-year in the inequity ratio, that represents how fast or slow we are moving towards equity as a society based on this statistic. We refer to these three

[1] https://www.aauw.org/resources/article/equal-pay-day-calendar/

aspects of the pay gap as the *proportional* (current state), integral (historical or long-term), and *derivative* (change). We can identify policies that provide feedback proportional to each term. For example, the salary history question on job applications contributes to keep future salaries (and inequity) close to current salaries and inequity [1], and thus is a proportional feedback mechanism. Gendered career roles are built over time – you are less likely to pursue a career if you don't see many people like you in that career – which is related to the inequity over a lifetime of people who entered that field. We thus see this as an integral mechanism. However, as women enter at higher rates into a profession previously dominated by men, the wages in that profession decrease [32]. Further, reactionary political movements use people's resentments about lost privilege to gain power and reverse policies that helped to reduce the gap [41]. We can see these as derivative mechanisms because the more progress towards equity that is made, the more each effect increases pay inequity.

1.1 Our Contributions.

We can summarize the contributions of this paper as follows.

- We present a method for modeling feedback in societal systems based on the PID framework from (linear) systems theory.
- We demonstrate with an extensive list of examples the ways in which the PID framework effectively captures real-world examples of moves towards (and away from) equity.
- We demonstrate the working of this model using three case studies involving historical and persistent inequity.
- We demonstrate how the model can be used to evaluate the effects of policy shifts and interventions.

2 TYPOLOGY OF FEEDBACK IN SYSTEMS OF INEQUITY

How do systems of inequity maintain themselves over time? Why doesn't awareness of the impacts on people due to systemic racism, sexism, heterosexism, transphobia, ableism, and classism make societies rectify the disparities immediately? We ask this question not to be naive, but to help us explicitly consider the range of mechanisms that maintain privilege and oppression over time.

We list 19 specific examples of feedback mechanisms in Table 1 which either maintain society's inequity or help to reduce inequity over time. We consider well-publicized inequities in education, employment, criminal justice, political representation, housing, and income. Although many more mechanisms exist, we attempt to give examples that describe a variety of feedback types. In particular, we posit in the rightmost column how next year's inequity is related to current and past inequity:

- *Proportional*: future inequity is a function of current inequity (rows 1-6),
- *Integral*: future inequity is a function of a sum of historical inequity (rows 5-15), or
- *Derivative*: future inequity is a function of the current change in (slope of) inequity (rows 4 and 15-19).

We note that future inequity due to one mechanism can be a function of multiple feedback types. We also note that feedback mechanisms are from different sources, including government policies (rows 5, 7, 11, 18), organization policies (rows 8, 9, 10, 17), laws (rows 2, 14, 19), algorithmic decision system (rows 5, 10, 11, 13), economic rule (rows 7, 12), activism (rows 3, 6), or human psychology (rows 4, 9, 15, 16).

While our analysis of patterns of feedback is systemic and "in the aggregate", we note that the impetus to resist change, or even to adopt particular modes of change that may be ineffective, are often rooted in well-studied group dynamics that typically contribute to the derivative element of feedback. These include a) *backlash*: "The resistance of

Table 1. Policies, algorithms, laws, and activism provide feedback in multiple societal systems which exhibit inequities, both to maintain and push back against oppression over time.

#	System	Description	Equity?	Mechanism
1	Education	Edu. software is fed unequal student data from oppressive educational contexts; tracking and 'at-risk' labelling keeps students stuck in their current track [37].	Anti	Proportional
2	Employ-ment	The US EEOC 4/5 rule allows legal remedy if, in part, the policy exhibits >20% disparity in hiring within a protected group.	Pro	Proportional
3	Surveil-lance	Publicity about large inequities in facial recognition by race and gender led to reduced disparities from targeted products [51]	Pro	Proportional
4	Income, Wealth	Support for progressive tax policy change can *decrease* when observing inequity, (e.g., from observing an unhoused person [55]) due to *belief in a just world*.	Anti	Proportional, Derivative
5	Criminal Justice	Since denying parole increases the rate of re-offending after release [65], the current & past racial inequity in parole leads to future inequity in re-offense.	Anti	Proportional, Integral
6	Criminal Justice	The #BlackLivesMatter movement was spurred both by specific incidents of violence and long-term *systemic* violence against Black people [58, 61].	Pro	Proportional, Integral
7	Housing, Wealth	The effects of discriminatory housing policies accumulate over time via lower property value growth in Black and Latinx neighborhoods, which also leads to mortgages with worse terms.	Anti	Integral
8	Higher Ed	Inequity of people admitted to college today will have an effect decades into the future via legacy admits [9].	Anti	Integral
9	Employ-ment	Discrimination in a profession over decades means that there are few examples of a minoritized group in that profession, which then makes members of that group feel less welcome in that profession.	Anti	Integral
10	Employ-ment	Automated hiring models use data from the history of past hires, thus may learn to repeat past discrimination [4].	Anti	Integral
11	Criminal Justice	Future police allocation to an area, and thus future discovered "incidents", is proportional to the cumulative history of incident reports [14].	Anti	Integral
12	Income, Wealth	Excess income (above consumption) adds to wealth in a cumulative sum over time, & earns money on itself (the gross rate of return on wealth) [36].	Anti	Integral
13	Health Care	Algorithms that allocate medical resources to reduce costs assign fewer re-sources to racial groups who historically received unequal treatment [47].	Anti	Integral
14	Income	The Lilly Ledbetter Fair Pay Act of 2009 allows lawsuits for wage discrimination at one's employer over one's entire career.	Pro	Integral
15	Income, Wealth	Exposure to historical data about *rising* wealth inequality in the U.S. tends to increase support for redistributive policies [40].	Pro	Integral, De-rivative
16	Income	As women become a higher percentage of a profession, employers reduce pay to that profession and value it less [32].	Anti	Derivative
17	Higher Ed	The DIF (in SAT future test planning) ensures slow change in race & gender gaps, rather than increasing the use of questions which defy those gaps [56].	Anti	Derivative
18	Voting Rights	Politicians can keep their power despite a changing population by redrawing their district boundaries to include people more likely to vote for them.	Anti	Derivative
19	Voting Rights	Roberts: Voting inequity exists today but is less than it was in the past, thus protection by the Voting Rights Act is not justified [53].	Anti	Derivative

those in power to attempts to change the status quo is a 'backlash', a reaction by a group declining in a felt sense of power" [39]; b) *reactance*: the pushback when people are confronted with threats to their freedom [5], including not being allowed to discriminate; and, in the case of racial inequity, c) color-blindedness: "Liberalism's very aspirations to color-blindness & equality – while admirable – can impede its goals, as they prohibit race-conscious attempts to right historical wrongs", i.e., change is slowed by them [48].

The purpose of Table 1 is to provide many examples of systems of inequity which are maintained and challenged in ways that can be modeled with proportional, integral, and derivative feedback. These mechanisms include existing and potential policies and algorithms, but any change in equity induced by their use would be subject to the other feedback mechanisms of that system. This paper provides a simple quantitative model for systemic feedback mechanisms that could be useful in analyzing changes.

Finally, we note what is left out of Table 1. There are multiple unpredictable ways in which inequity changes over time, e.g., due to a pandemic. In controls theory, this is called the *disturbance* or *process noise*, adding to the feedback as depicted in Figure 1, which is distinct from the errors in measuring inequity (which is referred to as *measurement noise*). Analyzing the stability of control mechanisms to keep inequity at a desired *set point* requires particular assumptions about the noise terms. One could imagine designing policies and algorithms which produce the required feedback, in parallel to other societal feedback mechanisms, to maintain a society at equity, but we do not attempt to study this here, since our purpose is primarily to provide a proof-of-concept of a type of feedback model.

3 DYNAMICAL PID STATE MODEL

We propose a model to quantify proportional, integral, and derivative (PID) feedback mechanisms in systems with inequity. Each PID term is represented as a state variable in the model and is incorporated as feedback on the future state of inequity.

Proportional. Equity is achieved when a societal system produces equal statistics across groups, for example: "racial equity is a state in which race no longer predicts outcomes" [15]. We measure inequity at time n as:

$$x(n) = \frac{\text{outcome measure for people in group A}}{\text{outcome measure for people in group B}} - 1, \tag{1}$$

where the "outcome measure" is a societal measure that should be equal across groups if equity is achieved. At equity, the value of $x(n)$ is 0. We choose the group in the numerator so that $x(n)$ is historically above 0, so that readers can consistently interpret $x(n)$ as 'inequity', and work to *reduce* $x(n)$ to 0 as pro-equity. For example, in 1964, from U.S. Census statistics, 70.7% of white Americans (which we set as group A) voted, and 58.5% of Black Americans (group B) voted [63], for a ratio of 1.209 and thus $x(n) = 0.209$. Our choice to subtract 1 in (1) is to ensure that minimizing $|x(n)|$ is a desirable goal.

Integral. The cumulative history of inequity is captured by the integral term, $\sigma(n)$. We choose to weigh the most recent history more heavily than the distant history by using an autoregressive filter. Then, the cumulative inequity at time n is given by:

$$\sigma(n) = x(n-1) + \alpha \cdot \sigma(n-1) \tag{2}$$
$$\text{with } 0 < \alpha < 1$$

The filter has an infinite impulse response, meaning weights in the cumulative sum will never be completely reduced to zero, but are instead proportional to a factor of α^n at time n. We discuss α further in Section 3.1.

Derivative. The derivative term at time n is the difference in inequity at time n and inequity at time $n-1$:

$$\dot{x}(n) = x(n) - x(n-1) \tag{3}$$

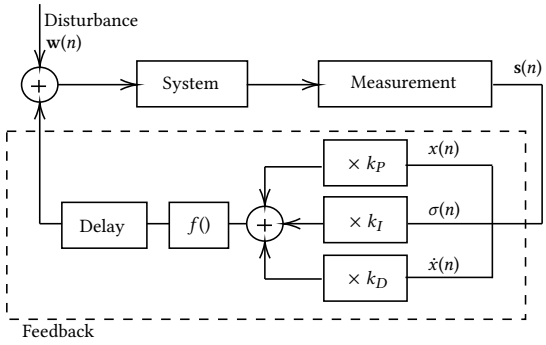

Fig. 2. A feedback model for societal control on a system with inequity $x(n)$, in which the control is linear in $x(n)$ itself (the proportional term), in its weighted sum $\sigma(n)$, and in the derivative $\dot{x}(n)$.

State Model. We define the state of our model to be the current proportional, integral, and derivative terms at time n, $\mathbf{s}(n) = [x(n), \sigma(n), \dot{x}(n)]^T$. The PID terms are incorporated as feedback in our model, as shown in Figure 2. We describe any other changes to the inequity that is not feedback from the system's outputs as part of the disturbance $\mathbf{w}(n) = [w_0(n), w_1(n), w_2(n)]^T$. We model the dynamics as linear, that is, a weighted sum of the three PID components of the inequity, as well as the disturbance. While it is possible to include non-linearities in the dynamical equations by including an arbitrary function f in the loop as shown in Figure 2, in this paper, we let $f(x) = x$ for simplicity. Then we model the societal feedback as a linear function of these terms:

$$\mathbf{k}^T \mathbf{s}(n) = k_P x(n) + k_I \sigma(n) + k_D \dot{x}(n), \tag{4}$$

where $\mathbf{k} = [k_P, k_I, k_D]^T$, which are the constants which describe how the state evolves. This linear sum, $\mathbf{k}^T \mathbf{s}(n)$ then adds to the current state, specifically, the slope $\dot{x}(n+1)$ at the next time $n+1$ is calculated as the current slope $\dot{x}(n)$ plus this feedback $\mathbf{k}^T \mathbf{s}(n)$ plus some disturbance:

$$\dot{x}(n+1) = \dot{x}(n) + \mathbf{k}^T \mathbf{s}(n) + w_2(n), \tag{5}$$

where $w_2(n)$ is the slope disturbance. The system also progresses by: 1) adding the current slope into the inequity for the next time, and 2) keeping track of the cumulative inequity by adding the current inequity to $\sigma(n+1)$. These state update equations are thus:

$$\begin{aligned} x(n+1) &= x(n) + \dot{x}(n) + w_0(n) \\ \sigma(n+1) &= x(n) + \alpha \cdot \sigma(n). \end{aligned} \tag{6}$$

These equations (5) and (6) implement the proportional, integral, and derivative feedback terms as described and justified prior. In short we can write

$$\mathbf{s}(n+1) = A\mathbf{s}(n) + \mathbf{w}(n) \tag{7}$$

where

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & \alpha & 0 \\ k_P & k_I & k_D + 1 \end{bmatrix}. \tag{8}$$

We note that we don't measure all of the state variables at each time, as we only measure directly the inequity $x(n)$ at each time n. Further, we measure only a noisy version of the inequity $x(n)$. We assume additive noise. To match the typical notation from linear systems, we write $\mathbf{s}(n)$

$$y(n) = C\mathbf{s}(n) + v(n), \tag{9}$$

where $C = [1, 0, 0]$ and $v(n)$ is the measurement noise at time n.

How feedback is incorporated. In equation (5) the feedback is assembled and used to update the *derivative* $\dot{x}(n)$. One might reasonably ask whether the feedback should be used instead to update the current state $x(n)$. However, our method estimates parameters $\mathbf{k}$ to best match the data, as we describe next in Section 3.2, and we find we end up with the exact same predictions from such a model as the one we give here, so it is just a matter of perspective. We find it more intuitive to think about feedback updating the slope, similar to how cruise control updates the acceleration (the derivative of speed) in its control of a vehicle's speed.

3.1 Autoregressive Time Constant Selection

To select a value of the parameter for the autoregressive filter on the integral portion, we choose an appropriate time constant, τ and calculate α as in (10).

$$\alpha = e^{-\frac{1}{\tau}} \tag{10}$$

The selection of τ is domain-specific and should be based on a reasonable estimate of the time for the impact of historical inequity to decay. For example, if we want terms in the cumulative sum to decay in 10 years and there is one time step per year, $\alpha \approx 0.9$. If we want the terms in the cumulative sum to decay in 100 time steps, then $\alpha \approx 0.99$.

3.2 Model Parameter Estimation

Given the dynamic model in (7) and a set of longitudinal data for $\{y(n)\}_n$, we want to estimate what parameters of the model would explain its temporal dynamics. As stated, there is noise in the measurement, and there are disturbances that contribute to the state that are not explained by the PID feedback model. How do we select values for the parameters $\mathbf{k}$ and α from a longitudinal data set?

We provide one method here. Some systems identification methods estimate the entire update matrix A from (8), but for our purposes, we only estimate the $\mathbf{k}$ parameters. For our three $\mathbf{k}$ parameters, k_P, k_I, and k_D, we derive a least squares estimator as follows. We define a vector $\Delta\mathbf{s}(n) = \mathbf{s}(n+1) - \mathbf{s}(n)$. An equation for $\Delta\mathbf{s}(n)$ can be written by subtracting $\mathbf{s}(n)$ from both sides of (7):

$$\Delta\mathbf{s}(n) = (A - I)\mathbf{s}(n) + \mathbf{w}(n), \tag{11}$$

where I is the 3x3 identity matrix. Focusing on the 3rd row of the vector $\Delta\mathbf{s}(n)$, since it is the one element that is a function of the unknown $\mathbf{k}$ parameters,

$$\dot{x}(n+1) - \dot{x}(n) = \mathbf{k}^T \mathbf{s}(n) + w_2(n). \tag{12}$$

Defining $\ddot{x}(n) = \dot{x}(n+1) - \dot{x}(n)$, we can then say that:

$$\begin{aligned} \ddot{\mathbf{x}} &= S\mathbf{k} + \mathbf{w}_2, \text{ where,} \\ \ddot{\mathbf{x}} &= [\ddot{x}(1), \ldots, \ddot{x}(N)]^T \\ \mathbf{w}_2 &= [w_2(1), \ldots, w_2(N)]^T \\ S &= [\mathbf{s}(1), \ldots, \mathbf{s}(N)]^T \end{aligned} \tag{13}$$

where we have recorded data from time $n = 0$ to $N + 1$. We could estimate $\mathbf{k}$ in multiple ways, but one easy way would be to use a least-squares approach. Defining superscript + to indicate the pseudoinverse of a matrix,

$$\hat{\mathbf{k}} = (S^T S)^+ S^T \ddot{\mathbf{x}}. \tag{14}$$

We note that this estimate is the maximum likelihood estimate in a Gaussian noise case. In short, if we have the full state $\mathbf{s}(n)$ for all times n, we can form the matrix S and vector $\ddot{\mathbf{x}}$ and compute an estimate for $\hat{\mathbf{k}}$.

However, we don't start out with a known state – we only measure $y(n)$ at all times. Thus it is necessary, in order to estimate the parameters $\mathbf{k}$, to first estimate the state $\mathbf{s}(n)$ for all time n. This creates a chicken-and-egg question. A standard approach is to use an expectation maximization (EM) approach to alternately 1) calculate the expected value of the sequence of states $\{\mathbf{s}(n)\}_n$ for all n, and then 2) find the system parameters which maximize the likelihood given the calculated states [17]. In our case, this second part is calculated with (14). The first part is described in Section 3.3.

3.3 State Estimation

Since we do not measure the state directly or in the absence of noise, our model says that we don't know exactly what the current inequity is, or its slope or cumulative sum. Given a historical set of data measuring the inequity, and known parameters $\mathbf{k}$, we use a Bayesian smoother to estimate the state [57]. We denote this state estimate as $\hat{\mathbf{s}}(n)$ for $n \in \{1, \ldots, N\}$.

As described above, from the state estimates we calculate the change in slope $\ddot{\mathbf{x}}$ which we use with the state S in (14) to re-estimate $\mathbf{k}$. We iterate this algorithm until convergence, which we note in practice takes less than 10 iterations.

4 EXPERIMENTS

We test our model on the following real-world datasets:

(1) *Earnings, Men vs. Women*: The inequity between men and women's earnings is commonly referred to as the gender pay gap, although we note that we do not have a data set inclusive of other genders. For the U.S., we use annual data from 1960 to 2018 [45]. Compiled from U.S. Census data, the data refers to the ratio of median income between men and women full-time, year-round workers. In 2018, women workers' median pay was $0.82 per dollar of men workers' median pay. Equivalently, we use the inverse, that is, median pay for men divided by the median pay for women, or 1.22, and subtract 1 to obtain an inequity of 0.22.

(2) *Voting, White vs. Black*: The percentage of white people who voted divided by the percentage of Black people who voted in the U.S., according to data collected by the U.S. Census Bureau [63]. This data is for national congressional or presidential elections, i.e., every even year, since 1964.
(3) *Income, Top 10% of Earners vs. 10% of All Income*: We take the total income of people in the top 10% by income and divide it by 10% of the sum of the income of all people in the U.S. This value is thus a ratio of how much more the people in the top 10% are paid than they would if income was split evenly among all people. The data comes from U.S. tax data collected by Piketty and Saez [50, 54].

We consider the following experimental questions:

- How well does the model predict future inequity?
- How can one interpret the model parameters?
- When does the model perform poorly?
- How does the model compare to existing simple models?
- How can we use the model to predict the impact of new policies or algorithms?

All code and data for the experiments can be found at this repository, which we attempted to anonymize for anonymous review: https://anonymous.4open.science/r/feedbackModeling-968E/.

4.1 Future Inequity Prediction and Parameter Interpretation

In this section, we divide the past into a training period and a test period in order to validate the model's extrapolation performance vs. real world changes in inequity in society. In other words, we essentially pick a threshold year for the purpose of evaluation; the model is trained on the data up to and including the threshold year, and the model then runs, starting with the next year through the present. Since we have data to the present (which was not used in the training) we can see how well the model predicts the "future".

Our results on our three data sets are shown in Figure 3. We test (in the left column) using the first half of data for training, and also (in the right column) using the first 2/3 of the data for training. The training is shown as a green solid line, and the actual reserved test data is shown with a green dashed line, and compared to a blue solid line for the model prediction.

We report the estimated model parameters in Table 2. We advocate for this model, in part, because the model parameters are interpretable as quantifying feedback types. As detailed in Equations (7) and (8), next year's slope, $\dot{x}(n+1)$, is k_P times the current inequity $x(n)$, plus k_I times the weighted cumulative sum of inequity, plus $k_D + 1$ times the current slope[2]. In all of our data sets, the inequity $x(n)$ is currently positive. Thus any $k_P < 0$ is pro-equity because it pushes the slope down (towards equity). Similarly, any $k_P > 0$ is anti-equity because it pushes the slope up (towards more inequity). Since we are looking at data with historical inequity, any $k_I < 0$ pushes the slope down toward equity, while $k_I > 0$ pushes the slope up and away from equity. However, the sign on the derivative term has a different effect. $k_D < 0$ indicates a push against the current change. If the current slope is negative, the effect of $k_D < 0$ is to slow down progress towards equity. In contrast, if $k_D > 0$, the derivative feedback reinforces the current direction of change in the system. For each model, we next describe the performance and interpret the parameters.

Earnings, Men vs. Women: The top row of Figure 3 shows predictions for the gender pay gap. We chose an integral term time constant α that corresponds to a decay time of 10 years. The model predicts the downward slope of the data closely when trained on the first half of the data set. When trained on the first 2/3, the model predicts the gender pay

[2]Note the +1 comes from the fact that the current slope stays the same in the absence of any feedback.

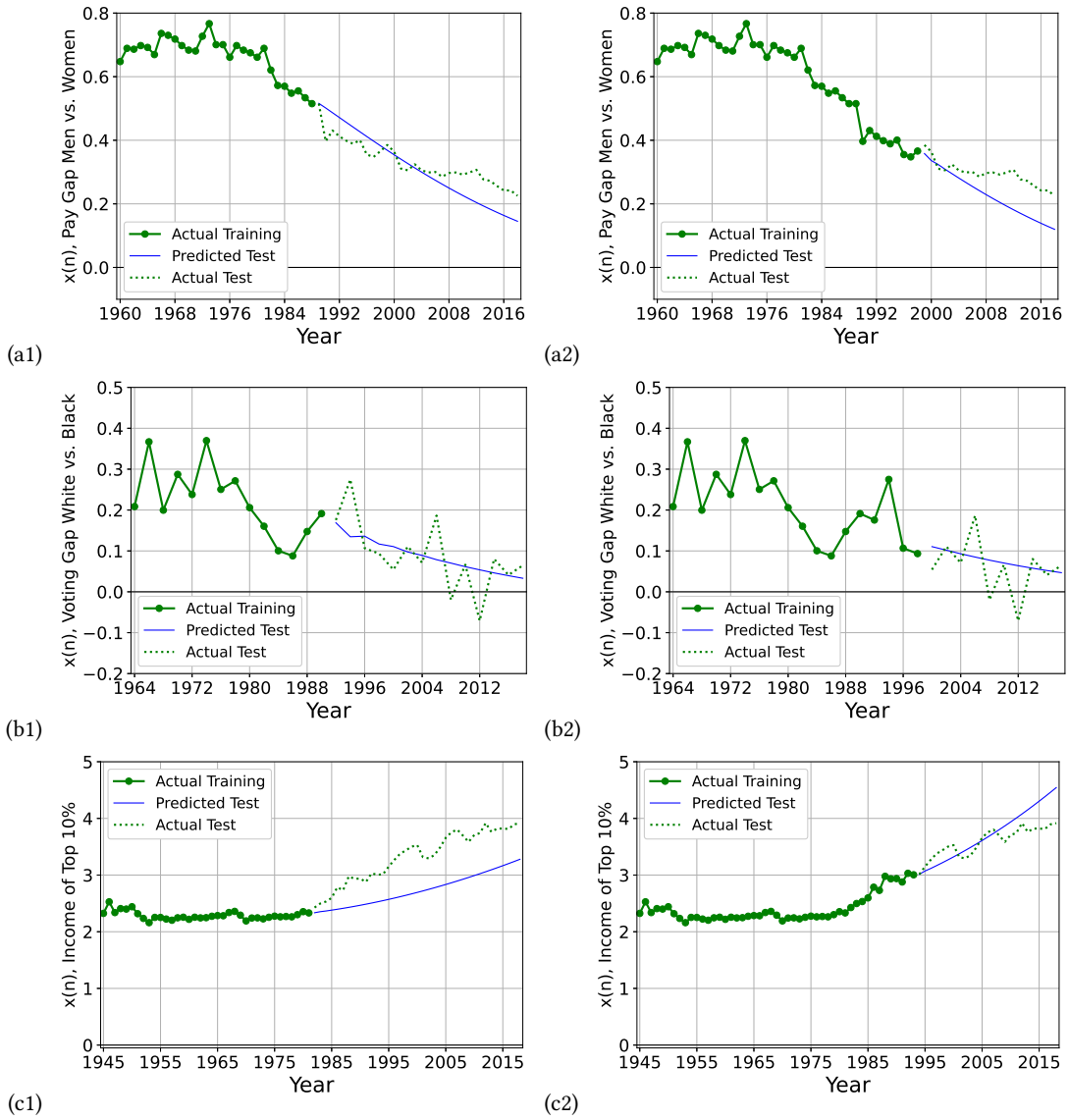

Fig. 3. Training the model only from the historical (Col #1) first 1/2 of data; (Col #2) first 2/3 of the data, we estimate PID model parameters. Then we extrapolate (simulate the model) to predict the remaining years, and compare to the actual test period data, when $x(n)$ is the U.S. (a) earnings inequity of men vs. women; (b) voting inequity of white vs. Black; (c) income inequity of top 10%. In all plots, $x(n)$ is as defined in (1), and a value of 0 (––) is equity.

gap to decrease more quickly than it actually did. Considering the parameters trained on the entire dataset, the positive sign on k_P serves to reinforce the current inequity, the cumulative effects of past inequity push toward lower inequity through k_I, and the negative sign on k_D acts in opposition to the current change. In other words, our model finds that

Data Set #	Name	τ (yrs)	Training Data Used	Proportional k_P	Integral k_I	Derivative k_D
1	Pay Gap Men vs. Women	10	First 1/2	0.0279	-0.0051	-1.07
			First 2/3	0.0372	-0.0066	-1.25
			All	0.0235	-0.0045	-1.08
			Second 1/2	-0.0173	-0.0012	-1.22
2	Voting Gap White vs. Black	20	First 1/2	-0.0511	-0.0075	-1.63
			First 2/3	-0.0729	-0.0033	-1.53
			All	-0.0656	-0.0062	-1.71
			Second 1/2	-0.0595	-0.5278	-0.96
3	Income of Top 10%	100	First 1/2	-0.0150	0.0008	-1.31
			First 2/3	-0.0218	0.0014	-1.25
			All	-0.0129	0.0006	-0.80
			Second 1/2	0.0160	-0.0005	-0.72

Table 2. PID model parameters estimated from training. Values are interpreted as: Next year's slope increases by k_P times the current inequity, increases by k_D times the current derivative, and increases by k_I times the current cumulative sum. All parameters with signs that *increase* inequity are red, those that decrease inequity are black, as detailed in Section 4.1.

only the cumulative income gap acts to push the gender earnings gap towards equity, while current pay differences and the decrease in the gap over time both serve as feedback against equity.

Voting, White vs. Black: For the voting gap, we chose a time constant such that there is a decay time of approximately 20 years. The voting gap data is particularly noisy, driven in part by different participation rates between presidential election years and non-presidential election years, as well as driven by particular candidates. For example, the 2008 and 2012 elections with President Barack Obama on the ballot had particularly high turnout among Black voters. Nevertheless, the shape of the model prediction closely matches the actual values in the test period. In the voting gap dataset, both k_P and k_I reduce inequity for every training set used, while the sign of k_D indicates a resistance to the decreasing inequity. Because the actual data oscillates election to election, i.e. a decrease in inequity one timestep is followed by an increase in inequity the next, the magnitude of k_D is generally large. Overall, increasing Black participation in voting (relative to white participation) is found to result in resistance.

Income, Top 10% of Earners vs. 10% of All Income: We predicted that the cumulative effect of past inequity would persist much longer into the future in the case of income inequality because excess income is accumulated without loss over time. Therefore, we selected a time constant of 100 years. Notably, this model does not perform as well when trained on the first half of data, i.e., a period of constant and low inequity from 1945-1981. It may be because the model parameters changed dramatically between the first half and the last half of the data set. We can observe, comparing the model parameters when trained on the "First 1/2" vs. on the "Second 1/2", that the proportional parameter switches from negative to positive, while the integral parameter switches from positive to negative. As stated previously, positive values of k_P mean that the slope increases (towards higher inequity) while there is current inequity. We interpret this as saying in the period 1945-1981, the current inequity *reduces* inequity, while the cumulative history of inequity increases inequity. In contrast, in the period 1982-2018, the dynamics of income inequality indicate that current inequity increases future inequity, while the cumulative history of inequity will work to reduce inequity. The smaller magnitude of k_I estimated by the second half of the data suggests that income inequality is less impacted by cumulative inequality and the smaller k_D indicates there is less resistance to rising inequality. These changes coincide with a wave of tax, benefit, and unionization policy changes starting in the US after 1980 [19]. These changes in the structure of the feedback from

the first half of the data mean that any prediction of the future that extrapolates from those first years is unlikely to be accurate.

4.2 How does the model compare to existing simple models?

In this section, we compare the PID model test predictions to those generated by other simple regression models that can be learned from a sequence of one-dimensional historical data. We calculate the root mean squared error for linear regression, polynomial interpolation, and decision tree regressors [3] and compare to the PID model. The results for each dataset and training set are shown in Figure 4.

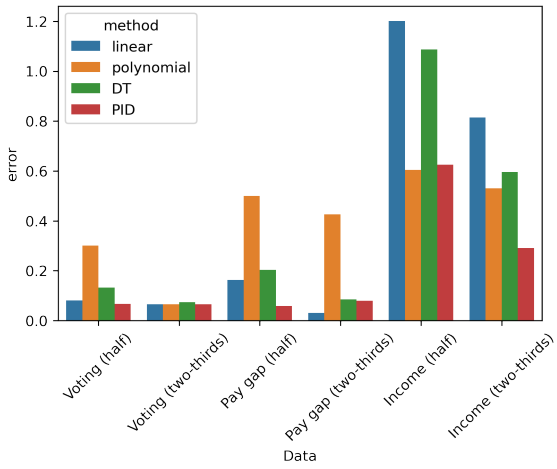

Fig. 4. RMS errors with respect to test data for each of the comparison methods as well as the PID model for each of the datasets and training combinations.

Overall, we find that although there is no single model with lowest error across all datasets and training options, the PID model has one of the lowest error values in general. Given this, as well as the interpretability and manipulatability of the model as described previously in this section, we believe that PID is a useful addition to the set of existing simple models.

4.3 What is the Impact of New Policies?

Consider the example of salary history being used to determine the offered salary for a candidate for a job. The salary history question is believed to perpetuate pay gaps as workers who are currently underpaid tend to get offered lower salaries when taking a new job. For example, consider a university that by policy limits the salary increase for in-university transfers to a maximum of 15%. For people without the privilege to start their career at a highly paid position (often women of color), climbing the ladder can lead to a higher position but, due to this policy, sometimes absurdly low pay compared to others in the same job.

When California banned employers from using salary history in 2018, it is estimated that this led to a 1% improvement in the gender pay inequality ratio over a synthetic control in the studied year [20]. It can be seen that the feedback effect

[3] All are implemented using sklearn's packages and default parameters, with degree of 3 chosen for the polynomial interpolation.

of the salary history question is proportional – current pay inequity leads directly to future pay inequity. Let us use this example in the PID framework to investigate the long-term effects of the policy. Let us define the PID model of the pay gap system using the parameter estimates from the entire dataset as shown in Table 2, $k_P = 0.0235$, $k_I = -0.0045$, and $k_D = -1.08$. We define the policy to have its own PID terms, k'_P, k'_I, k'_D. If we assume that the salary history ban policy effects only the proportional term, then we can assume $k'_I = 0, k'_D = 0$. To calculate k'_P, we consider the the additive effect of the policy parameters on the system parameters. In 2019, the system with the policy would have a 1% lower output than the system alone. Using (6), we derive an equation to solve for k'_P.

$$\frac{x(2018) + \dot{x}'(2018)}{x(2018) + \dot{x}(2018)} = 0.99 \tag{15}$$

where $\dot{x}'(n) = \dot{x}(n-1) + (k_P + k'_P)x(n-1) + k_I\sigma(n-1) + k_D\dot{x}(n-1)$. By substituting in the data on the gender pay gap, we find that $k'_P = -0.0536$.

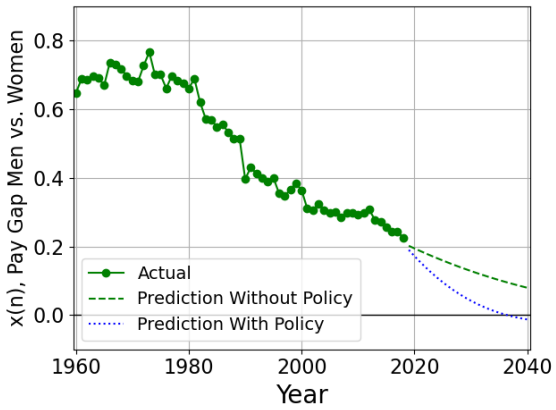

Fig. 5. Predicted gender pay inequity with the salary history ban in effect and without the salary history ban.

When we simulate the gender pay gap system with and without the salary history ban, we can see that the policy causes the system to approach equity much more quickly than the system without the policy. While the policy only causes a 1% decrease in inequity initially, over time, the effects of the policy are expected to be larger. However, it is important to note that the annual 1% improvement in the first year does not continue indefinitely. The PID model here naturally accounts for the societal feedback that, over time, acts against the initial change, such that the *slope* in year 2040 is approximately the same with or without the policy.

5 RELATED WORK

5.1 Long-term Fairness

Our work is situated within the study of long-term fairness effects in the presence of feedback, which now admits a growing literature. For more on the broader framework of sequential decision-making (and the associated feedback) see also the survey by Zhang and Liu [69] and the review article by Chouldechova and Roth [8].

In a general sense, much of the prior work on long-term effects of fairness has focused on a single decision system with somewhat explicit modeling of agent behavior. Prior work has either focused on two-stage pipelines (where one decision

causes a reaction followed by another) [23, 29, 33] or finite or infinite-horizon decision making [12, 21, 25, 34, 44, 68]. These approaches are primarily model-based. Other model-based approaches use Markov decision processes (MDPs) to capture agent behavior and use simulation techniques to analyze a system [11, 46]. MDPs can also be formally analyzed for long-term effects on group and individual fairness as explored by D'Amour et al. [11], Jabbari et al. [26], Joseph et al. [28], Wen et al. [66].

For a more 'model-free' approach, we turn to the effect of feedback in the context of predictive policing [14]. The interaction between predictive policing software and policing itself are analyzed using a discrete urn model, and the feedback is shown to be positive, i.e., resulting in divergence; Police end up vastly over-policing one neighborhood, regardless of the neighborhood crime rates [14]. The model in this paper adds complementary tools; it provides a continuous-valued model rather than a discrete-valued model, and it provides a connection to analysis methods within linear feedback systems theory [31] that help analyze dynamics and stability.

Model-free (or model-based) learning of systems that adapt is also the world of *reinforcement learning*. The survey by Recht [52] presents a beautiful overview of the connections between reinforcement learning and control.

5.2 Economic Models of Inequality

Most of the economic literature on inequality is about the relationship between growth and inequality. The literature refers to either political economy or wealth effect arguments [3] in which the economy is populated by a continuum of agents who are evolving over time (using agent-behavior modeling) to either maximize individual gain or to bring about economic growth. In addition, many such studies of inequality are built upon wealth distributions where some form of general-equilibrium or quantitative models with heterogeneous agents are in place [3, 6, 27]. Other models to forecast economic inequality require a concrete understanding of the macroeconomic explanatory parameters of the system. The model requires explanatory parameters to fit historical data and forecast future inequality. Examples of such parameters include human capital attainment, labor force indicators and macroeconomic indicators, e.g., GDP and inflation [18]. Note that the Lorenz Curves [35], the Gini coefficient [43], and Theil index [42] are some of the most well-known inequality measures, but are not models that can be used to predict the trajectory of future inequities.

In our approach, we do not need to have such detailed information about the macro-economic and explanatory parameters of the system (which might not even be available or extractable from the data). In addition, our view point is broader than individual-based optimization, allowing forecasting of the production and long-term sustainability of equity in a social system.

The area of "systems dynamics" applies feedback modeling to study the complex dynamical behaviors of economic and social systems, for example, the interaction between road construction, recycling, and mining [38]. Specific feedback mechanisms, including delays, differential and/or integral effects, are assumed to exist, and specified with each model. In model-building in social work, community engagement can be used to elucidate all of the possible feedback loops in the system [24]. Our paper is similar in that it mathematically models feedback with a systems approach, but we don't attempt to model each loop explicitly, but rather build a simplified model with historical data.

6 LIMITATIONS

We provide some discussion of the limitations of our model.

Portability trap. Are we falling into the "portability trap" [59]? We train our model for each domain / data set, and do not make assumptions about the particular structure of any one system of inequality. However, we are making model

assumptions that may not hold in every case — we do not anticipate that a linear feedback model will be sufficient, or that proportional, integral, and derivative terms are best to model the actual mechanisms that keep systemic inequality in place in every type of inequity.

Perception vs. Reality. We use measurements of inequity as the driver for societal feedback mechanisms. However, people's estimates of the level of inequity are inaccurate in the U.S. and U.K., and their estimates are heavily influenced by how much inequality they see locally [22]. If people support policies based on perceived inequity, and perceived inequity is not proportional to measured inequity, this could affect the accuracy of our model. Our model presumes that, society-wide, future changes in inequity are at some level influenced by present, historical, and past changes in measured inequity.

Multidimensionality of Oppression. We model only one inequity measure at a time. In reality, multiple factors contribute to the totality of oppression [16]. For example, the income gap leads to the wealth gap, which then increases the health equity gap. As another example, inequity in the educational system produces future inequity in the criminal justice system [67] via a mechanism called the school-to-prison pipeline. We can imagine extending the model to include multiple measures, and the feedback between different states, although the model complexity would grow.

Setpoint of Equity. In our model, the set point is equity, meaning that a system currently at equity with no historical inequity will not change unless there is a disturbance to the system. We use a set point of equity because we advocate for policies that "produce or sustain equity" between groups [30], but we recognize many people do not share a goal of equity [60]. We leave the development of models that do not have a set point of equity to future work.

7 CONCLUSION

We present arguments for, and methods to generate, a model for the feedback present in societal systems of inequity. Inequities in outcomes due to racism, sexism, classism, and other systems of oppression are preserved by feedback mechanisms which maintain the status quo, and are reduced by mechanisms which push to address disparities. We build a model with proportional, integral, and derivative feedback terms, and show how historical data can be used to estimate the model's parameters, which then quantify how much of each type of feedback exists in society. We use the model to predict future trajectories of the inequity and compare our model to alternatives, and show that the error is generally lower than other simple modeling methods. The parameters represent the particular mechanisms, which if changed, would quantitatively alter the trajectory. The model thus introduces a connection between linear systems theory and systemic oppression which could be useful in the modeling and analysis of policy and other mechanisms designed to address social inequity.

Answering the question, "when will we reach equity?" is not just an exercise. U.S. Supreme Court Justice Sandra Day O'Connor, writing the majority opinion in Grutter v. Bollinger that preserved affirmative action, wrote that the "Court expects that 25 years from now, the use of racial preferences will no longer be necessary" [64]. That was 19 years ago; racial inequity in college admissions persists, and O'Connor has since said "That may have been a misjudgement" [62]. But we hope that the model we introduce can do more than provide a likely imperfect window into the future. By explicit reporting of the proportional, integral, and derivative terms, we posit that those seeking equity and those interested in projecting into the future may be better able to reason about the relative impacts of current inequity, longstanding and accumulated inequity, and resistance to – or support for – change.

REFERENCES

[1] Torie Abbott Watkins. 2018. The ghost of salary past: Why salary history inquiries perpetuate the gender pay gap and should be ousted as a factor other than sex. *Minn. L. Rev.* 103 (2018), 1041.

[2] Christine Alksnis, Serge Desmarais, and James Curtis. 2008. Workforce Segregation and the Gender Wage Gap: Is "Women's" Work Valued as Highly as "Men's"? *Journal of Applied Social Psychology* 38, 6 (2008), 1416–1441.

[3] Abhijit V. Banerjee and Esther Duflo. 2003. Inequality and growth: What can the data say? *Journal of Economic Growth* 8, 3 (2003), 267–299.

[4] Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias. (2018).

[5] Sharon S. Brehm and Jack W. Brehm. 2013. *Psychological reactance: A theory of freedom and control.* Academic Press.

[6] Marco Cagetti and Mariacristina De Nardi. 2008. Wealth inequality: Data and models. *Macroeconomic Dynamics* 12, S2 (2008), 285–313.

[7] Nancy Cambria, Paul Fehler, Jason Q. Purnell, and Brian Schmidt. 2018. *Segregation in St. Louis: Dismantling the Divide.* Technical Report. Washington University in St. Louis.

[8] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[9] Ezekiel J. Dixon-Román. 2017. *Inheriting Possibility: Social Reproduction and Quantification in Education.* University of Minnesota Press.

[10] Frederick Douglass. 1857. *West India Emancipation Speech.*

[11] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. 525–534. https://doi.org/10.1145/3351095.3372878

[12] Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna Gummadi, and Patrick Loiseau. 2019. The Price of Local Fairness in Multistage Selection. *arXiv preprint arXiv:1906.06613* (2019).

[13] Paula England, Andrew Levine, and Emma Mishel. 2020. Progress toward gender equality in the United States has slowed or stalled. *Proceedings of the National Academy of Sciences* 117, 13 (2020), 6990–6997.

[14] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*. PMLR, 160–171.

[15] Forward Through Ferguson Commission. [n.d.]. A Path To Racial Equity Worksheet. https://forwardthroughferguson.org/get-involved/pathtoracialequitytool/ accessed 18 Jan 2022.

[16] Marilyn Frye. 1983. Oppression. In *The Politics of Reality: Essays in Feminist Theory*. The Crossing Press, 1–16.

[17] Zoubin Ghahramani and Geoffrey E. Hinton. 1996. *Parameter estimation for linear dynamical systems.* Technical Report. Technical Report CRG-TR-96-2, University of Toronto, Dept. of Computer Science.

[18] Marina Gindelsky. 2018. Modeling and Forecasting Income Inequality in the United States.

[19] Jacob S. Hacker and Paul Pierson. 2010. Winner-take-all politics: Public policy, political organization, and the precipitous rise of top incomes in the United States. *Politics & Society* 38, 2 (2010), 152–204.

[20] Benjamin Hansen and Drew McNichols. 2020. *Information and the persistence of the gender wage gap: Early evidence from California's salary history ban.* Technical Report 27094. National Bureau of Economic Research, Cambridge, Massachusetts.

[21] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *International Conference on Machine Learning*. 1929–1938.

[22] Oliver P. Hauser and Michael I. Norton. 2017. (Mis) perceptions of inequality. *Current Opinion in Psychology* 18 (2017), 21–25.

[23] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. In *International Conference on Machine Learning*. 2692–2701.

[24] Peter S. Hovmand. 2014. *Community Based System Dynamics.* Springer.

[25] Lily Hu and Yiling Chen. 2018. A Short-Term Intervention for Long-Term Fairness in the Labor Market. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1389–1398. https://doi.org/10.1145/3178876.3186044

[26] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 1617–1626.

[27] Hyeok Jeong and Robert Townsend. 2008. Growth and inequality: Model evaluation based on an estimation-calibration strategy. *Macroeconomic Dynamics* 12, S2 (2008), 231.

[28] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 325–333. http://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.pdf

[29] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream Effects of Affirmative Action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. 240–248. https://doi.org/10.1145/3287560.3287578

[30] Ibram X. Kendi. 2019. *How to be an antiracist* (1 ed.). One World.

[31] Jietae Lee, Su Whan Sung, and In-Beum Lee. 2009. *Process Identification and PID Control.* Wiley-IEEE Press.

[32] Asaf Levanon, Paula England, and Paul Allison. 2009. Occupational feminization and pay: Assessing causal dynamics using 1950–2000 US census data. *Social Forces* 88, 2 (2009), 865–891.

[33] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning*.

[34] Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. 2020. The Disparate Equilibria of Algorithmic Decision Making When Individuals Invest Rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 381–391. https://doi.org/10.1145/3351095.3372861

[35] Max O. Lorenz. 1905. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* 9, 70 (1905), 209–219.

[36] Qingyin Ma, John Stachurski, and Alexis Akira Toda. 2020. The income fluctuation problem and the evolution of wealth. *Journal of Economic Theory* 187 (2020), 105003.

[37] Michael Madaio, Su Lin Blodgett, Elijah Mayfield, and Ezekiel Dixon-Román. 2021. Beyond "Fairness": Structural (In)justice Lenses on AI for Education. *arXiv preprint arXiv:2105.08847* (2021).

[38] Rajib B. Mallick, Michael J. Radzicki, Martins Zaumanis, and Robert Frank. 2014. Use of system dynamics for proper conservation and recycling of aggregates for sustainable road construction. *Resources, Conservation and Recycling* 86 (2014), 61–73.

[39] Jane Mansbridge and Shauna L. Shames. 2008. Toward a theory of backlash: Dynamic resistance and the central role of power. *Politics & Gender* 4, 4 (2008), 623–634.

[40] Leslie McCall, Derek Burk, Marie Laperrière, and Jennifer A. Richeson. 2017. Exposure to rising inequality shapes Americans' opportunity beliefs and policy support. *Proceedings of the National Academy of Sciences* 114, 36 (2017), 9593–9598.

[41] Heather McGhee. 2021. *The sum of us: What racism costs everyone and how we can prosper together.* One World.

[42] Dilip Mookherjee and Anthony Shorrocks. 1982. A decomposition analysis of the trend in UK income inequality. *The Economic Journal* 92, 368 (1982), 886–902.

[43] James Morgan. 1962. The anatomy of income distribution. *The Review of Economics and Statistics* (1962), 270–283.

[44] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. 2019. From Fair Decision Making To Social Equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. 359–368. https://doi.org/10.1145/3287560.3287599

[45] National Committee on Pay Equity. [n.d.]. The Wage Gap Over Time: In Real Dollars, Women See a Continuing Gap. https://www.pay-equity.org/info-time.html accessed 2 June 2021.

[46] Pegah Nokhiz, Aravinda Kanchana Ruwanpathirana, Neal Patwari, and Suresh Venkatasubramanian. 2021. Precarity: Modeling the Long Term Effects of Compounded Decisions on Individual Instability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. https://arxiv.org/abs/2104.12037

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

[48] Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[49] George Orwell. 1949. *Nineteen eighty-four.* Secker & Warburg.

[50] Thomas Piketty and Emmanuel Saez. 2003. Income inequality in the United States, 1913–1998. *The Quarterly Journal of Economics* 118, 1 (2003), 1–41.

[51] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[52] Benjamin Recht. 2018. A Tour of Reinforcement Learning: The View from Continuous Control. arXiv e-prints, art. *arXiv preprint arXiv:1806.09460* (2018).

[53] C. J. Roberts. 2013. Shelby County, Alabama v. Holder, attorney general, et al. No. 12-96, Argued February 27, 2013—Decided June 25, 2013.

[54] Emmanuel Saez. 2019. Striking it Richer: The Evolution of Top Incomes in the United States (Updated with 2017 final estimates). https://eml.berkeley.edu/~saez/TabFig2018prel.xls UC Berkeley.

[55] Melissa L Sands. 2017. Exposure to inequality affects support for redistribution. *Proceedings of the National Academy of Sciences* 114, 4 (2017), 663–668.

[56] Maria Veronica Santelices and Mark Wilson. 2010. Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review* 80, 1 (2010), 106–134.

[57] Simo Särkkä. 2013. *Bayesian filtering and smoothing* (3 ed.). Cambridge University Press.

[58] Kendra Scott, Debbie S Ma, Melody S Sadler, and Joshua Correll. 2017. A social scientific approach toward understanding racial disparities in police shooting: Data from the Department of Justice (1980–2000). *Journal of Social Issues* 73, 4 (2017), 701–722.

[59] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 59–68.

[60] Christina Starmans, Mark Sheskin, and Paul Bloom. 2017. Why people prefer unequal societies. *Nature Human Behaviour* 1, 4 (2017), 1–7.

[61] Dominique Thomas. 2019. Black lives matter as resistance to systemic anti-Black violence. *Journal of Critical Thought and Praxis* 8, 1 (2019).

[62] Evan Thomas. 2019. Why Sandra Day O'Connor Saved Affirmative Action. https://www.theatlantic.com/ideas/archive/2019/03/how-sandra-day-oconnor-saved-affirmative-action/584215/

[63] U.S. Census Bureau. 2019. Historical Reported Voting Rates. https://www.census.gov/data/tables/time-series/demo/voting-and-registration/voting-historical-time-series.html

[64] U.S. Supreme Court. 2003. Grutter v. Bollinger. 539, No. 02-241 (2003), 306.
[65] Patrice Villettaz, Gwladys Gillieron, and Martin Killias. 2015. The effects on re-offending of custodial vs. non-custodial sanctions: An updated systematic review of the state of knowledge. *Campbell Systematic Reviews* 11, 1 (2015), 1–92.
[66] Min Wen, Osbert Bastani, and Ufuk Topcu. 2019. Fairness with Dynamics. *arXiv preprint arXiv:1901.08568* (2019).
[67] Maisha T. Winn. 2019. *Girl time: Literacy, justice, and the school-to-prison pipeline.* Teachers College Press.
[68] Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. 2019. Group Retention when Using Machine Learning in Sequential Decision Making: the Interplay between User Dynamics and Fairness. In *Advances in Neural Information Processing Systems*. 15243–15252.
[69] Xueru Zhang and Mingyan Liu. 2020. Fairness in Learning-Based Sequential Decision Algorithms: A Survey. *arXiv preprint arXiv:2001.04861* (2020).