Linear feedback system models for understanding systemic inequity

Anonymous

ABSTRACT

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, given a broader macro viewpoint of a social system, these study parameters become limited. In a societal system, neither a singular decision maker nor defined agent behavior rules exist. There is also the broader concern of achieving and maintaining equity in the face of strategic and other forms of resistance in a society. Using a key idea in the theory of feedback systems theory, namely *proportional-integral-derivative (PID) feedback and control*, we propose a framework to study the question of how to achieve and sustain equity in a social system. We illustrate how different components of the PID capture analogous aspects of societal interventions such as inequality representations, the long-term impacts of inequality, and the pace in which a society moves towards equity. PID does not take notice of the model that describes the underlying system, nor does it seek to capture individual-based actions. It is a system-based approach to study the global system 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 understand ways to produce and sustain equity in a social system by incorporating drivers of change, historical accumulation of inequity, as well as resistance to change.

KEYWORDS

equity, policy, systematic bias, linear systems models

ACM Reference Format:

Anonymous. 2018. Linear feedback system models for understanding systemic inequity. In *EAAMO '21: ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization, October 5–9, 2021, Virtual.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In the literature on algorithmic fairness, the bulk of research work focuses on what we will call "one-shot" fairness interventions, where the goal is to make a particular decision process fair (or less unfair) under some formal definition of fairness [15, 16, 21, 45, 68]. In recent years, this framework has been extended to "long-range" fairness, where a decision system operates over multiple epochs, and the goal is to design an intervention that (eventually) ensures that the decision system is fair [7, 26, 30, 39, 61].

A broader social system however differs from this closed world in crucial ways [60]. Firstly, there is no *single* decision system to be made fair: rather, the goal is a *society-wide* measure of *equity* between different demographic groups. Secondly, the system is not closed and does not have well-defined rules of engagement: actors that have power can change the rules or policies in reaction to enacted interventions. And finally, the goal is not just to achieve equity, but to sustain it in the face of resistance. As an example of such resistance, consider the goal of equity in voting rights and continued strategic and varying efforts to suppress voting rights in the United States.

> Near the end of her dissent in Shelby County v. Holder, Justice Ginsburg suggested a simple analogy to illustrate why the regional protections of the Voting Rights Act (VRA) were still necessary. She wrote that "[t]hrowing out preclearance when it has worked and is continuing to work to stop discriminatory changes is like throwing away your umbrella in a rainstorm because you are not getting wet." [33]

In such a setting, how might one design system-wide interventions that can move society towards equity and sustain it? And how might one recognize when "it has stopped raining"? Our contribution in this work is a framework for modeling the production and long-term sustainability of equity in a social system. Our key idea is to borrow elements from the theory of control – specifically the idea of PID control that is a fundamental concept in control theory.

We will show that the three different elements of PID control – the *proportional*, *integral*, and *derivative* terms – have natural analogues in how we intervene in social systems. Coarsely, the proportional term allows a representation of today's inequities, the integral term captures the lived impact of inequity over time, and the derivative term captures the pace of current progress, including improvements towards equity and resistance to these changes. The perceived weakness of PID control – its disregard of the model used to describe the underlying system – is in fact an advantage when dealing with social systems that we have little hope of modeling but can observe.

To illustrate the PID control based modeling we introduce in this paper, consider another example. The pay gap between men and women is the result of a complex system, impacted by cultural gender roles, K12 and higher education systems, inequity in recruiting, bias of managers and co-workers, the glass ceiling, and more. The current state of inequity (the proportional term) can be represented in our model by the current pay gap. The integral term represents the result of this pay gap over time; the difference in lifetime earnings based on gender. The derivative term incorporates broader societal resistance to increased equity. One example of this resistance term is the phenomenon where previously high-paying

EAAMO '21, October 5–9, 2021, Virtual

ACM ISBN XXXX-X/21/10…$0.00

https://doi.org/10.1145/1122445.1122456

fields that increase the percent of women experience a resulting decrease in pay [38].

Within this system, a single organization's particular change of policy, e.g., use of an automated hiring algorithm, can have a direct impact on hiring (and thus pay equity) for their organization, impacting the proportional term. The PID control model incorporates that information while recognizing that such changes would also have an indirect impact reflected in the other model terms, e.g., via feedback into the community of job seekers for such positions, in ways shown to be both positive and negative for equity [9].

As demonstrated by the pay gap example, the modeling approach we introduce here does not attempt to directly model the direct and indirect actions of the individual agents in the system (job seekers, employers, educators, etc.). In the face of complex sociotechnical systems that generate *systemic* inequities, our approach is a system-based model where changes in the measured inequity affect the global system via feedback mechanisms. The amount of feedback, i.e., the degree to which observable inequity pushes people to change the system, and the degree to which improving equity causes resistance, can be estimated from historical behavior. Such models can be used to extrapolate systems to the future, or quantify the long-term impact of a short-term change.

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" [65]. That was 18 years ago; racial inequity in college admissions persists, and O'Connor has since said "That may have been a misjudgement" [64]. 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.

2 MODEL BACKGROUND

As discussed above, our model is intended to have minimal complexity yet quantitatively describe some feedback mechanisms within a complicated system. These mechanisms have historically operated to maintain inequity, as well as to drive toward a more equitable future. We want the model to be data-driven, but have a small number of parameters so that it is able to be built with recent historical data that represents the societal systems and structure that exist today.

We've chosen to have a feedback model with three components (and thus parameters): *proportional*, *derivative*, and *integral* terms:

- Proportional: a term describing change in future inequity based on current inequity
- Integral: describes impact on future inequity of cumulative inequity over time
- Derivative: describes change in future inequity based on current directional change in inequity

We've chosen to model every other change to the system as the *disturbance*. These are the other things influencing equity besides those which are modeled in the controller. For example, a pandemic has impact on equity. This "process noise" is different from measurement noise, as the disturbance actually changes the underlying state (which thus has direct and lasting impact), while measurement noise only changes the instantaneous measurement of the state. Measurement noise would be higher, e.g., if we sample less of the population to quantify the current inequity.

Why these three parameters? Together, they provide a second order difference equation for the dynamics of the system. The behaviour and stability of 2nd order dynamical systems is well-studied. In particular, we name our three components for the components of a proportional-integral-derivative (PID) controller. Although we are proposing system modeling, not control, a large volume of literature addresses the stability (i.e., if the system output stays reliably near the set point) of PID controllers [36].

As its purpose is to help extrapolate the future, we might compare it to simpler methods of extrapolation. For example, one might simply draw a straight line through the recent historical inequity using linear interpolation. A line uses two parameters, and is easy to compute, but it excludes feedback, i.e., it assumes the slope will stay the same despite societal pressure and push-back. It would be possible to model system dynamics with more parameters, but this would add complexity to the parameter estimation task. Further, one could model system dynamics with other methods (e.g., frequency domain models as popular in controls applications) but they may be difficult to justify or relate to real-world mechanisms in systemic oppression.

We next describe societal motivations for choosing these model terms.

2.1 Proportional

The proportional term describes societal reaction to current inequity, representing the idea that the more inequity there is, the more likely it is that people are aware of that inequity and working to rectify it. For example, publicly announcing large inequities in facial recognition performance by skin color and binary gender led to reduced disparities from targeted products [53].

People often hold a "belief in a just world" [37] despite evidence to the contrary. However, seeing unambiguous evidence of injustice while believing in a just world leads to *cognitive dissonance*, one's mental state when one's beliefs and values don't align with one's behavior. The outcome of this dissonance can be to change one's behavior, to justify the behavior in some new way, or to deny the truth of the new information. Cognitive dissonance theory can help to explain both how inequality is justified in a society with a stated value of equality and justice, as well as why people change their behavior more when the evidence of inequity and injustice is stronger and thus less avoidable [20].

The power of the legal system to enforce equal justice is also, often, a function of how extreme the discrimination is. The rule-of-thumb used by the U.S. Equal Employment Opportunity Commission to define "substantially different" is that the difference in rate of selection between the over-represented and under-represented groups is more than 20% [8]. A larger discrepancy, beyond 20%, makes it easier to prove that the difference violates the law.

We note that individual support for policy change may actually decrease when observing inequity, due to belief in a just world. For example, observing an unhoused person makes a person less willing to endorse a progressive tax policy [57] compared to not observing an unhoused person. Our model itself is agnostic about whether the sign of the feedback is positive or negative. However, other work shows that exposure to historical data about rising wealth inequality in the US tends to increase support for redistributive policies [44]. We address these longer term trends and their impact on societal action via the next term in the model.

2.2 Integral

Does an inequity change as a function of the cumulative sum of past inequity? There are certainly aspects of the inequity which can lead to more inequity. A group that experiences lower average income, for example, then accumulates less wealth over time, since wealth can be approximated as a cumulative sum of excess income. Wealth provides power, which then can be used by the group with more power to maintain aspects of the system most favorable to themselves, thus increasing future inequity [67].

In contrast, cumulative inequity over time can lead to social movements that demand change. The #BlackLivesMatter movement was spurred not only by specific incidents of violence but also by the understanding that these incidents are part of a long-term *systemic* history of violence against Black people in the United States [59, 63]. The integral term aims to capture this awareness and societal response to cumulative long-term inequity.

2.3 Derivative

Psychological reactance is the pushback when people are confronted with threats to their freedom [4]; not being allowed to discriminate could be experienced as a "threat to freedom", as articulated by the common refrain, "when you're accustomed to privilege, equality feels like oppression". The society-wide impact of reactance could be modeled as feedback proportional to the derivative of inequity. We can see this reactance playing out in judicial decisions. Even though voter suppression has been a continual battle in the U.S. and voting rates are currently inequitable between Black and White citizens, conservative justices in Shelby County vs. Holder rejected a pro-equity voting policy because "The Fifteenth Amendment is not designed to punish for the past" [55]. The decision says that inequity is insufficient to justify a pro-equity law because today's inequity is lesser in magnitude than it was in the past [55]. In other words, the downward slope of voting inequity is used to justify allowing States greater freedom to implement policies that will increase inequity in the future, as States did immediately following the decision [41].

2.4 Why not a PD model?

It seems plausible a model that captures proportional and derivative terms (modeling the magnitude of change needed as well as resistance) might be sufficient to capture (linear) feedback and would require one less parameter. However, one of the lessons from PID control is that controllers with only proportional and derivative (PD) terms are unable to maintain a steady-state output at the set point (in our paper, the set point is equity) when there are non-zero forces pushing away from the set point [71]. One can see this by contradiction: assume a PD controller has a steady-state value of equity. The "steady-state" implies a zero derivative, and at the set point, the proportional term is zero. Thus there would be zero feedback, either P or D, to counter the forces pushing away from the set point. A PD controller could have a steady-state not at the set point, or pass through the set point while not at steady-state, but not both. A simulated example is plotted in Figure 1 – the additional integral term feedback allows the system to converge to the set point of 1.0.

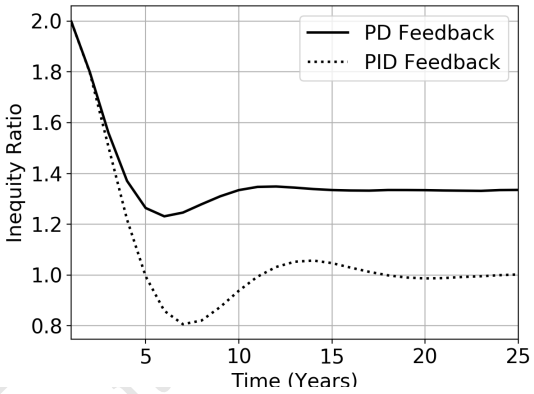

Figure 1: Simulation of inequity when feedback has PD components ($k_P = -0.3$ and $k_D = -0.8$) vs. PID components ($k_P = -0.3$, $k_I = -0.05$, and $k_D = -0.8$), in the face of a disturbance adding 0.1 to the slope at each time. Feedback with an integral term allows the system to converge to equity.

In terms of modeling inequity, the integral term in a PID feedback model has a very important purpose, namely to provide the feedback needed to keep the long-term output at the equity, i.e., to "sustain equity" [34] in the realistic case that there are forces that push for (or benefit from) inequity. In societal terms, *policies and mechanisms that push back in proportion to the cumulative historical inequity are necessary to sustain long-term equity.*

3 DYNAMICAL PID STATE MODEL

In this section, we provide a model for a single measure of inequity, $x(n)$, for integer times n. This measure would be equal to some value, or *set point*, if society was equitable. In this paper we consider examples which monitor a ratio which would indicate equity if its value was 1.0. In these cases the 'inequity' $x(n)$ is the ratio minus the set point of 1.0. Further, in our examples we choose the group in the numerator to put make the ratio historically above 1, so that readers can consistently interpret $x(n)$ as 'inequity', and work to reduce $x(n)$ is pro-equity. For example, in 1964, from US Census statistics, 70.7% of White Americans voted, while 58.5% of Black Americans voted [1], a ratio of 1.209. With a set point / equity value of 1.0, thus $x(n) = 0.209$.

3.1 State Model

We model a system with proportional, integral, and derivative (PID) feedback components, as shown in Figure 2. The system may have

multiple outputs, but here we measure the equity of the outputs of the system, for example, the wage gap between men and women. We assume that there is feedback, that is, the inequitable outputs of the system affect people and their resources, which then change the system itself over time. For example, seeing inequitable outputs might change support for policies that affect equity; or lower than average income leads to lower wealth and thus less power to change the system. We describe any other changes to the inequity that is not feedback from the system's outputs as part of the disturbance $w(n)$. We model the dynamics as linear, that is, a weighted sum of the proportional, derivative, and integral 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.

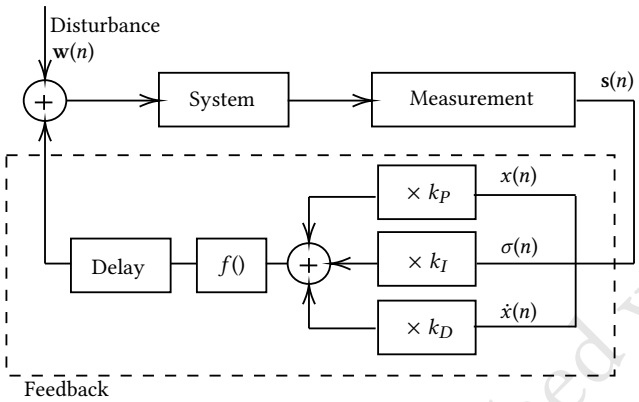

Figure 2: A feedback model for societal control on a system with inequity $x(n)$, in which the control is linear in $x(n)$ itself (the proportional term), in the derivative $\dot{x}(n)$, and in the integral of $x(n)$, $\sigma(n)$.

We define a state $\mathbf{s}(n) = [x(n), \sigma(n), \dot{x}(n)]^T$ where:

- $x(n)$ is the inequity at time n,
- $\sigma(n)$ is the cumulative inequity since time 0, thus $\sigma(n) = \sigma(n-1) + x(n)$.
- $\dot{x}(n)$ is the difference $x(n) - x(n-1)$,

Then we model the societal feedback as a linear function of these terms:

$$\mathbf{k}^T \mathbf{s}(n) = k_P x(n) + k_I \sigma(n) + k_D \dot{x}(n), \tag{1}$$

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:

$$\dot{x}(n+1) = \dot{x}(n) + \mathbf{k}^T \mathbf{s}(n) + w_2(n), \tag{2}$$

where $w_2(n)$ is the disturbance (also called process noise). 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:

$$\begin{aligned} x(n+1) &= x(n) + \dot{x}(n) + w_0(n) \\ \sigma(n+1) &= \sigma(n) + x(n+1) = \sigma(n) + x(n) + \dot{x}(n) + w_1(n). \end{aligned} \tag{3}$$

These equations (2) and (3) are our state update equations that implement the proportional, integral, and derivative feedback terms as described and justified prior. In short we can write

$$\mathbf{s}(n+1) = A\mathbf{s}(n) + \mathbf{w}(n) \tag{4}$$

where $\mathbf{w}(n) = [w_0(n), w_1(n), w_2(n)]^T$ is the disturbance, and

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ k_P & k_I & k_D + 1 \end{bmatrix}. \tag{5}$$

Finally, 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 $\mathbf{s}(n)$

$$y(n) = C\mathbf{s}(n) + v(n), \tag{6}$$

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

How feedback is incorporated. In equation (2) 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 $x(n)$. In a physical control system, the choice of which component to update is determined by the system itself: for example, a cruise control system that maintains car speed at a set point controls the application of the accelerator, which controls the gas and thus the acceleration of the car (the derivative of the speed). We argue here that for social systems, updating the derivative is an appropriate modeling choice. Since our goal is to move towards equity, the state variable $x(n)$ is the measure of equity. It is difficult to enforce policies that affect the measure of equity directly, rather than affecting the rate at which one might approach equity. Policies around affirmative action, prohibition of discrimination, and the like all fall into this category. One example of a policy that affects a set point directly is a raise in the minimum wage as it pertains to income gaps between different segments of the population. But even in this case, the policy affects only the minimum wage, and will only gradually have an effect (if at all) on wage gaps beyond the lowest level.

3.2 Numerical Simulation

To make concrete how the model incorporates feedback, we give the following example. For this example, we run our model in simulation for known parameters. For simplicity, we make the example noise-free by setting $w(n) = 0$ in (4). We use parameters $k_P = -0.05$, $k_I = 0.02$, and $k_D = -0.5$ and run the update equation for 10 time steps. We initialize the state with proportional term $x(0) = 1$, integral term $\sigma(n) = 0$, and slope term $\dot{x}(0) = -0.2$. At each step, we follow (4) with noise $w(n) = 0$. Specifically, at time $n = 1$, we calculate $\mathbf{s}(1) = A\mathbf{s}(0)$,

$$\mathbf{s}(1) = A\mathbf{s}(0) = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ -0.05 & 0.02 & -0.5 + 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -0.2 \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.8 \\ -0.15 \end{bmatrix}. \tag{7}$$

Thus the inequity at time 1 has reduced to 0.8, the cumulative inequity is now 0.8, and the slope is −0.15. The new inequity is 0.2 below the prior inequity value, as given by the previous slope of -0.2. Further, the current $\dot{x}(1)$ has updated to add in a linear combination of the k vector multiplied by the state values $x(n)$. The update in (4) repeats to calculate $\mathbf{s}(n)$ for $n = 2, \ldots, 9$.

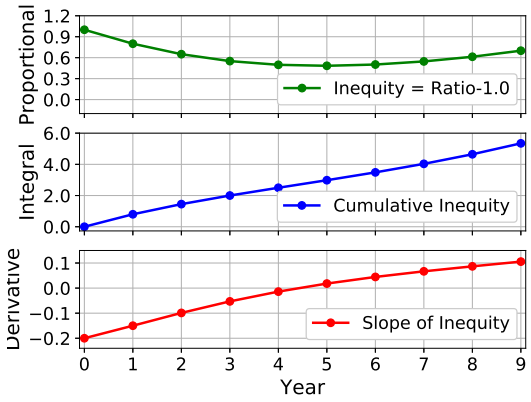

Figure 3: A noise-free simulation of the model with parameters $k_P = -0.05$, $k_I = 0.02$, and $k_D = -0.5$. The simulation starts with proportional term $x(0) = 1$, integral term $\sigma(n) = 0$, and slope term $\dot{x}(0) = -0.2$. At each step, we follow (4) with noise $w(n) = 0$.

Over 10 time steps, the system changes from decreasing inequity to rising inequity, as shown in in the results in Figure 3. This change comes from the rising slope. While the initial slope was negative (−0.2), it rises for two reasons, *despite* the $k_P = -0.05$ which decreases the slope by 5% of the inequity each time step. Initially the slope rises primarily because of the negative k_D, which cuts the slope in half at each time step. Later, it rises because of the positive $k_I = 0.02$ because the cumulative inequity is getting higher and higher, and $k_I = 0.02$ means that 2% of the cumulative inequity is added into the slope. We can see that in this model:

(1) The proportional term acts like interest. For $k_P < 0$, it directly pushes the slope down towards equity, $k_P \times 100\%$ of the current inequity.
(2) The integral parameter k_I is unimportant at the start when the cumulative inequity is small, but for a persistent inequitable outcome, it has outsized influence at later times.
(3) Any $-1 \leq k_D < 0$ tends to "push on the brakes" by bringing the slope closer to 0. We note that $-2 \leq k_D < -1$ leads to a ringing effect where the slope sign may change from time n to $n + 1$.

The simulation procedure we describe, with $w(n) = 0$, also provides the expected value of the state as long as noise is zero-mean. In this paper we use this simulation procedure as an extrapolation procedure. That is, given parameters $\mathbf{k}$ and the current state, we use (4) to generate the expected value of the state into the future.

3.3 Model Parameter Estimation

Given the dynamic model in (4) and a set of longitudinal data for $\{y(n)\}_n$, we would 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 estimate the parameters $\mathbf{k}$ from a longitudinal data set?

We provide one method here. Some systems identification methods estimate the entire update matrix A from (5), but if we do that we will be estimating all 9 elements of A, and this would complicate our model. Here, we derive a least squares estimator for our three parameters k_P, k_I, and k_D.

We do this 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 (4):

$$\Delta\mathbf{s}(n) = (A - I)\mathbf{s}(n) + \mathbf{w}(n), \tag{8}$$

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 parameters,

$$\dot{x}(n+1) - \dot{x}(n) = \mathbf{k}^T\mathbf{s}(n) + w_2(n). \tag{9}$$

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

$$\begin{aligned} \ddot{\mathbf{x}} &= S\mathbf{k} + \mathbf{w}_2, \text{ where,} \\ \ddot{\mathbf{x}} &= [\ddot{x}(1), \ldots, \ddot{x}(N)]^T \\ \mathbf{w}_2 &= [w_2(1), \ldots, w_2(N)]^T \\ S &= [\mathbf{s}(1), \ldots, \mathbf{s}(N)]^T \end{aligned} \tag{10}$$

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,

$$\hat{\mathbf{k}} = (S^TS)^+S^T\ddot{\mathbf{x}}. \tag{11}$$

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 [18]. In our case, this second part is calculated with (11). The first part is described next.

3.4 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 [58]. We denote this state estimate as $\hat{\mathbf{s}}(n)$ for all $n \in \{0, \ldots, N\}$.

As described above, from the state estimates we calculate the change in slope $\ddot{\mathbf{x}}$ which we use with the state S in (11) to re-estimate

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 the annual data since 1960 from [49]. Compiled from US Census data, the data refers to the ratio of median income between men and women full-time, year-round workers. Currently, women workers' median pay is $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, which is currently 1.22. The set point for equity would be 1.0.
(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 [1]. This data is for national congressional / presidential elections, i.e., every other 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 [52, 56].

We consider the following experimental questions:

(1) How well does the model predict future inequity?
(2) When does the model perform poorly?
(3) How much history do we need to accurately predict the future?
(4) How can one interpret the model parameters?
(5) How would the system evolve if the feedback was quantitatively different?
(6) How does the model compare to existing simple models?

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/PID_equity_modeling/.

4.1 How well does the model predict future inequity?

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 4. We test (in the left column) using the 1st 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. The performance with each model can be described here.

- *Earnings, Men vs. Women*: The top row of Figure 4 shows predictions for the gender pay gap. While the model does predict the sign of the slope, it accelerates faster towards equity than we actually see in the most recent 20 years.
- *Voting, White vs. Black*: The voting gap 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.
- *Income, Top 10% of Earners vs. 10% of All Income*: Notably, the model, trained on the 1st half of data, i.e., a period of constant and low inequity from 1945-1981, predicts that inequity will increase significantly starting in 1982, as it in fact does. We note that the rise is under-predicted using the 1st half of data as training, and the prediction using the 1st 2/3 of data as training shows inequity accelerating more than it ends up accelerating after 2005.

4.2 When does the model perform poorly?

Why did this model perform poorly on the pay gap data? It may be because the model parameters changed dramatically between the first half and the last half of the data set. In Table 1 we show the parameters estimated for the model from the 1st half, the first 2/3, all of the data, and from the 2nd half of the data. 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 positive to negative, while the other parameters keep the same sign. Again, positive values mean that the slope increases (towards higher inequity) while there is current inequity. We interpret this as saying, in the period 1960-1989, the dynamics of the pay gap indicate that current inequity produces more inequity, although the effect is attenuated by the negative k_I which causes inequity to decrease eventually due to the large cumulative inequity over time. In contrast, in the period 1990-2018, the current inequity *reduces* inequity; but a negative but smaller magnitude k_I value means that we are less impacted by the cumulative inequity and a larger negative k_D means that society is pushing back more against change. These changes in the structure of the feedback from the first 30 years of the data mean that any prediction of the future that extrapolate from those first years are unlikely to be accurate.

4.3 How much history do we need to accurately predict the future?

In the previous section, we compared using 1/2 of the data, vs. 2/3 of the data for training. While the error is generally lower when using 2/3 for training, there are two potential reasons: 1) more years of training, and 2) fewer years of extrapolation. In this section, we set the number of years of extrapolation as constant in an effort to isolate the effects of training years.

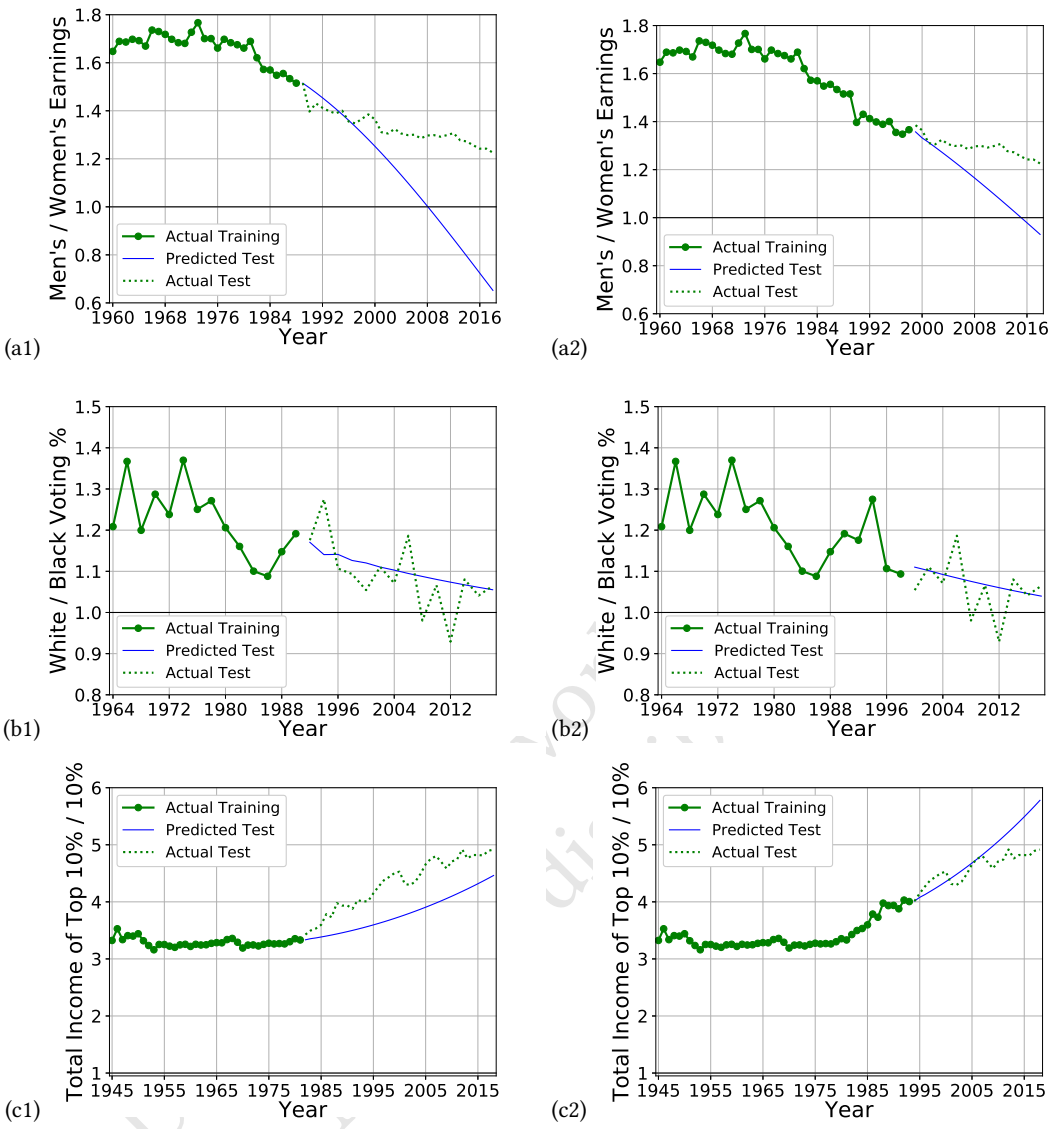

Figure 4: Training the model only from the historical (Col #1) first 1/2 of data; (Col #2) first 2/3 of the data, we estimate PID model parameters. Then we extrapolate (simulate the model) to predict the remaining years, and compare to the actual test period data, for U.S. (a) earnings ratio of men vs. women; (b) voting gap ratio white vs. Black; (c) total income of top 10% of earners (divided by 10%). In all plots, a ratio of 1.0 (––) is equity.

To address this, we vary the number of training set data length, i.e., the number of data values used for training, and estimate the model parameters. Next, we extrapolate and estimate the following 10 data values (as the test set). We compute the average absolute error between these 10 estimates and the actual values, and normalize by the average inequity in the combined training and test data, which we call the *average normalized extrapolation error*. Figure 6 shows the results.

First, we note that dataset 2, the voting gap data set, has values that change quickly. For example, the values vary by ±0.1 from one year to the next, similar in magnitude to the inequity (the ratio - 1.0), which is about 0.1-0.2. In short, the proposed model will not result in low average normalized error even if it estimates the general shape of the trajectory of the inequity, as it appears to do in Figure 4(b).

For the other datasets, the normalized extrapolation error is notably higher for training set length < 10. However, above 20, data set length does not appear to improve results. This may indicate that the system evolves over time on the order of decades, preventing further refinement; or it may indicate the limits of a three-parameter model.

Data Set #	Name	Training Data Used	Proportional k_P	Integral k_I	Derivative k_D
1	Pay Gap Men vs. Women	First 1/2	0.0136	-0.00154	-1.07
		First 2/3	0.0064	-0.00107	-1.12
		All	-0.0050	-0.00030	-0.91
		Second 1/2	-0.0198	-0.00033	-1.22
2	Voting Gap White vs. Black	First 1/2	-0.0819	-0.00097	-1.58
		First 2/3	-0.0790	-0.00145	-1.52
		All	-0.0900	-0.00133	-1.68
		Second 1/2	-0.5387	-0.02707	-1.09
3	Income of Top 10%	First 1/2	-0.0140	0.00063	-1.31
		First 2/3	-0.0200	0.00109	-1.27
		All	-0.0079	0.00039	-0.80
		Second 1/2	0.0153	-0.00041	-0.72

Table 1: 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 4.4.

4.4 How can one interpret the model parameters?

We report the estimated model parameters in Table 1. We advocate for this model, in part, because the model parameters are interpretable as feedback mechanisms. As detailed in Section 3.1, next year's slope increases by k_P times the current inequity, plus k_I times the cumulative sum of inequity, plus k_D times the current slope.

In all of our data sets, there is current inequity, i.e., the ratio is above the set point of 1. 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 worse inequity).

The sign on the integral term is similar in effect. Since we are looking at data with historical inequity, the cumulative sum of the inequity (the integral of the plot above 1.0) is positive. Any $k_I < 0$ is pro-equity because it pushes the slope down (towards equity), while $k_I > 0$ pushes the slope up and away from equity.

However, the sign on the derivative term is different in effect. If we are currently on a pro-equity slope, that is, the slope is negative, then a $k_D < 0$ results in something positive being added to the slope, that is, a push towards inequity. Similarly, if we are currently on a pro-equity slope, $k_D > 0$ is pro-equity. However, if we are currently moving away from equity and thus the slope is positive, then the effect of the sign of k_D is opposite. That is, the derivative term always "pushes on the brakes" when $k_D < 0$.

As the magnitude of a parameter increases, it signifies that there is *more feedback* via this mechanism. For example, we compare the parameters for the income of the top 10% data when trained on the "First 1/2" vs. "All" of the data. Note that the first half was a period of low inequity, with increasing inequity at the end; while the second half includes rising inequity throughout. There is a lower k_P value when training on all of the data because the system presumably responds less to inequity when it contains a long period of rising inequity.

4.5 How would the system evolve if the feedback was quantitatively different?

We propose that the use of three feedback terms, each tied to particular societal mechanisms, can help evaluate how the evolution of the inequity might change if the these mechanisms were to change.

For example, we might wonder about how college admissions might be made more equitable by ending legacy admissions policies [11]. As legacy policies preferentially admit descendants of graduates of the same college, they would presumably impact feedback via the integral term, admitting students proportional to the cumulative sum of past inequity. How would a college's admissions equity improve if the k_I term was reduced?

To quantitatively address questions like this, we first train a model with the first half of a data set. We test using the Voting Black vs. White data set. We assume that some mechanism changes the feedback structure and thus changes a model parameter. We are particularly interested in imagining alternate futures which are more equitable, so we change parameters in a direction towards equity. In the voting gap data set, we can increase equity by making k_P and k_I *more* negative, and making k_D less negative. All three parameters are estimated to be negative on the first 1/2 of this data set (see Table 1), so we change the parameters in three ways: 1) by multiplying k_P by 2; 2) by multiplying k_I by 2; and 3) by dividing k_D by 2. We run these three alternate future simulations and show the results in Figure 7 along with the extrapolation given the estimated **k**.

We see in Figure 7 that reducing k_D has the greatest impact on equity — it would have produced equity, on average, by about 2016. While increasing the magnitude of the negative proportional and integral feedback values as tested would have improved equity, neither one alone would have resulted in equity in this period.

However, note that this simulation does not address the effort or cost to be able to reduce or increase feedback; it may be more feasible to work towards equity by reducing k_P rather than increasing k_D, for example. The described model doesn't address the difficulty

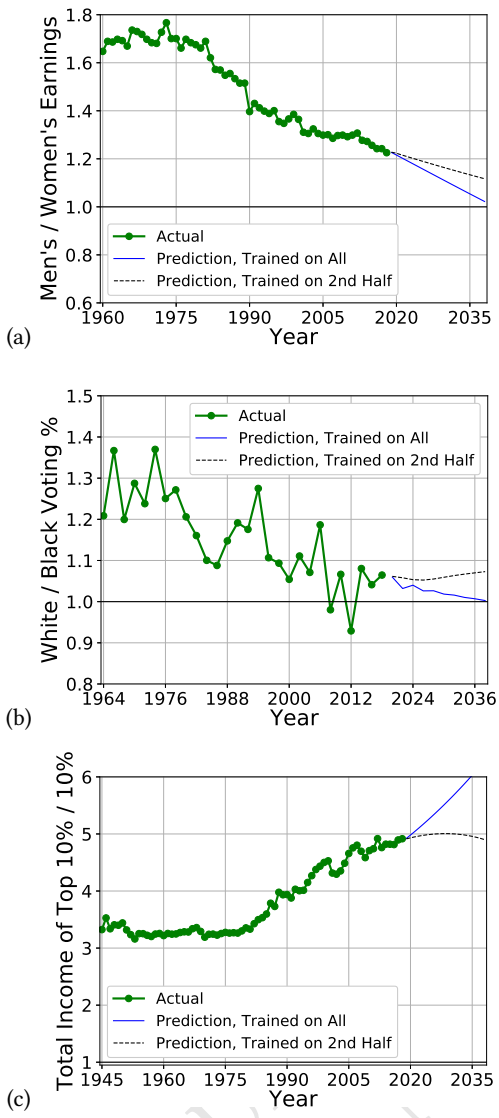

Figure 5: Training model from all historical data, then extrapolating / simulating to predict future 10 values, for U.S. (a) earnings ratio of men vs. women; (b) voting gap ratio white vs. Black; (c) total income of top 10% of earners (divided by 10%).

of changing systemic mechanisms that produce inequity, but it can model the impact of different types of changes.

4.6 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 consider

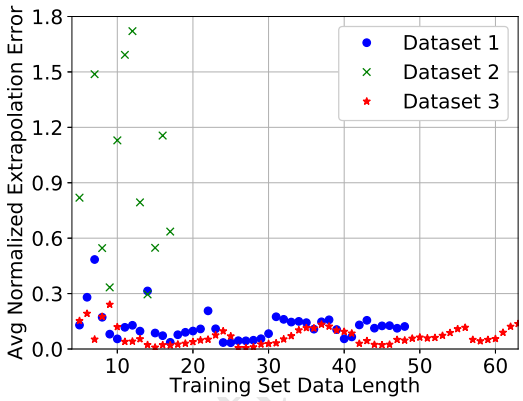

Figure 6: Average absolute error of a 10-length model extrapolation, normalized to average inequity, using a variable length of data to train the model. For datasets 1 & 3, error is notably higher for data length < 10.

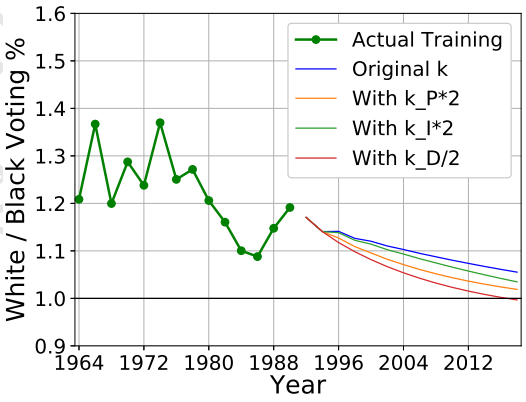

Figure 7: Simulations of alternate futures if the estimated k is altered, 1) by multiplying k_P by 2; 2) by multiplying k_I by 2; and 3) by dividing k_D by 2, compared to the future simulated by the estimated k.

linear regression, polynomial interpolation, and decision tree regressors.[1] The results can be seen in Figure 8. We also determine the root mean squared error of each of the comparison methods as well as the PID model for each of the datasets and training set combinations. These results are shown in Figure 9.

Overall, we find that there is no single model with lowest error across all datasets and training options. The PID model has one of the lowest error values across models for the voting gap and income datasets while linear regression performs well on the voting gap and pay gap datasets. Polynomial interpolation and PID share a similar model shape, but PID performs similarly or better than polynomial interpolation across all datasets. Given this, as well as

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

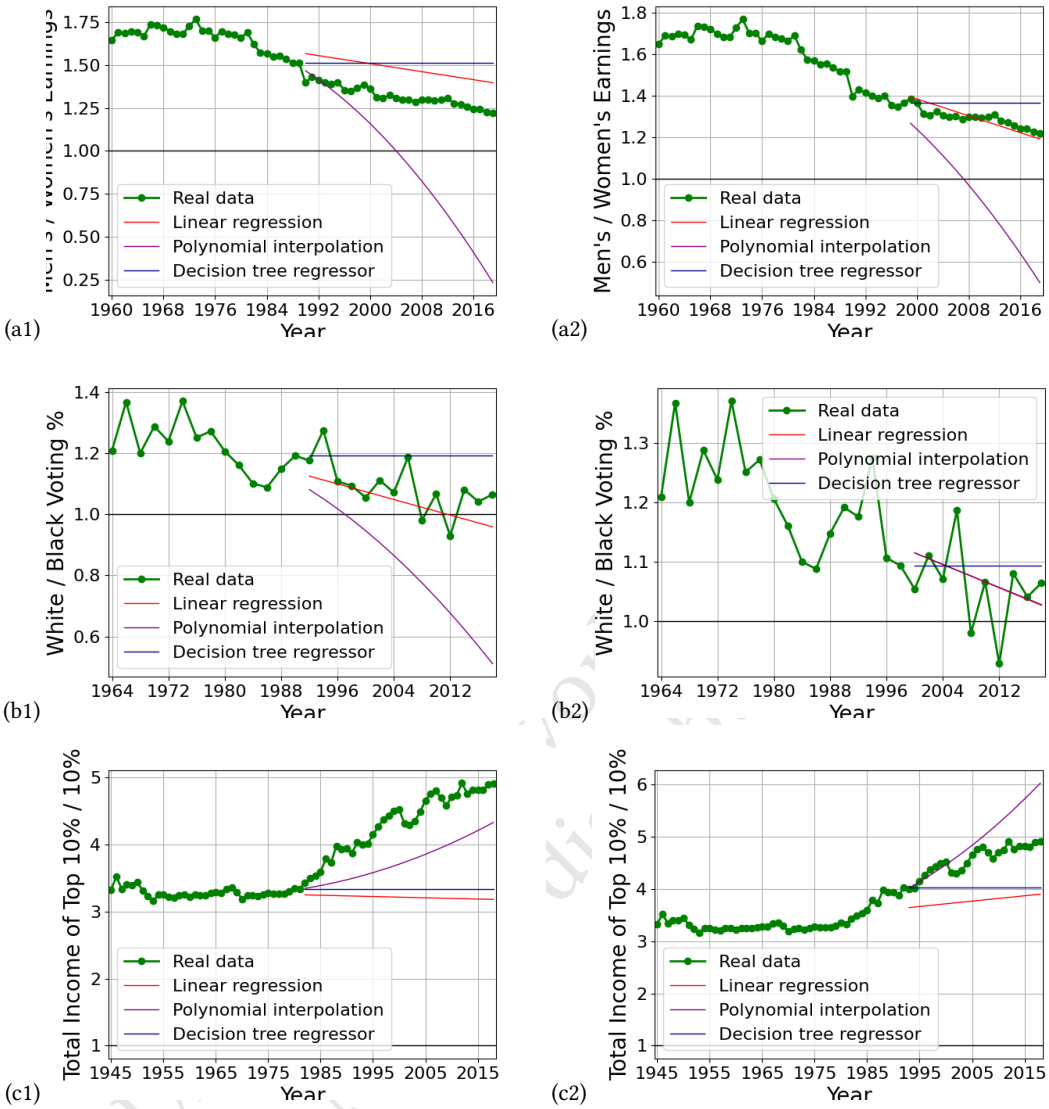

Figure 8: Training the models from the historical first half of the data (column 1) or first 2/3 of the data (column 2), we compare linear regression (red), polynomial interpolation (purple), and a decision tree regressor (blue) to the real-world data (green).

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.

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 [70] and the review article by Chouldechova and Roth [7].

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) [24, 32, 39] or finite or infinite-horizon decision making [13, 22, 27, 40, 48, 69]. These approaches are primarily model-based. Other model-based approaches use Markov Decision processes to capture agent behavior and use simulation techniques to analyze a system [12, 50]. MDPs can also be formally analyzed for long-term effects on (group and individual fairness) as explored by D'Amour et al. [12], Jabbari et al. [28], Joseph et al. [31], 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

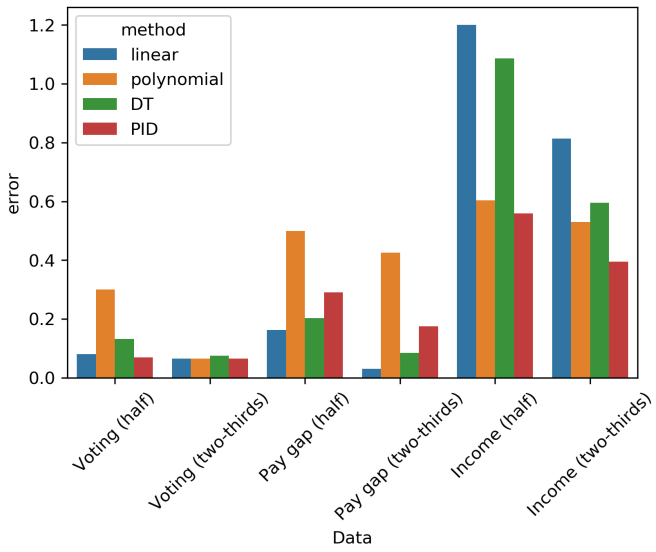

Figure 9: RMS errors with respect to test data for each of the comparison methods as well as the PID model for each of the datasets and training combinations.

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 overpolicing 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 [36] 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 [54] 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 [2] 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 [2, 6, 29]. 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 [19]. Note that the Lorenz Curves [42], the Gini coefficient [47], and Theil index [46] are some of the most well-known inequality measures, but are not models that can be used to predict future inequalities.

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 [43]. 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 [25]. 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" [60]? 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.

Regarding linearity, some feedback mechanisms do contain nonlinearities. For example, people's support of policy change in the face of inequity may not be linear. The "token" effect describes how people in a dominant group can be convinced there is no systemic bias when there exists a single example or *token* person of an oppressed group in a powerful position [3]. A token example thus reduces support for a redistributive policy much more than its marginal impact on equity.

Perception vs. Reality. We further use measurements of inequity as a proxy for the degree of inequity that drives the societal feedback mechanisms. If people support policies based on perceived inequity, there is a difference between the two. People perceive inequity differently based on their "social dominance orientation" [35], or their "belief in a just world" [20], both of which decrease one's perception of discrimination. However, when discrimination is unambiguous [20], and obtains more media coverage [10], people observe more discrimination. People's estimates of the level of inequality are inaccurate in the US and UK, and their estimates are heavily influenced by how much inequality they see locally [23]. While we don't dispute the difference between measured and perceived inequity, our model presumes that, society-wide, future changes in inequity are at some level influenced by present, historical, and past changes in measured inequity.

Power and Oppression. Our model does not make value judgements about the feedback mechanisms, rather, it reflects them. We

should be clear that feedback is connected to power; the ability to preserve inequity for the benefit of one's own group is made possible by one's own political, social, and/or financial power. Oppression is systemic, and rather than addressing particular actors (and thus motivations), we provide a modeling tool which quantifies the mechanisms by which power (and inequity) is preserved.

Multidimensionality of Oppression. We study only one inequity measure. In reality, systems of oppression intertwine in the "birdcage" analogy of [17]. For example, the income gap leads to the wealth gap, which then increases the health equity gap. 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. Should the set point be at equity? The model could be used with a different set point. People may want more equity, but are not likely to advocate for equal wealth among all people [51]. Within group wealth equity has a different meaning than equity between two groups. We use a set point of equity because we advocate for policies that "produce or sustain equity" between groups [34], but we recognize many people do not share a goal of equity [62]. An individual's measure of racial injustice is some combination of past-rooted, i.e., how much progress has made since some past date, and future-rooted, i.e., how far we have to go to get to an ideal future [5]. White Americans tend to be more past-rooted than Black Americans, and the more a past-rooted person believes that "we've come so far", the less they support affirmative action policies [5]. Such people's push to decrease inequity diminishes as inequity decreases, and thus can still be approximated with proportional feedback.

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 three 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 lower, on average, than other extrapolation 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.

REFERENCES

[1] 2019. Historical Reported Voting Rates. https://www.census.gov/data/tables/time-series/demo/voting-and-registration/voting-historical-time-series.html U.S. Census Bureau.

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

[3] Manuela Barreto, Naomi Ellemers, and Maria Soledad Palacios. 2004. The backlash of token mobility: The impact of past group experiences on individual ambition and effort. *Personality and Social Psychology Bulletin* 30, 11 (2004), 1433–1445.

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

[5] Amanda B Brodish, Paige C Brazy, and Patricia G Devine. 2008. More eyes on the prize: Variability in White Americans' perceptions of progress toward racial equality. *Personality and Social Psychology Bulletin* 34, 4 (2008), 513–527.

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

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

[8] US Equal Employment Opportunity Commission. 1979. Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures. OLC Control Number EEOC-NVTA-1979-1.

[9] Jessica Dai, Sina Fazelpour, and Zachary C Lipton. 2021. Fair Machine Learning Under Partial Compliance. In *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES-2021).* https://arxiv.org/abs/2011.03654.

[10] Matthias Diermeier, Henry Goecke, Judith Niehues, and Tobias Thomas. 2017. *Impact of inequality-related media coverage on the concerns of the citizens.* Number 258. DICE Discussion Paper.

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

[12] 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)*. Association for Computing Machinery, New York, NY, USA, 525–534. https://doi.org/10.1145/3351095.3372878

[13] 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).

[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] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.* 259–268.

[16] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 329–338.

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

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

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

[20] Carolyn L Hafer and Becky L Choma. 2009. Belief in a just world, perceived fairness, and justification of the status quo. In *Social and Psychological Bases of Ideology and System Justification,* John T. Jost, Aaron C. Kay, and Hulda Thorisdottir (Eds.). Oxford University Press New York, NY, 107–125.

[21] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems.* 3323–3331.

[22] 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.

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

[24] 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.

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

[26] 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.* 1389–1398.

[27] 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

[28] 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.

[29] Hyeok Jeong and Robert Townsend. 2008. Growth and inequality: Model evaluation based on an estimation-calibration strategy. *Macroeconomic Dynamics* 12, S2 (2008), 231.
[30] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in learning: classic and contextual bandits. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 325–333.
[31] 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
[32] 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)*. Association for Computing Machinery, New York, NY, USA, 240–248. https://doi.org/10.1145/3287560.3287578
[33] Ellen D Katz. 2014. Justice Ginsburg's Umbrella. *A Nation of Widening Opportunities* (2014).
[34] Ibram X. Kendi. 2019. *How to be an antiracist* (1 ed.). One World.
[35] Nour S Kteily, Jennifer Sheehy-Skeffington, and Arnold K Ho. 2017. Hierarchy in the eye of the beholder: (Anti-) egalitarianism shapes perceived levels of social inequality. *Journal of Personality and Social Psychology* 112, 1 (2017), 136.
[36] Jietae Lee, Su Whan Sung, and In-Beum Lee. 2009. *Process Identification and PID Control*. Wiley-IEEE Press.
[37] Melvin J Lerner and Carolyn H Simmons. 1966. Observer's reaction to the "innocent victim": Compassion or rejection? *Journal of Personality and Social Psychology* 4, 2 (1966), 203.
[38] 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.
[39] 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*.
[40] 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
[41] Tomas Lopez. 2014. 'Shelby County': One Year Later. *Brennan Center for Justice* (2014).
[42] Max O Lorenz. 1905. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* 9, 70 (1905), 209–219.
[43] 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.
[44] 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.
[45] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8 (2021).
[46] Dilip Mookherjee and Anthony Shorrocks. 1982. A decomposition analysis of the trend in UK income inequality. *The Economic Journal* 92, 368 (1982), 886–902.
[47] James Morgan. 1962. The anatomy of income distribution. *The Review of Economics and Statistics* (1962), 270–283.
[48] 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)*. Association for Computing Machinery, New York, NY, USA, 359–368. https://doi.org/10.1145/3287560.3287599
[49] 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.
[50] 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*. Association for Computing Machinery, New York, NY, USA. https://arxiv.org/abs/2104.12037
[51] Michael I Norton and Dan Ariely. 2011. Building a better America — One wealth quintile at a time. *Perspectives on Psychological Science* 6, 1 (2011), 9–12.
[52] Thomas Piketty and Emmanuel Saez. 2003. Income inequality in the United States, 1913–1998. *The Quarterly Journal of Economics* 118, 1 (2003), 1–41.
[53] 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.
[54] Benjamin Recht. 2018. A Tour of Reinforcement Learning: The View from Continuous Control. arXiv e-prints, art. *arXiv preprint arXiv:1806.09460* (2018).
[55] C. J. Roberts. 2013. Shelby County, Alabama v. Holder, attorney general, et al. No. 12-96, Argued February 27, 2013—Decided June 25, 2013.
[56] 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.
[57] Melissa L Sands. 2017. Exposure to inequality affects support for redistribution. *Proceedings of the National Academy of Sciences* 114, 4 (2017), 663–668.
[58] Simo Särkkä. 2013. *Bayesian filtering and smoothing*. Number 3. Cambridge University Press.
[59] 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.
[60] 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.
[61] Dylan Slack, Sorelle A Friedler, and Emile Givental. 2020. Fairness warnings and fair-MAML: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 200–209.
[62] Christina Starmans, Mark Sheskin, and Paul Bloom. 2017. Why people prefer unequal societies. *Nature Human Behaviour* 1, 4 (2017), 1–7.
[63] Dominique Thomas. 2019. Black lives matter as resistance to systemic anti-Black violence. *Journal of Critical Thought and Praxis* 8, 1 (2019).
[64] 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/
[65] US Supreme Court. 2003. Grutter v. Bollinger. 539, No. 02-241 (2003), 306.
[66] Min Wen, Osbert Bastani, and Ufuk Topcu. 2019. Fairness with Dynamics. *arXiv preprint arXiv:1901.08568* (2019).
[67] Felicia Wong, K. Sabeel Raham, and Dorian Warren. 2020. Democratizing Economic Power to Break the Cycle of American Inequality. *Stanford Social Innovation Review* (2020).
[68] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th Intl. Conf. on World Wide Web*. 1171–1180.
[69] 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.
[70] Xueru Zhang and Mingyan Liu. 2020. Fairness in Learning-Based Sequential Decision Algorithms: A Survey. *arXiv preprint arXiv:2001.04861* (2020).
[71] Karl J. Åström and Richard M. Murray. 2005. *Feedback Systems: An Introduction for Scientists and Engineers* (2 ed.). Princeton University Press.