VIRGINIA JOURNAL OF LAW & TECHNOLOGY

SUMMER 2020 UNIVERSITY OF VIRGINIA VOL. 24, NO. 1

Beyond Bias: Artificial Intelligence and Social Justice

ROBERT H. SLOAN†

RICHARD WARNER††

† Professor and Head, Department of Computer Science, University of Illinois at Chicago.

†† Professor of Law, Chicago-Kent College of Law. We presented an earlier draft at the 2019 Northeast Privacy Scholars Conference, and we gratefully thank the commentator and presenters for their insightful and helpful comments.

ABSTRACT

Artificial intelligence (AI) systems can discriminate against protected classes—a fact that has sparked an extensive literature about bias in AI. Bias, as important as it is, is a special case of the overall problem of social justice. *Beyond Bias* focuses on the general problem. It incorporates contributions from the extensive discussion of AI and fairness in the computer science literature. In particular, it draws on *Fairness Through Awareness*, an influential article by the Harvard computer scientist Cynthia Dwork and her co-authors. Adapting Dwork's approach, *Beyond Bias* reexpresses intuitive, well-motivated fairness constraints in a more mathematical way that shows how to apply the constraints to mathematically and computationally complex AI systems. The mathematics nonetheless uses only elementary arithmetic (unlike Dwork et al.).

Beyond Bias adapts the fairness constraints that it reexpresses from the Yale economist John Roemer. As Roemer notes in *Equality of Opportunity*, a conception of "equality of opportunity . . . prevalent today in Western democracies . . . says that society should do what it can to 'level the playing field' among individuals who compete for positions." *Beyond Bias* shows that AI systems can unfairly tilt the playing field. The reason lies in the pervasive (and unavoidable) use of "proxy variables"—e. g., using credit ratings to predict driving safety (as many insurance companies do). The credit ratings are the substitute—the proxy—for details about individuals' driving practices. *Beyond Bias* is the first article to apply a level playing field concept of fairness to issues of fairness in AI systems.

Beyond Bias briefly reviews the history of the use of proxy variables to evaluate consumers from the late Nineteenth Century to the present. It was already clear at the close of the Nineteenth Century that proxy-driven analysis could make seemingly unrelated aspects of one's life "have a profound impact on [one's] future potential in matters economic or social," as Dan Bouk notes in HOW OUR DAYS BECAME NUMBERED: RISK AND THE RISE OF THE STATISTICAL INDIVIDUAL. The concern was that proxy-driven analysis would unfairly tilt the playing field, and that concern continues to this day. *Beyond Bias* outlines a regulatory approach that ensures level playing field fairness by incorporating its mathematical constraints on AI systems.

TABLE OF CONTENTS

I. INTRODUCTION

Artificial intelligence (AI) systems can discriminate against protected classes—a fact that has sparked extensive literature about bias in AI.[1] Bias, as important as it is, is a special case of the overall problem of social injustice. We focus on the broader problem—more precisely, on one aspect of it: the use of proxy variables ("proxies"). A proxy variable is an easily measured input variable used instead of a desired input variable that is unobservable, or perhaps too costly to measure.[2] Suppose, for example, a teacher is interested in the length of time students pay attention in class. She cannot directly measure paying attention, so she uses proxies: taking notes, looking at material displayed on the board, and so on. As long as the proxy variables have some correlation with the unobservable variable, using them will improve the accuracy of

---

[1] *See, e.g.*, Margot Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 SOUTH. CALIF. LAW REV. 1529, 1537, n.14-17 (2019). *See also* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. LAW REV. 671 (2016); Danielle Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Joshua A. Kroll, et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017). Computer science literature has written on the topic as well. *See* Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV180800023 CS (2018), http://arxiv.org/abs/1808.00023; Cynthia Dwork et al., *Fairness Through Awareness*, *in* PROCEEDINGS OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE 214 (2012), http://doi.acm.org/10.1145/2090236.2090255; Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, *in* PROCEEDINGS OF INNOVATIONS IN THEORETICAL COMPUTER SCIENCE (ITCS) (2017), https://arxiv.org/abs/1609.05807v2; Maranke Wieringa, *What to Account for when Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability*, *in* PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1–18 (2020), https://doi.org/10.1145/3351095.3372833 (analyzing 243 articles in English from 2008 up to and including 2018).

[2] *See infra* Section I for important refinements of this preliminary explanation.

the prediction. The use of proxies is not new. The commercial credit reporting agencies that arose in the 1840s in the United States made extensive use of proxies to predict the likelihood of payment, for example.[3] The difference today is that proxy use has "gone viral." AI's "all-encompassing scope already reaches the very heart of a functioning society,"[4] and so do the proxies it uses. The consequence is that data from virtually any area of your life may serve as a proxy to make predictions about another seemingly disconnected area.[5] Imagine, for example, that Sally declares bankruptcy defaulting on a $50,000 credit card debt. The debt was the result of paying for lifesaving medical treatment for her eight-year-old daughter, and despite her best efforts, she could not make even the minimum payments. Assume post-bankruptcy Sally is a good credit risk—her daughter having recovered, but given her bankruptcy, a credit scoring system predicts that she is a poor risk. Her insurance company, which uses her credit rating as a

---

[3] *See infra* Section I.

[4] ERIC SIEGEL, PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE 293 (2016). Siegel identifies one hundred and forty seven examples of different types of use. *Id.* xv-xiv (listing examples, including the extension of credit, marketing and advertising, judicial sentencing and parole decisions, searching travelers, auditing taxpayers, police scrutiny of individuals and neighborhoods, welfare and financial aid, public health decisions, employee hiring, visa decisions, political campaign decisions, business planning and supply chain management, call center treatment, employee scheduling, evaluation of teachers, and ranking of the value of customers for differential treatment). *See also* STEVEN FINLAY, ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR BUSINESS: A NO-NONSENSE GUIDE TO DATA DRIVEN TECHNOLOGIES 9 (2nd ed. 2017) ("Today, machine learning is being applied to a huge range of problems. In fact, almost any aspect of life that involves decision making in one form or another"); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (Reprint ed. 2016); CATHY O'NEIL, ON BEING A DATA SKEPTIC (2013); Kroll et al., *supra* note 1.

[5] *See, e.g.*, HANNAH FRY, HELLO WORLD: BEING HUMAN IN THE AGE OF ALGORITHMS 47 (2018) ("the reach of these kinds of calculations now extends into virtually every aspect of society"). *See also infra* Section I.

proxy for safe driving (as many insurance companies in fact do)[6], increases her premium.[7]

Is it fair that saving her daughter's life makes Sally pay more for car insurance? In what sense of fairness? There are different concepts.[8] We confine our attention to one: a particular interpretation of fairness as equality of opportunity. As the economist John Roemer notes, a conception of "equality of opportunity . . . prevalent today in Western democracies . . . says that society should do what it can to 'level the playing field' among individuals who compete for positions."[9] Roemer, for example, notes that lower socio-economic status tilts the playing field against access to educational opportunities and proposes mechanisms to equalize access.[10] Most people think there are some attributes for which society should level the playing field, although they often disagree about which ones they are.[11] Our point is that the "Which attributes?" question arises in a sweeping, across the board way as contemporary proxy-driven AI makes what happens in one area of one's life reverberate through the rest in ways that can dramatically tilt the playing field.

What proxies ought to be allowed? We take it for granted that social and political processes settle that normative question. What are the appropriate processes? We offer four

---

[6] Consumer Reports, *Special Report: Car Insurance Secrets*, CONSUMER REPORTS, 2015, https://www.consumerreports.org/cro/car-insurance/auto-insurance-special-report/index.htm.

[7] *See, e.g.*, Leslie Scism & Mark Maremont, *Insurers Test Data Profiles to Identify Risky Clients*, WALL ST. J., November 19, 2010, http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html.

[8] Abigail Z. Jacobs & Hanna Wallach, *Measurement and Fairness*, ARXIV191205511 CS, 15–19 (2019), http://arxiv.org/abs/1912.05511 (noting that there are different conceptions of fairness and analyzing the impact of the failure to distinguish them on the computer science literature on fairness).

[9] JOHN E. ROEMER, EQUALITY OF OPPORTUNITY 1 (2000).

[10] *Id.* at 74-83.

[11] *See, e.g.,* ERIK OLIN WRIGHT & MICHAEL BURAWOY, HOW TO BE AN ANTICAPITALIST IN THE TWENTY-FIRST CENTURY 10 (2019).

criteria of adequacy and propose a regulatory process that meets them. It bears emphasis that our approach makes preserving informational privacy a constraint on the fairness of AI systems. Informational privacy consists of your ability to control what others do with information about you.[12] AI's use of proxies is a dramatic example of the loss of that control. Our approach restores control. Privacy concerns are sometimes best addressed from the broader perspective of social justice.

A note on terminology: We use "artificial intelligence" in a way that it is frequently used in business and the popular press as being equivalent to or including machine learning.[13] An accurate description of machine learning is that it is "the use of mathematical procedures (algorithms) to analyze data. The aim is to discover useful patterns . . . between different items of data. Once the relationships have been identified, these can be used to make inferences about the behavior of new cases."[14]

Section I gives a fuller explanation of proxy variables and briefly reviews their use from the Nineteenth Century into

---

[12] *See* ALAN WESTIN, PRIVACY AND FREEDOM 7 (1967); *see also* DOJ v. Reporters Comm. for Freedom of the Press, 489 U.S. 749, 763 (1989) ("both the common law and the literal understandings of privacy encompass the individual's control of information concerning his or her person"); JAMES B. RULE, PRIVACY IN PERIL: HOW WE ARE SACRIFICING A FUNDAMENTAL RIGHT IN EXCHANGE FOR SECURITY AND CONVENIENCE 3 (2007) (defining privacy "as the exercise of an authentic option to withhold information on oneself"); Michael Froomkin, *The Death of Privacy*, 52 STAN. L. REV. 1461, 1463 (2000) ("I will use 'informational privacy' as shorthand for the ability to control the acquisition or release of information about oneself").

[13] *See, e.g.,* Louis Columbus, *Roundup Of Machine Learning Forecasts And Market Estimates,* FORBES (Feb. 18, 2018), https://www.forbes.com/sites/louiscolumbus/2018/02/18/roundup-of-machine-learning-forecasts-and-market-estimates-2018/ (noting that "61% of organizations most frequently picked Machine Learning/Artificial Intelligence as their company's most significant data initiative for next year").

[14] STEVEN FINLAY, ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FOR BUSINESS: A NO-NONSENSE GUIDE TO DATA DRIVEN TECHNOLOGIES 6 (2nd ed. 2017).

the Twenty-First. Section II explains and motivates our appeal to level playing field fairness and uses it to formulate two conditions of adequacy on any approach to regulating AI systems. The section concludes by recasting those two conditions as three requirements formulated in a way that facilitates their application to computationally and mathematically complex AI systems. The reformulation draws on an influential computer science article, Dwork et al.'s *Fairness as Awareness*.[15] Section III uses that reformulation to address the question of how regulators can acquire the information they need to assess the fairness of AI systems. That section adds a fourth requirement to the previous section's three. The fourth requirement puts the burden on the user of an AI system to provide regulators sufficient information to assess the fairness of the system. Section IV offers four criteria on any adequate regulatory approach to assessing an AI system's fairness and suggests the Federal Trade Commission plausibly meets those criteria. Section V concludes by emphasizing the urgency and importance of finding an effective way to ensure the fairness of AI systems.

II. LEVEL PLAYING FIELD FAIRNESS

We do not offer our version of a level playing field fairness as a comprehensive theory of social justice. We offer it only as a plausible component of social justice in societies in which market economies allocate socio-economic positions based on a person's attributes such as talent and degree of effort.

A. Explanation and Motivation

To motivate and explain the conception, let us say that persons are advantaged to the extent that they have attributes that improve their likelihood of material success in the market economies typical of Western democracies. Persons are disadvantaged to the extent they lack those attributes or have

[15] Dwork et al., *supra* note 1.

attributes that decrease the likelihood of success.[16] One levels the playing field with regard to a set of attributes to the extent one structures society to reduce the advantage or disadvantage those attributes confer. The basic rationale Roemer gives for leveling the playing field is that people may be advantaged or disadvantaged as a result of circumstances beyond their control.[17] One relatively uncontroversial example is education.[18] People are advantaged by an adequate education, and whether a person acquires one is dependent on a number of factors beyond that person's control. There is widespread agreement that compulsory and voluntary educational opportunities are appropriate ways to narrow the gap between the more educated and the less educated.[19] Additional, more controversial, examples of attributes include access to health care, and the availability of unemployment insurance.[20]

Two further points are in order. The first is that not all examples of leveling the playing field fit comfortably with the "beyond one's control" rationale. Bankruptcy is a case in point.

There are multiple reasons to have the institution of bankruptcy, but leveling the playing field is one. As the United States Supreme Court notes, bankruptcy "gives to the honest but unfortunate debtor . . . a new opportunity in life and a clear field for future effort, unhampered by the pressure and discouragement of preexisting debt."[21] Despite the Court's characterization of bankrupt debtors as "honest but unfortunate," not all bankruptcies are the result of circumstances beyond bankrupts' control. Sally arguably faces such circumstances when her daughter's life is at stake, but compare Roger. He declares bankruptcy after defaulting on $50,000 of credit card debt, which he incurred by spending

---

[16] A more sophisticated approach would sort attributes into different types and combinations with different likelihoods of success, but we need not do so here. *See* ROEMER, *supra* note 9.
[17] *Id.* at 19.
[18] *Id.* at 16 and Chapters 9 and 11.
[19] *See, e.g.*, *id.* at 54.
[20] *Id.* at Chapters 8 and 10.
[21] Local Loan Co. v. Hunt, 292 U.S. 234, 244 (1934).

well beyond his means on five-star hotels and expensive restaurants. His plan was to declare bankruptcy when the borrowed money ran out. One plausible "level playing field" rationale for bankruptcy is the severity of the consequence of lacking a favorable credit rating. One may struggle to buy or rent a home or a car, start a business, pay for higher education, find employment, or buy insurance.

Some suggest formulations that expand the reach of level playing field fairness well beyond the confines of what is beyond a person's control. For example, "In a just society, all persons would have broadly equal access to the material and social means necessary to live a flourishing life."[22] There is no need to settle on an exact formulation. It is enough to note that most people think there are some attributes with regard to which society should level the playing field.[23] People of course disagree about which attributes.[24]

The second point concerns the way in which people traditionally interpret level playing field views. As Roemer explains:

> Among the citizens of any advanced democracy, we find individuals who hold a spectrum of views with respect to what is required for equal opportunity, from the nondiscrimination view at one pole to pervasive social provision to correct for all manner of disadvantage at the other. Common to all these views, however, is the precept that the equal-opportunity principle, at some point, holds the individual accountable for

---

[22] WRIGHT & BURAWOY, *supra* note 11, at 10.

[23] *Id.*

[24] *See* ROEMER, *supra* note 9, at 2 ("More specifically, different people have different conceptions about where the starting gate should be, or about the degrees to which individuals should be held accountable for the outcomes or advantage they eventually enjoy. My purpose is to propose an algorithm which will enable a society (or a social planner) to translate any such view about personal accountability into a social policy that will implement a kind or degree of equal opportunity consonant with that view").

> the achievement of the advantage in question, whether that advantage be a level of educational achievement, health, employment status, income, or the economist's utility or welfare. Thus, there is, in the notion of equality of opportunity, a "before" and an "after": before the competition starts, opportunities must be equalized, by social intervention if need be, but after it begins, individuals are on their own.[25]

Should it be the case that "individuals are on their own" after competition starts provided opportunities are equalized? The answer matters greatly to our eventual regulatory proposal. Our answer is *no*. Even if opportunity is equal before the competition starts, competition can still unfairly tilt the playing field. Some may object that this is not possible if the market is sufficiently competitive and opportunity is equal before competition starts. As the political economist Charles Lindblom notes:

> In our day, perhaps most people would not defend the rule [that individuals are on their own after competition starts] unless it were supplemented by other rules and procedures, such as those of the welfare state. Some people nevertheless defend the rule as itself sufficient without such supplements as pensions and unemployment compensation.[26]

Lindblom continues that, although people offer several defenses of the rule that individuals are on their own, "all are flawed; and the flaws represent misunderstandings of the

---

[25] *Id.*

[26] CHARLES E. LINDBLOM, THE MARKET SYSTEM: WHAT IT IS, HOW IT WORKS, AND WHAT TO MAKE OF IT 115–116 (2002). We substituted "the rule that individuals are on their own" for Lindblom's "the rule of quid pro quo." The idea that, once opportunities are equalized, individuals are on their own is a special case of what Lindblom means by "the rule of quid pro quo," and all of Lindblom's criticisms apply.

market system."[27] Relying on Lindblom's critique, we take it for granted that market mechanisms will not be sufficient to prevent all "after start" unfair tilts.[28]

We add that bankruptcy's goal of providing "a new opportunity in life and a clear field for future effort, unhampered by the pressure and discouragement of preexisting debt"[29] is an instance of society recognizing a need to adjust equality of opportunity *after* competition starts. More generally, the history of consumer surveillance since the Nineteenth Century rise of credit reporting is in part a history of concern about unfairly tilted playing fields as consumer surveillance make it increasingly likely that one area of one's life would impact other areas. The concern is clearly with "after start of competition" unfairness. The worry is that as your market activity unfolds some combination of circumstances in one area will—unfairly—reverberate to your disadvantage throughout a wide range of other areas. There are plausible examples of such unfairness. One is the Sally bankruptcy example. Another is the often used example of American Express's 2009 lowering of credit limits based on the stores in which the cardholder shopped. American Express determined that cardholders who shopped in certain stores were less likely to repay than cardholders who did not patronize those stores.[30] This disadvantaged the cardholders (in our sense of reduced likelihood of market success). They not only had less borrowing power, but also likely had their credit rating lowered and the borrowing costs increased. Some may not find

---

[27] *Id.* at 116.

[28] *See id.* at Chapter 8 (discussing how market competition can create unfairness). *See also* Adam Pham & Clinton Castro, *The Moral Limits of the Market: The Case of Consumer Scoring Data*, 21 ETHICS INF. TECHNOL. 117 (critiquing consumer scoring systems). *See also* Alan Rubel, Clinton Castro & Adam Pham, *Agency Laundering and Information Technologies*, 22 ETHICAL THEORY MORAL PRACT. 1017 (2019) (providing a moral critique regarding the use of AI systems).

[29] Local Loan Co. v. Hunt, 292 U.S. 234, 244 (1934).

[30] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 156-157 (reprt.ed. 2016) (noting that American Express left affected cardholders "careening into a nasty recession with less credit").

these examples convincing. Although people will disagree on particular examples, there is widespread agreement that proxy-driven AI systems can generate "after start" unfair tilts of the playing field.[31]

B. Two Conditions On AI Systems

We formulate two conditions on any adequate process for regulating proxy-driven AI systems. The first is that the process should identify "before start" attributes whose use in AI systems would unfairly tilt the playing field and then prohibit their use in AI systems. For example, suppose a society attempts to level the playing field in part by ensuring that everyone has an equal access to educational opportunities. To this end, it uses remedial reading programs to ensure that all elementary school children have the same basic reading skills. It would be unfair for an AI system to assign a lower score (which would translate into being less likely to be hired) to applicants who participated in a remedial reading program.

The second condition concerns the regulation of "after start of competition" attributes. One cannot regulate "after start of competition" tilts in the way we suggested regulating "before start" tilts, where we simply prohibited the use of certain attributes. Consider the attribute of bankruptcy. It plausibly unfairly tilts the playing field in some cases in which an AI car insurance program assigns higher premiums to applicants who declared bankruptcy. Compare a program that uses the attribute of bankruptcy plus a combination of others to distinguish between bankruptcies like Sally's and bankruptcies like Rogers' and then avoids assigning a higher premium to the Sally-like applicants on the basis of their bankruptcy. The second program arguably avoids unfairly tilting the playing field against the Sally-like applicants. "After start" tilts are the result of the *particular way* the AI program uses variables to allocate costs and benefits. Some allocations tilt the playing

---

[31] *See supra* note 62. *See also* MULLER, *supra* note 36; CATHY O'NEIL, ON BEING A DATA SKEPTIC (2013); SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER (2019).

field, some do not. Thus, the second condition is that the process should identify which uses of "after start" attributes unfairly tilt the playing field and prohibit those uses.

How is a regulator supposed to implement these two conditions? Faced with the computational and mathematical complexity of AI systems, how do you tell when and how they unfairly tilt the playing field? To answer, we recast these requirements in a somewhat more mathematic mode by adapting ideas from the computer scientists Dwork et al.'s influential article *Fairness Through Awareness*.[32] Dwork et al. offer a mathematical formulation of a Roemer-inspired approach to fairness.

C. A Reformulation

An example is helpful in explaining the approach. We use a highly simplified model for setting auto insurance premiums. The model uses the four variables represented in bold type headings in the table below (excluding the identifier "Name"). We include the "Bankruptcy" heading for simplicity. A more realistic model would replace it with a heading for "Credit Rating" or something similar. Assume the model treats the attributes in the headings as proxy variables for something like how carefully a person drives. These assumptions are for illustrative purposes only. We do not claim they are valid.

Name	**Age**	**Income**	**Occupation**	**Bankruptcy**
Sally	35	88,000	Software Engineer	Yes
Roger	40	65,000	Police Officer	No

[32] Dwork et al., *supra* note 1.

The system uses an algorithm that assigns a numerical score to any combination of the four variables. Thus, Sally and Roger are each assigned a score. Assume the scores run from 1 to 100, with higher scores indicating higher premiums. Suppose Sally's score is 75 (given her bankruptcy) while Roger's is 10. Subtracting Roger's score from Sally's represents the system's proxy-based determination of the relevant difference between them in regard to the assignment of premiums—in this case, 65. (It is convenient to have the difference always be a positive number, so when subtracting Sally from Roger take the absolute value to also get 65). In this way, for any two applicants *x* and *y*, the system establishes a distance *d(x, y)* between x and y. We will call *d(x, y)* a system's *distance metric* (borrowing from Dwork et al.[33] and the mathematics of metric spaces[34]).

---

[33] *Id.* at 216 ("To introduce our notion of fairness we assume the existence of a metric on individuals"). Dwork et al. assume that a relevant distance metric is available for *each* AI system that classifies individuals in a way that will be used to allocate costs and benefits to individuals. They note that "one of the most challenging aspects of our work is justifying the availability of a distance metric." *Id.* at 223. We have avoided the complexities of that challenge by limiting our attention to systems that assign numerical scores and use them to allocate costs and benefits. Dwork et al. also note the availability of distance metrics in such systems. *Id.* at 224 ("The imposition of a metric already occurs in many classification processes").

[34] Wikipedia provides an informal explanation of distance metrics. "In mathematics, a metric space is a set together with a metric on the set. The metric is a function that defines a concept of distance between any two members of the set, which are usually called points. The metric satisfies a few simple properties. Informally:

- the distance from a point to itself is zero,
- the distance between two distinct points is positive,
- the distance from A to B is the same as the distance from B to A, and
- the distance from A to B (directly) is less than or equal to the distance from A to B via any third point C."

Metric Space, WIKIPEDIA https://en.wikipedia.org/wiki/Metric_space (last visited June 7, 2020).

We can reformulate the two conditions on regulating AI in terms of distance metrics. The first condition becomes the following requirement:

> *Requirement 1*: For any system, the system must generate its distance metric in ways consistent with "before start" equal opportunity.

The second condition is to identify and prohibit those uses of "after start" attributes that unfairly tilt the playing field. We divide this into two requirements. To formulate the first, let *A(x)* be the cost or benefit that the system allocates to *x*. Assume, for the moment, that one can measure the extent of the cost or benefit quantitatively—on a very broad understanding of what cost/benefit analysis can include. It can include "everything that matters to people's welfare, including such qualitatively diverse goods as physical and mental health, freedom from pain, a sense of meaning, culture, clean air and water, animal welfare, safe food, pristine areas, and access to public buildings."[35] For illustrative purposes, in the case of credit rating systems, *A(x)* could be a credit rating; in the case of car insurance, a premium.

A more mathematical formulation of the second condition is, roughly speaking, that for any two individuals *x* and *y* we have $A(x) - A(y) \leq d(x, y)$.[36] Two problems make this rough speaking—one trivial, one slightly less so. The trivial one is that $d(x, y)$ is always a nonnegative number while $A(x) - A(y)$ will be negative when $A(y) > A(x)$. The easy fix is to use the absolute value $|A(x) - A(y)|$. The slightly more difficult problem is that $|A(x) - A(y)|$ and $d(x, y)$ may measure on different scales. Suppose $|A(x) - A(y)|$ is a number from one into the thousands while $d(x, y)$ is a number from one to one hundred. The solution is easy: reexpress the measures on a common scale.[37] We will write $|A(x) - A(y)| \leq d(x, y)$, assuming

---

[35] CASS R. SUNSTEIN, THE COST-BENEFIT REVOLUTION 23 (2019).

[36] The allocation might be probabilistic, so a more precise description would require that the *expected value of A(x) – A(y)* be at most *d(x, y)*.

[37] For an accessible overview of scaling, *see Feature scaling*, WIKIPEDIA, https://en.wikipedia.org/wiki/Feature_scaling (last visited June 7, 2020).

the choice of some suitable scaling method that makes the comparison meaningful. Thus, we propose:

> *Requirement 2*: For a system S, it must be the case that $|A(x) – A(y)| \leq d(x, y)$.

Requirement (2) says that differences in allocations must be no larger than differences between individuals. Fairness requires this. Otherwise, the allocations do not treat like cases alike. Suppose, for example, an auto insurance system finds little difference between Alice and Bob, but assigns Bob a high premium and Alice a low one. [38] Requirement (2) prohibits such systems. Two further points are in order.

---

[38] Dwork et al. also require that like cases be treated alike. *Supra* note 1. They offer Definition 2.1 below to implement the "*fairness constraint*, that similar individuals are treated similarly." *Id.* at 214. As they explain: "To introduce our notion of fairness we assume the existence of a metric on individuals d: $V \times V \rightarrow$ **R.** We will consider randomized mappings M: $V \rightarrow \Delta(A)$ from individuals to probability distributions over outcomes. Such a mapping naturally describes a randomized classification procedure: to classify $x \in V$ choose an outcome a according to the distribution $M(x)$. We interpret the goal of 'mapping similar people similarly' to mean that the distributions assigned to similar people are similar. Later we will discuss two specific measures of similarity of distributions, D_∞ and D_{tv}, of interest in this work.

> Definition 2.1 (Lipschitz mapping). *A mapping* M: $V \rightarrow \Delta(A)$ *satisfies the* (D,d)-Lipschitz *property if for every* $x,y \in$ V, *we have* $D(Mx, My) \leq d(x, y)$. (1)

. . . We note that there always exists a Lipschitz classifier, for example, by mapping all individuals to the same distribution over A." *Id.* at 216.

Definition 2.1 applies far more generally than our Requirement 2, which we apply only for consumer scoring systems meeting the conditions (a) and (b) given earlier. *See supra* text accompanying note 47. Dwork at al. handle the general case in which there is a probabilistic assignment of individuals to a set of outcomes. *Id.* at 215 Consider for example a system that allocates the display of advertisements to website visitors. It might, for example, assign Alice a probability p1 of being shown ad 1, a probability p2 of being shown ad 2, and a probability p3 of being shown ad 3. Dwork et al. require that if two individuals are pretty similar, then their respective p1, p2, and p3 should be pretty similar. *Id.*

First, in formulating Requirement (2) we assumed that the allocations *A(x)* were quantifiable. Meaningful quantification of costs and benefits is often both possible and desirable when framing policies that affect millions,[39] but there are of course "values that are difficult or impossible to quantify, including equity, human dignity, fairness, and distributive impacts."[40] The list could be much longer. One could add—to take just a few examples—friendship, beauty, meaningful work, kindness, and "capacity shown, in some form or other, by humans in all cultures to live under rules and values and to shape their behavior in some degree to social expectations, in ways that are not under surveillance and not directly controlled by threats and rewards."[41]

So how does this apply when *A(x)* is not quantifiable? Replace $|A(x) - A(y)|$ with a qualitative comparison of the allocations, and a qualitative comparison of how well the difference in allocations matches the difference the system encodes in $d(x, y)$. Comparisons of non-quantifiable values are routine in framing public policy. For example, Executive Order 13563—Improving Regulation and Regulatory Review requires regulators to "take into account benefits and costs, both quantitative and qualitative."[42]

---

[39] For relevant considerations, *see* SUNSTEIN, *supra* note 69.

[40] Exec. Order No. 13,563, 76 Fed Reg. 3821 (Jan. 18, 2011).

[41] BERNARD WILLIAMS, TRUTH AND TRUTHFULNESS: AN ESSAY IN GENEALOGY 24 (2004).

[42] *Supra* note 74. *See also* STUART HAMPSHIRE, INNOCENCE AND EXPERIENCE (1991) (offering a well worked out view of how to include qualitative considerations in debates about public policy. *But cf*. SUNSTEIN, *supra* note 69. Sunstein acknowledges that "It is true that moral commitments often signal values that are not adequately captured by private willingness to pay." *Id.* at 104. He nonetheless argues that regulators should consider quantifying the cost of meeting or violating moral commitments—at least in a wide range of cases. His argument rests on examples of a vast increase in the cost of legislation to protect a moral concern of relatively small importance. We agree that the cost of the legislation argues against protecting the moral concern, but do not think one can infer from such examples to the *general* conclusion that should consider quantifying the cost of meeting or violating moral commitments.

The second point is that Requirement (2) is not sufficient to ensure that a system does not unfairly tilt the playing field. We give an example of a system that obeys Requirement (2) but is not fair. Start with a system S, and assume for sake of argument, that S's allocation function makes *fair* assignments. Assume in particular that S assigns the high car insurance premium H to all individuals in category C1, who have a history of traffic violations, and S assigns the low premium L to all individuals in C2, who have no history of traffic violations (nor any other indicia of careless driving). Now create a new system S' which is like S except S' assigns L to the careless drivers in C1 and H to the careful drivers in C2. It is *still* true that $|A(x) - A(y)| \leq d(x, y)$. However—grant for the moment—treating the careful drivers in category C2 as if they were the careless drivers in category C1 unfairly tilts the playing field against the careful drivers. The problem is that Requirement (2) does not guarantee the *direction* of the allocations—which category gets more or less.[43] If you do not find changing the insurance premiums sufficient for a clear example of unfairness, add additional circumstances and consequences as we did in the Sally example to make the effects more severe. To address the fact that Requirement (2) does not ensure fairness, we add a third requirement:

> *Requirement 3*: S's allocation function $A(x)$ does not unfairly tilt the playing field.

This may seem disappointing. We set out to explain how a regulator could fit fairness requirements onto the computational and mathematical complexity of AI systems, but, at a crucial point, we still require a judgment of fairness. This is unavoidable. One cannot reduce questions of fairness to

[43] The direction of the allocations matters when treating like cases alike fails to award people what they merit. The careful drivers in C2 merit lower premiums than the careless drivers in C1, but S' fails to make assignments in accord with merit. Compare this example with the display of advertising example, outlined *supra* note 72. There may be no clear sense in which individuals merit the display of one ad instead of another.

a mathematical test.[44] What one can do is provide a workable framework in which regulators are to make that judgment. Conditions (1) and (2) provide part of that framework. We add one more condition in the next section.

The next section begins with the question of whether regulators will be able to learn enough about AI systems to apply (1) - (3). Assessing compliance with (1) – (3) requires relevant information about a systems distance metric *d(x,y)* and its allocation function *A(x)*. How will regulators have access to that information? The discussion leads to the addition of a fourth requirement.

III. TRANSPARENCY

There is a standard answer to the question of how regulators should learn what they need to know about AI systems: namely, the systems should be "transparent." The problem is to explain what transparent means. "Legal scholars have argued for twenty years that automated processing requires more transparency, but it is far from obvious what form such transparency should take."[45] We explain our use of "transparent" by analogy. A physical thing is transparent if you can see through it. We explain our use of transparency by answering two questions. What do *regulators* need to see when they look into AI systems? And, what do *consumers* need to see? We begin with the question about regulators.

A. False Hope: Knowledge of Source Code Will Suffice

[44] Doing so is not possible where it would require meaningful quantification of attributes that cannot be meaningfully quantified. *See generally* notes 73-75 and accompanying text; for additional considerations, *see* Ben Green & Lily Hu, *The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning* (Int. Conf. on Machine Learning 2018).

[45] Kroll et al., *supra* note 1, at 638.

It is common to assume that revealing an algorithm's source code will reveal to regulators what they need to know.[46] This is false. As Kroll et al. note, "The source code of computer systems is illegible to nonexperts."[47] If the code were legible to experts, their reports could make algorithms consumer-transparent However, "even experts often struggle to understand what software code will do, as inspecting source code is a very limited way of predicting how a computer program will behave."[48] Indeed, some approaches, including some that are very popular, such as support vector machines and deep learning of neural nets, give predictive models that are quite difficult for humans to comprehend.[49]

Fortunately, access to source code is not necessary for regulators to have sufficient information about a system's distance metric and allocation function.

B. Regulator Access to Information

To see how regulators can have access to relevant information, we distinguish traditional machine learning from deep learning:

> The traditional machine learning approach is characterized by practitioners investing the bulk of their efforts into engineering features. This feature engineering is the application of clever, and often elaborate, algorithms to raw data in order to preprocess the data into input variables that can be readily modeled by traditional statistical techniques. These techniques . . . are seldom effective on unprocessed data, and so

---

[46] *Compare* Citron & Pasquale, *supra* note 1, at 14 (insisting on the need to know source code), *with* Kroll et al., *supra* note 1, at 642–656 (discussing the assumption that one needs to know the source code and pointing out its difficulties), *and* Devan R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide To Algorithms And The Law*, 31 HARV. J. L. & TECH. 1, 9-12 (2017) (discussing difficulties with insisting on knowing source code).

[47] Kroll et al., *supra* note 1, at 638.

[48] *Id.*

[49] FINLAY, *supra* note 46, at 126.

> the engineering of input data has historically been a prime focus of machine learning professionals.[50]

In such cases, the proxy variables (at least a significant number of them[51]) will be readily available—at least as long as the system's creators have adequately documented their process of creation.[52]

Contrast deep learning approaches where the "practitioner typically spends little to none of her time engineering features, instead spending it modeling data with various artificial neural network architectures that process the raw inputs into useful features automatically."[53] In deep learning, the proxy variables may not be readily identifiable.[54] There is however recent work devoted to explaining the operation of deep learning systems.[55]

We propose placing the burden the users of AI systems to ensure that sufficient information is available to regulators. If a user fails to meet burden, the system is presumptively unfair. Thus:

> *Requirement 4*: Users of an AI system S have the burden to provide sufficient about S's distance metric and allocation function to assess compliance with the

---

[50] JON KROHN, GRANT BEYLEVELD & AGLAÉ BASSENS, DEEP LEARNING ILLUSTRATED: A VISUAL, INTERACTIVE GUIDE TO ARTIFICIAL INTELLIGENCE 44–45 (2019).

[51] The operation of the system may generate additional variables.

[52] *See, e.g.*, Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 33 (2020), https://doi.org/10.1145/3351095.3372873.

[53] KROHN, BEYLEVELD & BASSENS, *supra* note 84, at 45.

[54] *See, e.g.*, FINLAY, *supra* note 46, at 126.

[55] *See, e.g.*, Pieter Jan Kindermans et al., *Learning How to Explain Neural Networks: PatternNet and PatternAttribution* (2018), https://openreview.net/forum?id=Hkn7CBaTW; Chun-Hao Chang et al., *Explaining Image Classifiers by Counterfactual Generation*, ARXIV (2019), http://arxiv.org/abs/1807.08024.

requirements (1) – (3). Failure to do so makes the system presumptively unfair.

We argue for (4) in the next subsection.

C. Why Lack of Information Is Presumptively Unfair

The rationale for treating lack of relevant information as presumptively unfair is consumers' need for informational privacy. Informational privacy consists of the ability to control what others do with information about you.[56] It is important to restore an adequate degree of informational privacy. Like many, we think that adequate informational privacy is essential to human freedom and well-being,[57] but we will not argue the point here. We take the importance of informational privacy for granted.

Proxy-driven AI systems significantly reduce that control in which informational privacy consists. By way of illustration, suppose Sally is searching online for the best buy on an air purifier. She realizes that consumer scoring systems operating in the background use information about her. She could avoid that by not searching online at all, but she is not willing to give up the convenient access to information. She strongly prefers that they use her information fairly, but she lacks any way to control the systems. She lacks information about how the systems work, and, even if she had it, she lacks the knowledge and expertise to analyze and understand it. Further, even if she had the relevant knowledge and expertise, she does not have time to spend on the analysis. She is already committed to a variety of goals—raising her daughter, pursuing

---

[56] *See supra* note 12.

[57] The connection between informational privacy and freedom and the self is a standard theme in the privacy literature. *See, e.g.*, DANIEL J. SOLOVE, UNDERSTANDING PRIVACY 112 (2008) ("Theorists have proclaimed the value of privacy to be protecting intimacy, friendship, individuality, human relationships, autonomy, freedom, self-development, creativity, independence, imagination, counterculture, eccentricity, thought, democracy, reputation, and psychological well-being").

her career, enjoying her friends, and so on—and the time she is willing to allot to buying an air purifier is relatively short.

Under these conditions, how can one provide Sally with control over what the systems do with her information to ensure adequate informational privacy? She is not *individually* in a position to control what the systems do, but that does not preclude control through *collective* political action, in particular the collective action of the relevant regulatory processes. In the next section, we outline such a regulatory process. It controls AI systems to ensure that they use information fairly, its authoritative assurances of fairness provide Sally—and consumers generally—with the knowledge they want and need.

Before we turn to that task, two questions remain. First, our proposal is to treat the failure to provide sufficient information to assess a system as making the system unfair. But could an AI system not be fair even though no one has the information necessary to know that it is? To answer, suppose that you are subject to decisions of AI systems, and suppose—through no fault or failure on your part—that you do not know whether they are fair. If you knew you were on a fair playing field, you would know (or could know[58]) that the systems do not use certain "before start" attributes and that their use of "after start" attributes is constrained in ways that constrain the systems' use of information about certain areas of your life to assign costs and benefits in other areas. In light of that knowledge, you could plan accordingly. We take that interference *itself* to be unfair. It is unfair to be denied knowledge of whether you are on a fair or unfair field. The

---

[58] We argue in the next section that the regulatory agency that determines whether AI systems are fair should provide adequate reasons for their judgments. Those reasons make it possible for consumers to know why systems are fair or unfair. There are interesting and important questions about what consumers know, how they know it, and how they use that knowledge. Those questions lie far beyond the scope of this article. *See, e.g.*, ROBERT H. SLOAN & RICHARD WARNER, THE PRIVACY FIX: HOW TO PRESERVE PRIVACY IN THE ONSLAUGHT OF SURVEILLANCE, Cambridge University Press (forthcoming 2021).

claim is not controversial. Suppose Roger invites Sally to play poker. Roger informs her that Alice, who is observing, may or may not declare certain cards that Roger holds wild. Roger says he knows what Alice will do, but when Sally asks Roger to pass that knowledge on to her, Roger replies, "That is for me to know and you to find out." Sally responds, "That is not fair!", and she is right. Roger's proposal leaves her not knowing whether she is playing a fair game without intervention by Alice, or one that Alice unfairly biases in Roger's favor. That lack of knowledge *itself* makes Roger's proposed game one unfairly tilted against Sally.

The second question is why failing to provide information relevant to assessing fairness creates only a *presumption* of unfairness. The rationale is to allow users to answer a charge of unfairness by appealing to the consequences of using the system. The argument would be: (1) In the past, the benefits of using the system significantly outweighed the costs, and (2) it is likely that that the benefits will continue to do so in the future. We note in passing that such justifications may be far less readily available than it may at first appear. It is *possible* for (1) to be true. Imagine systems that cure diseases, restore the climate, eliminate starvation, and order social relations in ways that yield a vibrant culture in which all have satisfying opportunities for self-realization. But AI systems' results are likely to be much more mixed, as we have argued elsewhere.[59] (2) is also problematic. Past decisions may provide little indication of future ones. A system's predictions are a function of the data it takes as input, and the algorithm it employs.[60] Both are likely to change over time. "Predictive models tend to deteriorate over time—their ability to predict gets worse as economic, market and social change occurs. The relationships that were found between the predictor

---

[59] Robert H. Sloan & Richard Warner, *Algorithms and Human Freedom*, 35 SANTA CLARA HIGH TECH. L. J. 1 (2019).

[60] *See, e.g.*, JOHN D. KELLEHER & BRENDAN TIERNEY, DATA SCIENCE 143–144 (2018) ("Two major factors contribute to the [prediction] . . . that an ML [machine learning] algorithm will generate from a data set. The first is the data set the algorithm is run on . . . The second . . . is the choice of ML algorithm").

data and the outcome data when the model was originally constructed no longer apply."[61]

IV. A REGULATORY PROPOSAL

We propose four criteria that any adequate regulatory scheme should meet and then suggest that regulation by—something like—the Federal Trade Commission (FTC) is a plausible way to meet them. The "something like" qualification is necessary because our proposal significantly expands the powers of the actual FTC. The actual FTC serves as a useful yardstick to measure the extent of the changes need to address AI systems.

A. Four Criteria of Adequacy

An adequate regulatory scheme in this context is sufficiently well equipped to evaluate AI systems in terms of the requirements (1) – (4). The first two criteria concern the ability to acquire and understand sufficient information about a system's distance metric *d(x, y)* and its allocation function *A(x)*.

1. Broad investigative powers

Businesses increasingly use ever more sophisticated AI systems in a rapidly changing technological and economic environment. Any regulatory approach that hopes to keep pace with the developments will have to be an agile one with broad investigative powers. There is a historical parallel with the creation of the Federal Trade Commission in 1914. The end of the Nineteenth Century saw reduced competition and price-fixing as a result of business growth and consolidation along with various forms of cooperation and collusion.[62] Congress perceived the need for a quick acting, flexible organization with broad investigative powers that could effectively regulate

---

[61] FINLAY, *supra* note 47, at 79.

[62] CHRIS JAY HOOFNAGLE, FEDERAL TRADE COMMISSION PRIVACY LAW AND POLICY 4 (2016).

those developments.[63] A similar need exists today in the case of AI systems.

2. Access to expertise

At the turn of the twentieth century, the complexities of regulating market competition called for regulation guided by adequate expertise.[64] At the turn of the twenty-first century, the rapidly increasing use of increasingly complex AI systems similarly argues for regulatory access to relevant technical and business expertise.

3. Market freedom within the constraints of fairness

Our third requirement is that a system's allocation function *A(x)* should not unfairly tilt the playing field. Determining the fairness of an AI system can involve assessing the effects of allocations of costs and benefits over large numbers of individuals. Those effects may be difficult to determine, and fairness judgments based on them can involve complex tradeoffs.[65] Among the factors a regulatory body should consider is market freedom. We take it for granted that, in a market economy, it is desirable, as Dwork et al. note, to "permit the entity that needs to classify individuals, which we call the vendor [of an AI system], as much freedom as possible . . . [to allow it] to benefit from investment in data mining and market research in designing its classifier."[66] We understand Dwork et al.'s "as much freedom as possible" to mean as much

---

[63] *Id.* at 4–10.
[64] *Id.* at 9.
[65] Fairly allocating resources can be quite complicated even when the only question is how to distribute a certain well defined benefit in a limited, specific context. *See, e.g.*, H. PEYTON YOUNG, EQUITY IN THEORY AND PRACTICE (1995). Peyton's discussion of demobilization after World War II is a good illustration of a complex distribution question. *Id.* at 23–27. For an example of a broader question of social justice, *see, e.g.*, VICTOR R. FUCHS, WHO SHALL LIVE? HEALTH, ECONOMICS AND SOCIAL CHOICE (Expanded ed. 2011); SUNSTEIN, *supra* note 69.
[66] Dwork et al., *supra* note 1, at 214.

freedom as possible consistent *with the demands of fairness* (which is Dwork et al.'s understanding as well[67]).

4. Reason-giving

Our fourth requirement is that, on pain of unfairness, users of an AI system must provide sufficient information about its distance metric and allocation function. The corresponding fourth criterion of adequacy is a constraint on how the regulatory body uses that information: namely, they should use it to articulate reasons why a system is fair or unfair. An agency could, of course, simply announce that a system is fair *without* giving supporting reasons. That may be acceptable when confidentiality is important. For example, when a university committee makes decisions about tenure, very few have access to the reasons for the decisions; but as long as those affected have adequate reasons to trust the process, they may find the decisions acceptable. Regulatory decisions about AI systems, however, should not put a premium on confidentiality. It comports better with the requirements of legitimate democratic governance to provide supporting reasons. "[D]ecision making based on reason … , not on preference or faith, is crucial for legitimacy."[68]

B. A Role for the Federal Trade Commission

We propose regulation by the FTC—more accurately, by an FTC-like agency—as a plausible way to meet the criteria of adequacy. The agency we envision is politically empowered and adequately funded with significantly expanded powers to make and enforce judgments of fairness. We envision regulation under 15 U.S. Code § 45(n), the FTC's standard for an unfair business practice. A system should not cause or be likely to cause "substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to

---

[67] *Id.* at 214 (claiming that their approach provides an "absolute guarantee of fairness").

[68] Steven Burton, *Reaffirming Legal Reasoning: The Challenge from the Left*, 36 J. LEGAL EDUC. 358, 368 (1986).

competition."[69] Our proposal is to see the consequences of unfairly tilting the playing field as constituting a "substantial injury," at least when those costs are great enough.

This proposal greatly expands the power of the FTC to make judgments of fairness. In the FTC's current practice "[s]ubstantial injuries to consumers usually . . . involve monetary harm, coercion into the purchase of unwanted goods or services, and health or safety risks."[70] Does our proposed expansion of what counts as a substantial injury grant the FTC too unconstrained a power to make judgments of fairness?[71] We leave that question unanswered. Our point is that AI systems raise pressing issues of level playing field fairness. Given, as we have assumed, that market mechanism will not be sufficient to ensure fairness, some more or less constrained regulatory processes will be needed to make the necessary judgments of fairness. We suggest a suitable interpretation of the FTC's substantial injury requirement as a plausible approach.

1. Broad investigative powers

The FTC has extremely broad powers to initiate investigations into a company's practices. "The FTC's investigatory power . . . is akin to an inquisitorial body. On its own initiative, it can investigate a broad range of businesses without any indication of a predicate offense having occurred."[72]

2. Access to expertise

The FTC has ready access to experts.[73]

---

[69] 15 U.S.C § 45(n) (2012).
[70] HOOFNAGLE, *supra* note 96, at 132.
[71] For a discussion of pros and cons, *see* HOOFNAGLE, *supra* note 96.
[72] HOOFNAGLE, *supra* note 96, at 102.
[73] *Id.* at 30.

3. Market freedom within the constraints of fairness

Earlier, we agreed with Dwork et al. that it is desirable to allow creators and users of AI systems as much freedom as possible consistent with the requirements of fairness "to benefit from investment in data mining and market research."[74] The FTC's fairness standard in 15 U.S. Code § 45(n) gives considerable weight to users' interest in freedom by balancing substantial injuries to consumers against the "countervailing benefits to consumers or to competition."

4. Reason-giving

Administrative agencies are well suited to reason-giving. The administrative law expert Jerry Mashaw notes that

> The path of American administrative law has been the path of the progressive submission of power to reason. The promise of the administrative state was to bring competence to politics. It is the institutional embodiment of the enlightenment project to substitute reason for the dark forces of culture, tradition, and myth. Administrators must not only give reasons; they must give complete ones. We attempt to ensure that they are authentic by demanding that they be both transparent and contemporaneous. "Expertise" is no longer a protective shield to be worn like a sacred vestment. It is a competence to be demonstrated by cogent reason-giving.[75]

One may take a less sanguine view of agency reason-giving. As the law professor Chris Hoofnagle notes in regard to agencies' use of their rule-making power, a "1979 assessment

---

[74] Dwork et al., *supra* note 1, at 214.

[75] JERRY L. MASHAW, REASONED ADMINISTRATION AND DEMOCRATIC LEGITIMACY: HOW ADMINISTRATIVE LAW SUPPORTS DEMOCRATIC GOVERNMENT 11 (2018) (citing Citizens to Preserve Overton Park, Inc. v. Volpe, 401 U.S. 402 (1971)).

of agency rule-making found that the average one created a record of over 40,000 pages."[76] Hoofnagle observes, that it "is popularly believed that if a general, online privacy rule-making were to be started, it simply would be stale by its implementation date."[77] Our point however is simply that, as Mashaw notes, "cogent reason-giving" is a hallmark of agency decision-making.

V. CONCLUSION

We conclude that regulation by the FTC (or an FTC-like) agency is a plausible way to ensure that AI systems meet the following four requirements. Where S is an AI system with a distance metric $d(x, y)$ and an allocation function $A(x)$:

> *Requirement 1*: S must generate $d(x, y)$ in ways consistent with "before start" equal opportunity.
>
> *Requirement 2*: It must be true that $|A(x) – A(y)| \leq d(x, y)$.
>
> *Requirement 3*: The allocation function $A(x)$ must not unfairly tilt the playing field.
>
> *Requirement 4*: A failure to provide adequate information in regard to (1) – (3) makes S presumptively unfair.

A dominant theme in the history of consumer surveillance has been consumer's *acceptance* of surveillance, or at least their acceptance of the products and services surveillance makes possible.[78] Fairness concerns have been a counterpoint, but so

---

[76] HOOFNAGLE, *supra* note 96, at 102. (quoting Hybrid Rulemaking Procedures of the Federal Trade Commission, ADMINISTRATIVE CONFERENCE OF THE UNITED STATES (1979), https://www.acus.gov/recommendation/hybrid-rulemaking-procedures-federal-trade-commission).

[77] *Id.*

[78] *See* BOUK, *supra* note 28 at 240 (remarking that "conflicts over traditional, predictive risk making spur the creation of new forms of risk

far, a largely ineffective one. Without effective regulation, the story is likely to continue in the same way.

---

making supposed to change fates, such that more risks get made from more people in new ways"). *See also* LAUER, *supra* note 25, at 215 (noting that the mid-1960s outrage soon died down). For insightful studies of consumers' acceptance of statistical analyses, *see* SARAH E. IGO, THE AVERAGED AMERICAN: SURVEYS, CITIZENS, AND THE MAKING OF A MASS PUBLIC (2007); MULLER, *supra* note 35. For studies focused on consumer acceptance in the twenty-first century, *see* ANDREAS BERNARD, THE TRIUMPH OF PROFILING: THE SELF IN DIGITAL CULTURE (Valentine A. Pakis tran., 2019); STEFFEN MAU, THE METRIC SOCIETY: ON THE QUANTIFICATION OF THE SOCIAL (2019).