When a Small Change Makes a Big Difference: Algorithmic Fairness Among Similar Individuals

Jane R. Bambauer,[†*] *Tal Zarsky,*[**] *and Jonathan Mayer*[***]

If a machine learning algorithm treats two people very differently because of a slight difference in their attributes, the result intuitively seems unfair. Indeed, an aversion to this sort of treatment has already begun to affect regulatory practices in employment and lending. But an explanation, or even a definition, of the problem has not yet emerged. This Article explores how these situations — when a Small Change Makes a Big Difference ("SCMBDs") — interact with various theories of algorithmic fairness related to accuracy, bias, strategic behavior, proportionality, and explainability. When SCMBDs are associated with an algorithm's inaccuracy, such as overfitted models, they should be removed (and routinely are). But outside those easy cases, when SCMBDs have, or seem to have, predictive validity, the ethics are more ambiguous. Various strands of fairness (like accuracy, equity, and proportionality) will pull in different directions. Thus, while SCMBDs should be detected and probed, they should not necessarily be removed.

 We thank David Abrams, Courtney Bowman, James Grimmelmann, Adam Kolber, Helen Nissenbaum, Ted Parsons, Andrew Selbst, Seana Shiffrin, Lior Strahilevitz, Mark Verstraete, Eugene Volokh, Felix Wu, and participants in the following discussions: the Cornell Tech Digital Life Initiative Fellows Workshop, the Cornell Tech Digital Life Initiative seminar, the Princeton University Center for Information Technology Policy seminar, the UCLA Law faculty workshop, and the Privacy Law Scholars Conference.

* Professor of Law, University of Arizona.

** Vice Dean and Professor of Law, University of Haifa – Faculty of Law and Visiting Scholar, University of Pennsylvania (2019-2020).

*** Assistant Professor of Computer Science and Public Affairs, Princeton University.

TABLE OF CONTENTS

INTRODUCTION

If an automated scoring or decision-making algorithm allows a small difference between two people to result in dramatically different treatment, has something gone wrong, morally or legally?

Consider two neighbors, Alex and Barbara, who are the same age, gender, race, ethnicity, and sexual orientation. Both practice the same religion. Both recently worked for the same employer, in the same type of job. But Alex moved into her home seven months ago, while Barbara moved in a month later. Suppose next, the police arrest Alex and Barbara for committing the same nonviolent crime on the same day. That small difference — just one month in move-in date — can make a big difference in outcome. A public algorithm might score Alex as a low-risk defendant, resulting in automatic release on her own recognizance, while Barbara is held in custody because the system has scored her as a medium-risk defendant. Can one say with confidence that an algorithmic process is unfair if it treats Alex and Barbara very differently from each other? Is it acceptable for a small change to make a big difference ("SCMBD")?

This scenario may sound fanciful, but the fact pattern is grounded in a real pretrial risk assessment tool and pretrial detention law.[1] Moreover, with the rapid adoption of statistics and machine learning in decision-

[1] The Ohio Risk Assessment System Pretrial Assessment Tool ("ORAS-PAT") uses a numerical scoring system for evaluating pretrial detention. EDWARD LATESSA, PAULA SMITH, RICHARD LEMKE, MATTHEW MAKARIOS & CHRISTOPHER LOWENKAMP, UNIV. OF CINCINNATI CTR. OF CRIM. JUST. RSCH., CREATION AND VALIDATION OF THE OHIO RISK ASSESSMENT SYSTEM: FINAL REPORT 49 (2009), https://www.ocjs.ohio.gov/ORAS_FinalReport.pdf [https://perma.cc/DF9D-WPE5]. Alex would receive 2 points for being unemployed, resulting in a low-risk (0–2 points) determination. Barbara would receive 2 points for being unemployed plus 1 point for moving in the past six months, resulting in a medium-risk (3–5 points) determination. A recent California law — which did not take effect because of a ballot referendum — would have generally required prearraignment release for low-risk defendants but permitted detention for medium-risk defendants. California Money Bail Reform Act of 2017, S.B. 10, 2017-2018 Leg., Reg. Sess. (Cal. 2018); *see Recent Legislation: Criminal Law — Bail Reform — California Replaces Money-Bail System with Pretrial Detainment System. — S.B. 10, 2017-2018 Leg., Reg. Sess. (Cal. 2018) (Enacted) (Codified at Cal. Gov't Code § 27771 and scattered sections of Cal. Penal Code)*, 132 HARV. L. REV. 2098, 2099-100 (2019) (summarizing the California bail reform law and describing the difference in treatment for low-risk and medium-risk defendants). As of 2017, seventeen California counties were using the ORAS-PAT for pretrial risk assessments. PRETRIAL DETENTION REFORM WORKGROUP, JUD. BRANCH OF CAL., PRETRIAL DETENTION REFORM: RECOMMENDATIONS TO THE CHIEF JUSTICE 101-02 (2017), https://www.courts.ca.gov/documents/PDRReport-20171023.pdf [https://perma.cc/4L9F-Q25X].

making systems,[2] these scenarios will become more common. Increasingly, a slight change in circumstances — with little foundation in factual, legal, or moral intuition — could result in dramatically different treatment of individuals by government agencies and private businesses.

Scholars in the budding field of algorithmic fairness, accountability, transparency, and ethics (commonly referred to as "FATE") recognize that an algorithm's output may be hypersensitive to changes in inputs, and that the phenomenon is endemic to machine learning.[3] But the literature almost exclusively treats this hypersensitivity as a diagnostic perspective for examining widely recognized types of unfairness, especially bias on the basis of protected characteristics, rather than as a potentially distinct type of unfairness.[4]

An explanation, or even a definition, of the SCMBD problem has not yet been fully elucidated. We aim to fill the gap. Our goal is primarily positive and taxonomical rather than prescriptive, though we do identify the outer boundaries of the normative case for circumstances under which firms and regulators should permit SCMBDs to be used. We also identify normative problems that run in the other direction when machine learning algorithms are designed to automatically detect and smooth out SCMBDs in their predictive functions. Sometimes those *corrections* will unwittingly introduce unfairness. To do all this, we explore how these SCMBD situations interact with various theories of algorithmic fairness related to accuracy, bias, strategic behavior, proportionality, and explainability.

Assessing SCMBD through the various and sometimes-competing notions of fairness simultaneously illuminates the nature of SCMBD and that of "fairness" itself. The inquiry takes us to the bleeding edge of the study of algorithmic decision-making in the fields of computer science and law, and at the same time, is rooted in concepts of equality dating back to Aristotle. Thus, the philosophical questions wrapped up in the debates about AI fairness are timeless.

---

[2] DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 53 (2020), https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf [https://perma.cc/9H5L-QHJM].

[3] There is substantial literature on explaining algorithm behavior and developing algorithms that are interpretable by human analysts, which we discuss in Part IV. For specific examples, see *infra* notes 12–13.

[4] *See, e.g.*, Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841, 853-54 (2018) (describing counterfactual explanations, a method for understanding an algorithm's behavior based on how it would treat similar individuals, as a possible means of identifying bias based on sensitive traits).

Nevertheless, they are also timely, as the stakes have new urgency. After all, the human experience is facing a tectonic shift in how we make decisions. Prior scholarship has ably characterized how advances in big data and artificial intelligence are transforming private- and public-sector decision making.[5] Briefly, and at the risk of overgeneralization, important decisions used to be guided by subject-matter experts and premised, ideally at least, on their experience and causal theories. That expertise is being supplemented — and in some contexts supplanted — by systems that make predictions based on large volumes of data and complex interactions within that data.[6] The shift to algorithmic prediction is transforming all areas of decision-making at once, enhancing the scale of its consequences (both beneficial and harmful). In the public sphere, algorithms are influencing decisions about arrests, criminal sentencing, the allocation of public benefits, and the administration of educational and medical services.[7] In the private sphere, algorithmic decision-making systems are increasingly common in finance, marketing, retail, and medicine, among many other fields.[8] In all of these realms, automated decision-making can potentially make marked improvements in efficiency without disturbing (and possibly even improving) equity and other fairness considerations. But fairness is an ambiguous concept. The scholarly literature is in a self-aware struggle to define what it means for an algorithm to be fair, or to at least to harmonize the many definitions floating around in the field.[9]

[5] ENGSTROM ET AL., *supra* note 2, at 20. *See generally* Kenneth A. Bamberger, *Technologies of Compliance: Risk and Regulation in a Digital Age*, 88 TEX. L. REV. 669 (2010) (discussing how the public sector mandates risk management in the form of compliance regulations and how the private sector complies through the use of data and artificial intelligence); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019) (discussing how the public sector, specifically the criminal justice system, uses data and artificial intelligence to estimate the likelihood that a person will commit future crime).

[6] *See* Dan L. Burk, *Algorithmic Legal Metrics*, 96 NOTRE DAME L. REV. 1147, 1149-50 (2021).

[7] Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 84-85 (2019) (discussing COMPAS and its faults); *see, e.g.*, Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1160-67 (2017) (discussing various ways how decision-making systems are currently being used in the public sector and future efforts to create large volumes of data to support agency functions); Aziz Z. Huq, *Constitutional Rights in the Machine Learning State*, 105 CORNELL L. REV. 1875, 1878, nn.8–14 (2020) (providing examples of how algorithms are being used in the public sphere, such as in immigration cases and in bail/sentencing contexts).

[8] *See* W. Nicholson Price II & Arti K. Rai, *Clearing Opacity Through Machine Learning*, 106 IOWA L. REV. 775, 775 (2021).

[9] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI*, 41 COMPUT. L.

The legal literature on algorithmic fairness has, so far, gravitated toward two foci: the risks of opaque systems and the risks of hidden biases.[10] Scholars who focus on the former have called for establishing (and possibly even mandating) "explainable" algorithmic processes[11] or to "break open the black box"[12] of automated decisions. Scholars focused on the latter have helped reinvigorate the discourse on antidiscrimination theory and policy.[13]

Our project is fundamentally distinct from these two strands of scholarship. The problem that we study does not result from inadequate transparency (though lack of transparency may exacerbate the problem, and transparency may be a constructive response). Our focus also does not relate to discrimination on the basis of protected characteristics, where an algorithmic decision-making system by intent or effect exhibits impermissible biases (though the problem that we study may, in some instances, be correlated with those types of biases). To see why, we need not search beyond the pre-arraignment detention example we just presented: the algorithm is public, and the aspect of the algorithm that causes concern does not involve discrimination related to sensitive traits. Nevertheless, our collective discomfort with the example suggests that discussions of algorithmic fairness may be missing something.

& SEC. REV., July 2021, at 2-3; Arvind Narayanan, *Tutorial: 21 Fairness Definitions and Their Politics*, YOUTUBE (Mar. 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk [https://perma.cc/X7ZM-R9R5] (indicating the diversity of "fairness" definitions).

[10] See JULIE E. COHEN, BETWEEN TRUTH AND POWER 247 (2019), for a discussion of both the issue of discrimination and opacity as the key problems at this juncture.

[11] Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1280 (2017); Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 209 (2019).

[12] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 142 (2015) (setting forth transparency solutions to black box problems). For one of the many popular news articles using a variation of this theme, see Jason Bloomberg, *Don't Trust Artificial Intelligence? Time To Open the AI 'Black Box'*, FORBES (Sept. 16, 2018, 1:26 PM EDT), https://www.forbes.com/sites/jasonbloomberg/2018/09/16/dont-trust-artificial-intelligence-time-to-open-the-ai-black-box/#29a5bbd33b4a [https://perma.cc/HNG6-SC7H].

[13] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 681 (2016) [hereinafter *Big Data's Disparate Impact*]. For a critical discussion of the ability to correct algorithmic processes, see Mayson, *supra* note 5, at 2262-72. For a recent review of the legal and policy literature on these matters, see ENGSTROM ET AL., *supra* note 2, at 79-82. For a survey of the literature in the computer science context, see Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Krisitina Lerman & Aram Galstyan, *A Survey on Bias and Fairness in Machine Learning*, 54 ACM COMPUTING SURVS., July 2021, at 1-2, https://dl.acm.org/doi/pdf/10.1145/3457607 [https://perma.cc/24FS-46PM].

The SCMBD strand of potential unfairness has previously been identified in the technical literature on fair algorithms. In a seminal contribution to FATE scholarship, Cynthia Dwork et al. offered a formal mathematical definition of "individual fairness" that requires similar (probabilistic) treatment for similar individuals in a machine learning classifier.[14] But literature on individual fairness has been overshadowed by "group fairness" scholarship, which emphasizes measuring and mitigating discrimination and biases associated with sensitive characteristics such as race and gender.[15]

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel, *Fairness Through Awareness*, *in* PROCEEDINGS OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE 214, 215 (2012), https://dl.acm.org/doi/pdf/10.1145/2090236.2090255?casa_token=N8c3YnXkQjUAAAAA:fnCeHk8VlB0mz8tYcobUrd9e2X_ss92zMJ0dZ0zI7k2uwGStwrKnjYGXbd1_-NChUFh6dbXsW3z5dmw [https://perma.cc/5JTZ-CNR5]. In the Dwork et al. definition, a randomized machine learning classifier satisfies "individual fairness" if individuals with similar inputs (compared with a distance metric) have similar probabilistic outputs. Dwork et al. contrast their definition with "group fairness," such as statistical parity between minority and non-minority groups, and demonstrate ways in which individual fairness and group fairness can interact. While the goals of the paper are quite different from our own — exploring a new technical approach for constructing probabilistic machine learning classifiers — the focus on fairness among similarly situated individuals is analogous.

[15] Subsequent publications have proposed other formal definitions of "individual fairness," which involve measures at the individual rather than group level. Some of these definitions are also conceptually analogous to the SCMBD problems that we examine. *E.g.*, MICHAEL KEARNS, AARON ROTH & SAEED SHARIFI-MALVAJERDI, AVERAGE INDIVIDUAL FAIRNESS: ALGORITHMS, GENERALIZATION AND EXPERIMENTS 4 (2019) https://arxiv.org/pdf/1905.10607.pdf [https://perma.cc/VT4V-NH9F] (proposing a definition of individual fairness that probabilistically bounds differences in classification errors among individuals subject to repeat classification); Preethi Lahoti, Krishna P. Gummadi & Gerhard Weikum, *Operationalizing Individual Fairness with Pairwise Fair Representations*, 13 VLDB ENDOWMENT 506, 507-09 (2019) (proposing an approach to individual fairness that relies on pairwise comparisons rather than a distance metric); Richard Zemel, Yu Wu, Kevin Swersky & Toniann Pitassi, *Learning Fair Representations*, 28 ICML 325, 326 (2013) (proposing a "consistency" metric that compares classification of an individual to the classification of the nearest neighbors of the individual). Other technical contributions on individual fairness are more removed from SCMBD problems. *E.g.*, MATTHEW JOSEPH, MICHAEL KEARNS, JAMIE MORGENSTERN & AARON ROTH, FAIRNESS IN LEARNING: CLASSIC AND CONTEXTUAL BANDITS 3 (2016), https://arxiv.org/pdf/1605.07139.pdf [https://perma.cc/UNG8-8MLM] (proposing a definition of fairness for multi-armed bandit problems where arms are individuals from different groups and with high probability the algorithm always selects the arm with the highest expected payoff, grounded in the perspective that "it is unfair to preferentially choose one individual . . . over another if he or she is not as qualified as the other individual"); Asia J. Biega, Krishna P. Gummadi & Gerhard Weikum, *Equity of Attention: Amortizing Individual Fairness in Rankings*, 41 ACM SIGIR 405, 405-06 (2018) (proposing new measures and methods for fairness in recommendation ranking that are grounded in individualized attention and relevance metrics).

Analogizing our inquiry to constitutional concepts helps bring the project into relief. Prior work on algorithmic transparency is, in many respects, connected to notions of procedural due process: individuals subject to decision-making by an algorithm deserve notice and an opportunity to be heard.[16] Scholarship on algorithmic bias, meanwhile, has deep parallels to traditional concepts of equal protection: decisions by algorithm should be free of discriminatory intent or disparate impact.[17] Our focus, by contrast, is more analogous to substantive due process or equal protection review of policies that distinguish among individuals on a potentially arbitrary basis.[18] When an algorithm has a major effect on a person's life, our concern is that similar individuals may receive dissimilar treatment. This perspective is comparable to equal protection review involving a fundamental right, where courts apply heightened scrutiny and question policies that distinguish among similarly situated individuals, whether they are in a protected class or not.[19] Where an algorithmic decision-making system does not result in a highly significant consequence, our concern has less force. Nevertheless, even outside fundamental rights, treatment cannot be arbitrary or irrational — concepts that are, at times, well-matched to SCMBDs.[20]

As this Article will show, a formal exploration of SCMBD and proportionality is no simple task. After all, while proportional treatment is, intuitively, a hallmark of fairness, advances in AI will force us to question how we know if small differences between individuals really are "small" and why proportionality should be so highly valued in the first place — especially when achieved at the expense of accuracy or consistency. Ultimately, we find that fairness is best conceived as the final product of a careful assessment and weighing of various social goals and priorities. SCMBDs, like inaccuracy, bias, opaqueness, and other strands of unfairness, are only "bad" or "good" depending on how they fit in the balance of social goals. Another way to put this is that

---

[16] *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1252-54 (2008); Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 27-28 (2014); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 109 (2014).

[17] *See* Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 13, at 681.

[18] For a discussion that disentangles this form of rational basis review from other constitutional forms of scrutiny, see Jane R. Bambauer & Toni M. Massaro, *Outrageous and Irrational*, 100 MINN. L. REV. 281, 309-17 (2015).

[19] *See* City of Cleburne v. Cleburne Living Ctr., Inc., 473 U.S. 432, 440 (1985) (describing fundamental rights that initiate elevated scrutiny).

[20] *See, e.g.*, Williamson v. Lee Optical Inc., 348 U.S. 483, 487-88 (1955) (stating the rule that legislation needs to be a rational way to address a problem).

demands for explainable AI or for parity across demographic groups are easier to justify and to respond to when they dovetail with arguments against SCMBD. When this occurs, there are multiple, stacking reasons to allege an algorithm is unfair.

We proceed as follows: in Part I, we mark out the scope of the project. First, we offer motivating examples of SCMBD in the real world. We then attempt to clearly define SCMBD instances and illuminate the difficulties their analysis raises. The meaning of "small" and "big" differences in inputs or outputs will inevitably be context-dependent and contested at the margins, but there will often be consensus around the central meaning. At the very least, some automated techniques can be used to identify possible SCMBD candidates that a human team can further review to decide whether they fit the definition or not. Part I also describes how regulatory agencies have already begun to discourage SCMBDs, albeit without explicit recognition of the problem. This proves that SCMBDs are intuitively suspicious and deserve a theoretical frame to help classify which SCMBDs are undesirable and why. Finally, we explain how machine learning systems create new risks of SCMBD phenomena.

Part II explains why SCMBD dynamics are, or might be, bad. At times, they may be unfair because they are (a) inaccurate, possibly the product of overfitting, spurious correlations, or flawed training data; (b) biased against protected or marginalized groups, possibly due to reliance on tainted training data, poorly chosen objective functions, or other algorithm development shortcomings; (c) game-able, leading to strategic behavior and noise in a dynamic system; and (d) inherently unfair, despite accuracy, based on foundational concepts of moral philosophy requiring proportionality or parsimony.

Part III explains why SCMBD dynamics are, or might be, *not* bad. In some cases, SCMBD outcomes can enhance most other forms of fairness. Small changes may cause real differences in outcomes due to natural phenomena (consider phase transitions) or social phenomena (consider network effects and tipping points). There is a reason that men who are 5' 11" tall report that they are 6' 0" on their dating profiles,[21] and there is also a reason that a home mortgage applicant with a FICO credit score of 630 will get much better terms than an applicant with a score of 610.[22] Even if SCMBDs are inexplicable when

[21] *The Big Lies People Tell in Online Dating*, OKCUPID BLOG (July 7, 2010), https://theblog.okcupid.com/the-big-lies-people-tell-in-online-dating-a9e3990d6ae2 [https://perma.cc/82YW-DKLX].

[22] Fannie Mae only purchases mortgages on the secondary market for mortgage loans with borrowers who have a credit score of at least 620. *Selling Guide: B3-5.1-01, General*

first discovered, further investigation may lead to new insights and understanding. This Part also compares SCMBDs that occur in decision-making algorithms to SCMBDs that are the products of legal rules. The two are surprisingly dissimilar. Because SCMBDs that emerge from machine learning are latent and often improve the accurate categorization of people, they mostly have opposite qualities to SCMBDs in the law — which have the virtue of clear notice but the drawback of inaccuracy at the margins. Thus, deciding what to do about SCMBDs require us to grapple with foundational jurisprudential questions and principles rather than importing insights from the literature on bumpy laws (and the lumpiness they feature) or rules versus standards.

Part IV concludes with policy recommendations. A policy window is beginning to open as public consensus forms around distrust of automated decision-making.[23] Taking cues from legal scholarship, policymakers have introduced a range of regulatory[24] and legislative efforts[25] to constrain the use of opaque or unfair algorithms. We recommend that when an organization deploys an algorithmic decision-making system of consequence, it include audits for SCMBD as part of a program to ensure ethical and responsible operation of the system. We describe several possible approaches for identifying SCMBD phenomena in a system. Once found, a SCMBD can then be further tested for impact on accuracy and disparities across demographic

Requirements for Credit Scores, FANNIE MAE (Dec. 15, 2021), https://selling-guide.fanniemae.com/Selling-Guide/Origination-thru-Closing/Subpart-B3-Underwriting-Borrowers/Chapter-B3-5-Credit-Assessment/Section-B3-5-1-Credit-Scores/1032996841/B3-5-1-01-General-Requirements-for-Credit-Scores-08-05-2020.htm [https://perma.cc/KN97-RTXG].

[23] Aaron Smith, *Public Attitudes Toward Computer Algorithms*, PEW RSCH. CTR. (Nov. 16, 2018), https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/ [https://perma.cc/95QP-AUXC]. Journalism and popular nonfiction books have helped a rapid transmission of academic concern into the public sphere. *See generally* CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016) (describing how automated processes can reproduce both blatant and latent biases from the past); Shoshana Zuboff, *You Are the Object of a Secret Extraction Operation*, N.Y. TIMES (Nov. 12, 2021) (describing how algorithms used by online advertisers are "weaponized" against "unsuspecting" Internet users), https://www.nytimes.com/2021/11/12/opinion/facebook-privacy.html [https://perma.cc/3LLG-8L8U].

[24] For some early regulatory steps in New Zealand, see Charlotte Graham-McLay, *New Zealand Claims World First in Setting Standards for Government Use of Algorithms*, GUARDIAN (July 27, 2020, 2:00 PM EDT), https://www.theguardian.com/world/2020/jul/28/new-zealand-claims-world-first-in-setting-standards-for-government-use-of-algorithms [https://perma.cc/R6PG-FBS7].

[25] The Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (1st Sess. 2019).

groups, and the results of all of these tests can together help inform a decision about whether to adjust a model.

That's when the hard work really begins. When machine learning deprives decision-makers of the social balm of impossibility, they will have to make (human) judgment calls between competing values and incompatible priorities. While there may be some wrong answers, there are no definitively right answers. Thus, decisions about how to handle a SCMBD will never be entirely free from criticism.

I. WHAT IS SCMBD?

This Part sets the scope and motivation for the entire project. We start with two additional realistic examples of SCMBD, to further develop intuition about the problem. Each example has two versions, first as an instance of bias on the basis of a protected characteristic, and second as an instance of SCMBD. We then define our terms ("small change" and "big difference") with greater precision and provide examples of regulatory interventions that are, at least implicitly, driven by a concern of SCMBD. Finally, we discuss why SCMBD will likely be a recurring issue for machine learning systems.

A. Motivating Examples

In August 2019, Apple and Goldman Sachs introduced a new credit card (the Apple Card) that provides perks for cardholders who regularly purchase Apple products or use Apple Pay.[26] Later that year, an affluent couple using the card discovered that the husband's credit limit was twenty times higher than his wife's.[27] Her application for a credit increase was denied.[28] This outcome was surprising as the couple lived at the same address, the wife had a higher credit score and they filed a joint tax return.[29] The husband, who happened to be a prominent software developer, decided to vent to his 350,000 followers on Twitter.[30] Soon, other customers chimed in, including Steve Wozniak

[26] John S. Kiernan, *Apple Credit Card Review*, WALLETHUB (Nov. 24, 2021), https://wallethub.com/edu/cc/apple-credit-card-review/25979/ [https://perma.cc/3QYH-D8QE].

[27] Neil Vigdor, *Apple Card Investigated After Gender Discrimination Complaints*, N.Y. TIMES (Nov. 10, 2019), https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html [https://perma.cc/CWD3-9L5A].

[28] *Id.*

[29] *Id.*

[30] *Id.*; Reuters, *Apple Card Issuer Investigated After Claims of Sexist Credit Checks*, GUARDIAN (Nov. 9, 2019, 10:01 PM EST), https://www.theguardian.com/technology/2019/

(one of Apple's founders), reporting similar experiences.[31] The New York State Department of Financial Services began an investigation into the Apple Card and, in a turn of events that will surprise precisely no one, the wife's credit limit was promptly increased.[32] The investigation did not find any evidence of deliberate discrimination or disparate impact (but did find "deficiencies in customer service and transparency," which were subsequently corrected).[33]

This example raises a suspicion of gender or sex discrimination, and on those terms, we would not classify it as a SCMBD. A distinction on the basis of gender is normatively troubling (not to mention illegal), but most would agree that differences in gender are not "small." Gender carries sociological, cultural, and behavioral differences that make it a substantial and meaningful factor in many contexts. For all these reasons, the Apple Card controversy as a gender discrimination story would not trigger our attention for SCMBD. It is prudent at this point to remind readers that we do not intend SCMBD to be the only, or even the primary, means of assessing algorithmic fairness. Gender disparities are unfair for their own reasons.

Apple and Goldman Sachs reported (and the state investigators agreed) that they did *not* rely on gender in setting credit levels, nor were they aware of a gender bias in outcomes.[34] They also did not use any obvious proxy for gender when making the credit limit determinations (though of course what it means to be a "proxy" for gender or race is difficult to pin down in a machine learning system that learns subtle correlations from a large volume of data).[35] But consider this possible explanation: suppose that a slight difference between the spouses' spending on the card was the factor that drove the large difference in credit limits for the husband and wife. These slight differences were not

nov/10/apple-card-issuer-investigated-after-claims-of-sexist-credit-checks [https://perma.cc/FL4G-33PL].

31 Reuters, *supra* note 30.

32 Will Knight, *The Apple Card Didn't 'See' Gender — and That's the Problem*, WIRED (Nov. 19, 2019, 9:15 AM), https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/ [https://perma.cc/Z656-RRRB]; Vigdor, *supra* note 27.

33 N.Y. STATE DEP'T OF FIN. SERVS., REPORT ON APPLE CARD INVESTIGATION 1, 2 (2021), https://www.dfs.ny.gov/system/files/documents/2021/03/rpt_202103_apple_card_investigation.pdf/ [https://perma.cc/L5U5-KTJV]. The report further found, after extensive review, that men and women with "equivalent credit characteristics" had similar outcomes. *Id.* at 5.

34 *Id.* at 6 ("The Department did not find, for example, any policy providing for lower credit limits for women or evidence suggesting the Bank intentionally judged women and men by different standards.").

35 Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 13, at 681.

reported by the husband on Twitter because the details were forgotten, or perhaps were never even known to the couple. Indeed, there is some speculation in the commentary following the outcry that the reason for the disparity between the spouses' credit limits in this case was caused by a difference in the individual levels of spending using the card.[36] Generally speaking, higher spending leads to greater credit limits. If this was the only difference between the spouses — that the husband had spent somewhat more money than his wife while using the card — would the couple lose their moral claim to outrage? Similarly, the New York State investigation noted as possible reasons for the gender disparity the fact that only one of the spouses was named on their residential mortgage or that one spouse held multiple credits cards while the other spouse had merely a single card.[37] Is the outcome justifiable based on these differences?

If a small (but nontrivial) difference in spending or the number of credit cards issued to an individual causes a twenty-fold increase in credit limits, a SCMBD may be present. Of course, when minor differences tend to track gender, looking for SCMBD situations could help antidiscrimination goals, too, by providing a means to find factors that have excessive impact on a gender gap. This might be especially true if the reason for the difference is the naming of only one spouse on the residential mortgage which, by patriarchal tradition, is typically the male. But to understand the independent role of SCMBD in assessing algorithm fairness, let's make a tweak to the facts. Suppose a same-sex couple experienced a similar dynamic — a substantial credit limit disparity between the spouses caused by differences in spending patterns.[38] Removing gender from the fact pattern allows us to consider whether small differences between the parties that cause a twenty-fold difference in available credit would generate a normative concern *on its own terms*, without tapping into the form of unfairness that stems from group-based disparate impacts. We get a clean focus on the propriety of SCMBDs.

Even without gender bias, the SCMBD prompts an intuitive and visceral response. Most have an aversion to the breach of expectations

---

[36] This position was mentioned by some credit card experts. *See* Diane Harris, *Apple Card Gender Bias? Don't Assume Its Discrimination, Experts Warn*, NEWSWEEK (Nov. 12, 2019, 6:18 AM EST), https://www.newsweek.com/apple-card-gender-bias-credit-limit-goldman-sachs-1471146 [perma.cc/NQ44-YQ5U].

[37] N.Y. STATE DEP'T OF FIN. SERV., *supra* note 33, at 10-12.

[38] Anecdotally, looking at the press addressing this issue, it appears that many other couples indicated they had such a similar experience, but we did not find any such reports regarding same-sex couples, or that the dynamic was reversed with mixed-sex couples (meaning that it was the husband and not the wife that received the lower credit limit), all of which makes the issue more suspicious of including a gender bias.

of proportionality. SCMBDs seem to demand an explanation. This naturally opens questions about the sufficiency of explanations. Suppose a higher spending spouse's purchasing patterns place them in the top three percent of all cardholders and thus, makes them automatically eligible for VIP credit lines. Is this explanation enough? There are at least two possible responses. One is to accept steep cliffs between categories like this so long as this strategy fits the standard business practices of the organization. The other option, which we think is more desirable, is to use the SCMBD to probe whether crude categories (such as a high spending category) actually serve a business purpose, particularly when more granular treatment may be inexpensive for a company. If categories are just a holdover from the era when computation was costly, the identification of a SCMBD should help guide the company to revise their policies not only for the good of their baffled consumers, but for their own bottom line as well. This observation wouldn't necessarily require a company to smooth out *all* categorical distinctions completely; rather, it might highlight that the firm should add additional, intermediate pricing categories if for no other reason, to ensure customers will not feel they are being treated in an arbitrary way.

Let's move on to a second real life algorithmic SCMBD example. Anecdotal evidence suggests that e-commerce websites have occasionally recommended different products to Mac and PC users (a website can easily learn a user's operating system from their web browser).[39] To be clear, this is not a case of price discrimination (though we could imagine the same information could be valuable for that as well); rather, the selection of products shown to Mac users tended to be pricier options.[40]

Intuitively, it is not surprising that users of these platforms differ in their personality traits,[41] and thus in their shopping habits and preferences. Apple had a long-running ad campaign personifying the Mac as the hip metropolitan yuppy in contrast to the fuddy-duddy, suburban dad bod PC. Still, the operating system doesn't have any obvious direct

[39] *See, e.g.*, Dana Mattioli, *On Orbitz, Mac Users Steered to Pricier Hotels*, WALL ST. J., https://www.wsj.com/articles/SB10001424052702304458604577488822667325882 (Aug. 23, 2012, 6:07 PM EST) [perma.cc/XTA5-G7LM] (describing an apparent algorithm design that promoted higher priced hotels to Internet users on Apple devices).

[40] *Id.*

[41] Ana Sandoiu, *Do Android and iPhone Users Have Different Personalities?*, MED. NEWS TODAY (Nov. 27, 2016), https://www.medicalnewstoday.com/articles/314376 [perma.cc/FYK9-MJSS].

or necessary connection to personality, so its use is still surprising. The scattered reports of differences in the offers displayed for Mac and PC users prompted researchers to examine the efficacy of using the Mac/PC distinction (as well as its close cousin, the iOS/Android divide) in credit allocation.[42] They found that operating systems were correlated with creditworthiness, and combining similar "digital footprint" features could be highly predictive of credit scores.[43] But the Mac/PC distinction also correlates with class and race differences. Mac users are more likely to be high income and white than PC users.[44] Based on these findings, a recent Brookings report openly questions whether credit denial using this parameter should be prohibited under existing antidiscrimination law.[45] And yet, a lot of things that go into credit scores — both the traditional FICO-style scores and the new alternative scores developed in the Fintech field — have racially disparate impact.[46] Income and employment status differ by race but banning their use because of incidental impacts would dramatically alter credit scoring and the entire credit market. We suspect disparate impact alone is not what makes the Mac/PC distinction so troubling; it is the disparate impact combined with its seeming senselessness (even though the algorithm indicates predictive power).

If we consider this example through the SCMBD prism, we might reach a better diagnosis of the problem. Let's consider two similar applicants who differ only by the operating systems/hardware they use. Setting aside the racial proxy matter for just a moment, this binary difference could be considered "small." This categorization is debatable, but it seems to us that the Mac/PC distinction, while real, should mostly

[42] For a study indicating the limited differences between these groups, see Friedrich M. Götz, Stefan Stieger & Ulf-Dietrich Reips, *Users of the Main Smartphone Operating Systems (iOS, Android) Differ Only Little in Personality*, 12 PLOS ONE, May 3, 2017, at 9.

[43] Tobias Berg, Valentin Burg, Ana Gombovi & Manju Puri, *On the Rise of FinTechs — Credit Scoring Using Digital Footprints* 3-4 (Nat'l Bureau of Econ. Rsch., Working Paper No. w24551, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3170770# [https://perma.cc/BM9C-7HWE] (discussing how the difference in default rates between customers using iOS (Apple) and Android (e.g., Samsung) is equivalent to the difference in default rates between a median credit score and the eightieth percentile of the credit bureau score).

[44] *See* Aaron Klein, *Reducing Bias in AI-based Financial Services*, BROOKINGS (July 10, 2020), https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/ [perma.cc/2B47-5R48] [hereinafter *Reducing Bias in AI-based Financial Services*] ("Think about the potential to use whether or not a person uses a Mac or PC, a factor that is both correlated to race . . . .").

[45] Aaron Klein, *Credit Denial in the Age of AI*, BROOKINGS (Apr. 11, 2019), https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/ [perma.cc/L6QT-ARF3].

[46] Klein, *Reducing Bias in AI-based Financial Services*, *supra* note 44.

capture aesthetic differences and not differences in responsibility, particularly when other factors (like income) are already well-controlled. Without a better theory to explain the behavioral differences in Mac and PC users, one could reasonably regard this factor as suspect if too much weight is placed on it.

We can contrast the Mac/PC SCMBD with other examples of odd binary inputs that have outsize effects on output, but that may not be "small." For example, consumers who purchase floor-protecting pads for the legs of their furniture are scored as significantly more creditworthy than otherwise similar consumers who do not,[47] and auto insurance companies may charge higher rates to smokers than to otherwise similar non-smokers.[48] In both cases, the inputs (furniture pads, smoking) have no direct connection to the predicted outcome (paying back a large loan or driving well), and yet their *indirect* connection is easy enough to articulate. The decision to buy furniture pads or to not smoke demonstrates foresight and caution — qualities that are hard to observe directly. These present edge cases for SCMBDs — examples where reasonable minds may differ as to whether to categorize the changes as "small" or not. The decision to purchase a Mac, by contrast, has no clear connection of even an indirect sort to creditworthiness (at least when controlling for financial factors like income).

Thus, we would conclude that the Mac/PC distinction is a small change. Denial of credit or a significant difference in the terms of a loan is a big difference in outcome, so a SCMBD is found. The fact that this binary variable is correlated with class and race and that its use would most likely exacerbate racial gaps adds additional weight to the arguments against its continued use. But even if there weren't a risk of disparate impact, this SCMBD can be a signal of unfairness for independent reasons, as we explain in Part II.

---

[47] It is unclear whether this often-cited example is more than an urban legend. For discussion of this dynamic, see Charles Duhigg, *What Does Your Credit-Card Company Know About You?*, N.Y. TIMES (May 12, 2009), https://www.nytimes.com/2009/05/17/magazine/17credit-t.html [perma.cc/9ULK-LX6H]; Frank Pasquale, *Scores of Scores: How Companies Are Reducing Consumers to Single Numbers*, ATLANTIC (Oct. 14, 2015), https://www.theatlantic.com/business/archive/2015/10/credit-scores/410350/ [perma.cc/6TN3-2QXJ].

[48] *See* Jeffrey J. Sacks & David E. Nelson, *Smoking and Injuries: An Overview*, 23 PREVENTIVE MED. 515, 516 (1994) ("Compared with nonsmokers, smokers may have a 50% increased risk for MVCs . . . .").

B. From Intuition to Definition

The examples above intuitively track instances where similar people are treated too differently.[49] But, as with many intuitions, it is more difficult to rigorously define what "SCMBD" means than to recognize it. In this sense, it shares something with Justice Stewart's "I know it when I see it" First Amendment test for hardcore pornography.[50] In this Section, we offer definitions for a "small change" in inputs or a "big difference" in outputs. And we concede that a healthy level of subjectivity is involved.

1. Small Change

Let's start with the "small change" in inputs. For ease of discussion, we will assume that data used as inputs take the form of continuous or binary variables.[51] Continuous variables exist on a sliding scale that can theoretically take an infinite number of possible values if measurement were infinitely precise.[52] Height, BMI, age, and average pulse under stress are all continuous variables. Total credit card spending by Apple Card users would be a continuous variable, too. Binary inputs, by contrast, have only two possible values: 1 or 0.[53] They are either fulfilled or are not fulfilled for each data subject. A subject can be described as either a smoker or non-smoker, pregnant or not, with or without a college degree, or a Mac-user/PC (non-Mac) user.[54] We are leaving out categorical variables that take multiple discrete values in this taxonomy because they are typically encoded as binary or continuous features for machine learning systems, and the same SCMBD analysis applies.[55]

---

[49] These are of course only "second-best" examples. The best example would be showing how the same individual with only slightly different attributes is treated very differently in alternative universes. Yet, that "pure" example is currently unrealistic.

[50] Jacobellis v. Ohio, 378 U.S. 184, 197 (1964) (Stewart, J., concurring).

[51] This distinction is often addressed in this context, for instance, see David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC DAVIS L. REV. 653, 674-75 (2017).

[52] *Id.* at 673.

[53] *Id.*

[54] We will assume that categorical variables that are not ordinal will be converted to a set of binary variables, as is customary for popular machine learning methods such as neural networks. For a discussion comparing discrete variables to binary ones, see Adam J. Kolber, *Smooth and Bumpy Laws*, 102 CALIF. L. REV. 655, 660 (2014).

[55] If a categorical feature lacks ordinal properties (e.g., type of animal or industry), it is usually encoded as a binary feature for a machine learning system. If a categorical feature has ordinal properties (e.g., test scores, age in years, or years of education), it is usually encoded as a continuous feature.

The "change" is the difference between two individuals' values for a particular variable, of course, but when is that change "small"? Ideally, we would have an objective, measurable, and predictable threshold for changes that are sufficiently small. Continuous variables hold more promise for a quantitative definition of "small change" than binary variables, so let's start with them.

One natural candidate is to look at the absolute differences in values — for example, anything less than a $5,000 difference in annual household income is small. Another natural candidate is to look at relative changes so that "small" could be defined within a certain range of ratios (between 90%–110%, for instance). These are the most straightforward measures of "change," but the threshold for "small" will be subject to debate. More importantly, absolute and relative comparisons often won't capture what is really most important, which is how similar two people are compared to the overall distribution of the population. With normal distributions (the classic bell curve, that is), the same difference in absolute terms could have a very different significance depending on whether the change occurs near the mean or around the tails.[56] Thus, it may be better to analyze the size of a difference in terms of the distribution of the underlying population, such as by examining percentiles.

Take, for example, household income.[57] Is the difference between American households with $20,000 and $30,000 in income "greater" than the difference between households with $200,000 and $300,000? Even though the ratio (150%) would be the same, intuition says yes, and so does the distribution. After all, shifting from $20,000 to $30,000 in household income means moving from the 14th percentile to the 23rd, but moving from $200,000 and $300,000 would shift just five percentile points, from 91st to 96th.[58] Far enough out at the limits of the value range, percentiles might lose their appeal (e.g., we would expect there is a big difference between households earning $1 million

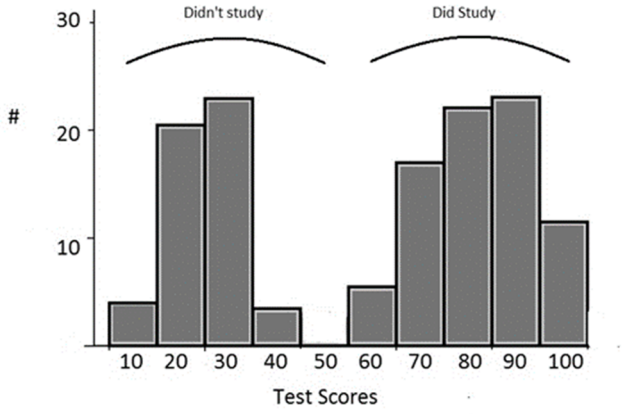
[56] *See* Solon Barocas, Andrew D. Selbst & Manish Raghavan, *The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons*, FAT* '20: PROC. OF THE 2020 CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, Jan. 2020, at 84, https://dl.acm.org/doi/pdf/10.1145/3351095.3372830 [https://perma.cc/ZHR9-966R] (explaining the need to "normalize" data when striving to compare inputs and outputs).

[57] Household income is not actually normally distributed; it has a thick long tail. The natural log of household income is closer to normally distributed, and thus, Ln(income) is often used in statistical analyses. For ease of discussion, we use dollars rather than log-dollars in this example.

[58] *See* PK, *Household Income Percentile Calculator for the United States in 2019*, DQYDJ, https://dqydj.com/2019-household-income-percentile-calculator/ (last visited Feb. 15, 2022) [https://perma.cc/99W4-GED9].

per year and $100 million per year even though they are both in the top 1%), but outside the extremes, percentiles arguably capture the differences between people in terms that are compatible with how we assess each other.

However, the *shape* of the distribution matters a good deal, too. In bimodal or multimodal distributions,[59] the definition of a "small" change should take into account the appropriate cluster. For example, suppose that the outcome of a spelling test looks like this:[60]



In this example, a person who received a grade of 47 may very well be more similar to a person who received a 29 than a person who earned a 61 even if they look more similar to the 61 in absolute, relative, and percentile terms.[61] Thus, the definition of small should have sensitivity to the shape of the underlying distribution. The important point is that the definition of a "small change" lends itself to quantification for a continuous input variable, and it would not be difficult to write an auditing program to identify potential candidates.

Matters get complicated analytically (although simpler mathematically) for binary variables because the change in value is *always* 1. Does this mean that binary variables always satisfy the "small

[59] A distribution with at least two distinct modes (or peaks). *See Multimodal Distribution*, WIKIPEDIA, https://en.wikipedia.org/wiki/Multimodal_distribution (last updated Sept. 10, 2021, at 9:21 PM UTC) [perma.cc/5GQ9-AG5P].

[60] Stephanie Glen, *Bimodal Distribution: What Is It?*, STAT. HOW TO, https://www.statisticshowto.com/what-is-a-bimodal-distribution/ (last visited Dec. 28, 2021) [perma.cc/5H3S-GE52].

[61] Alternatively, the difference between scoring forty-seven and sixty-one can be treated as "small" for the purpose of identifying a SCMBD, and later in an assessment process, the SCMBD can be deemed explainable and valid. See *infra* Part III, for a full articulation of legitimate or beneficial SCMBDs.

change" requirement of SCMBD or never do? Neither option is satisfying. On one hand, it is quite common for a response to a single binary question to change the outcome of an algorithm for very good reason. For instance, travel insurance premiums jump if a person provides an affirmative answer when asked whether they plan to engage in scuba diving or other extreme sports.[62] Whether someone has filed for bankruptcy in the recent past is another example of a single binary factor that can appropriately receive large weight on a credit score or lending decisions. And yet, some binary responses seem borderline-trivial and beg for inclusion in our scope. For instance, whether an individual has or has not used ALL CAPS in her online application or uses a Mac seem to be "small" changes that should not substantially affect one's credit standing.[63] Unlike continuous variables, the distribution of the population into each of the two values (0 or 1) does not help much, either. If only a tiny proportion of people fall into the 0 or 1 category, it might suggest that the factor is a highly unusual and relevant quality (e.g., going SCUBA diving), but it might not (e.g., using ALL CAPS).

We tentatively recommend treating binary variables as if they automatically satisfy the "small change." The binary variables that cause a "big difference" will have to be further analyzed by humans to determine whether the candidate variable is a SCMBD. This means including or excluding a particular binary variable in the meaning of "SCMBD" will inevitably capture some value judgments or policy preferences. So be it. If SCMBD is worth scrutinizing at all, it is worth scrutinizing binary variables, too, for at least two reasons. First, excluding binary variables takes too much off the table. Big data techniques often use sparse data that is coded as a huge collection of binary variables containing, for example, a variable for each individual item that a consumer might purchase.[64] These huge, sparse sets of

[62] *See, e.g.*, *Travel Insurance*, CTRS. FOR DISEASE CONTROL & PREVENTION: TRAVELERS' HEALTH, https://www.cdc.gov/travel/page/insurance (last reviewed June 22, 2021) [perma.cc/MC7H-KG99] (noting that when traveling out of country and performing adventure activities such as scuba diving or hang gliding, travel insurance is important).

[63] *See* Steve Lohr, *Banking Start-Ups Adopt New Tools for Lending*, N.Y. TIMES (Jan. 18, 2015), https://www.nytimes.com/2015/01/19/technology/banking-start-ups-adopt-new-tools-for-lending.html [perma.cc/P5CU-C2X2]; *see also* Duhigg, *supra* note 47. See discussion of this example in Katyal, *supra* note 7, at 98.

[64] A common method for representing machine learning input data is "one-hot encoding," where each possible value of a particular input type has a separate variable. Only one of the variables is set to one for an input instance. The rest of the variables are set to zero, which inherently creates sparse inputs. *See* John T. Hancock & Taghi M. Khoshgoftaar, *Survey on Categorical Data for Neural Networks*, 7 J. BIG DATA 28, 30

binary variables offer advantages in data collection efficiency *and* in predictive accuracy since the analysis does not require a prespecified theory of the functional relationship between inputs and outputs.[65]

Second, if binary variables were given a free pass from a SCMBD audit, the loophole would invite strategic behavior among firms that can plausibly justify structuring what would normally be a continuous variable into a series of binary ones.[66] Thus, identifying SCMBDs will require at least some amount of judgment and consensus. The guiding lights for this judgment, as many of our examples illustrate, are theoretically sound explanations linking the binary factor to the objective function and decisional outcome.

2. Big Difference

Now, let's consider the outputs — or the "Big Difference." Again, we should distinguish between binary and continuous outcomes. With continuous outcome variables such as a score, a salary, a prison sentence, or an insurance interest rate, the difference between two outcomes can be measured and given a sense of scale that is not possible with discrete or categorical outcomes. Whether a difference is "big" could be established in advance using absolute value terms (e.g., a 1 percentage point difference in interest rate), relative value terms (e.g., a 20% change in either direction), or in relative distributional terms (e.g., a 10 percentile point change, or a 0.5 standard deviation change, in predicted outcome).

When input *and* output variables are continuous, the small changes and big differences could be defined together. A SCMBD could be defined as any rule in an algorithm's model that produces a ratio between a change in inputs and a concomitant difference in outputs is too small. That is, if the change in inputs ("$\ _1$") is divided by the difference in outputs ("$\ _2$"), the ratio ($|\ _1/\ _2|$) could be used to define

(2020) (surveying deep learning scholarship and concluding that one-hot encoding is the most common method of representing categorical input data).

[65] By contrast, a regression equation must assume, for example, that an output variable has a linear or quadratic or other specific type of relationship with the input variable. With binary variables, the relationship can have discontinuities at every possible value of the input.

[66] Binary variables can often be transformed into continuous ones, too. For instance, the binary "smoking" attribute (i.e., whether or not one smokes) could be transformed to a continuous variable using a question asking about the average number of cigarettes consumed per day. Similarly, questions about earned degrees could be replaced with those inquiring about the years of education. *See* Kolber, *supra* note 54, at 661 (explaining this dynamic in the context of speeding).

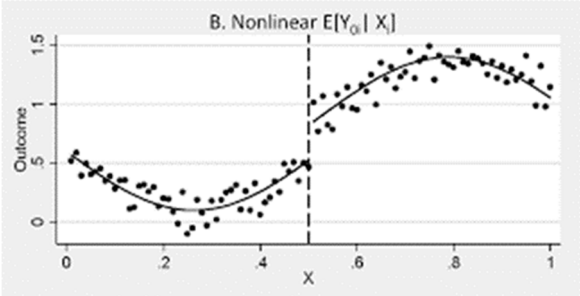
both small and large. A SCMBD could be flagged any time the ratio falls below a presumptively acceptable ratio ("AR") set in advance ($|\ _1/\ _2|$<AR). Using a ratio still requires discretion in deciding whether the differences ($\ _1$ and $\ _2$) will be measured in absolute, relative, or distributional terms, but it does not require the auditor to set a threshold for each input and output individually. This offers some conceptual advantages because a single cutoff ratio can detect imbalances of any sort between inputs and outputs.[67] It would flag instances in which a small change makes a big difference, and it would also flag instances when a medium sized change makes a *huge* difference.[68]

Data visualization could also be used to eyeball a SCMBD. If we were to look for SCMBDs in a very simple model that uses a single input

[67] In other words, it enables ranking that is "ordinal" — ranking which is relative to one another, as opposed to ranking which is nominal or cardinal which rely on the categories' inherent values. For more on this taxonomy, see Marion Fourcade, *Ordinalization: Lewis A. Coser Memorial Award for Theoretical Agenda Setting 2014*, 34 SOC. THEORY 175, 176-78 (2016).

[68] It also allows auditors to aggregate several input variables and detect instances when a set of tiny changes cause too big of a difference. The numerator of the equation can be composed from several (absolute value) changes that are each tiny, but that together cause a large difference in outcome (the denominator). There are other ways to test for the effect of several tiny changes, too. For example, subsets of the input variables with tiny changes could be dropped to see if any of the remaining variables produce a SCMBD. This is somewhat similar to a "feature selection" analysis that can be used to avoid overfitting in Machine Learning. *See* Lehr & Ohm, *supra* note 51, at 700-01. Other solutions might include creating an index variable generated from multiple inputs. If a cluster of input variables are highly correlated, an auditor could test a single variable that adds (or subtracts, where the effect on output is negative) the values of all of the correlated variables together to see whether small changes along the distribution of *that* index variable yield large differences. We are continuing to explore options for aggregating small changes because we suspect it will be an important aspect of SCMBD in the wild, and because this could also help identify attempts to game an audit for SCMBD.

variable to predict an outcome, a SCMBD would look like a sharp cliff or discontinuity in the function like so:



With multiple inputs, a possible approach would be to imagine a multidimensional hyperplane that represents outcomes over several variables at a time.[69] If we could picture this, a SCMBD would look like an out-of-place nodule, similar to what bad gerrymandering does to an election map. Indeed, the law has turned to mathematical tools to define, identify, and challenge precisely these sorts of violations of "compactness" as a result of redistricting to precisely identify unacceptable gerrymandering.[70]

Binary and crude categorical outcome variables are trickier because they function as cut-offs that inherently risk creating SCMBD problems wherever the line is drawn. However, in practice, algorithmic decision-making programs estimate outcomes on a continuum. Even if they report recommendations in stark or binary categories, a continuous score, probability, or estimated likelihood ratio usually underlies the recommendation. For example, a recommendation to hire a candidate or to release a criminal defendant on bail is typically derived from a much more fine-grained prediction about the probability of some specific events.[71] For ease of discussion, we will refer to the fine-grained

[69] We discuss conceptually related methods for identifying possible SCMBD problems in Part IV.

[70] For instance, see Aaron R. Kaufman, Gary King & Mayya Komisarchik, *How to Measure Legislative District Compactness if You Only Know It When You See It*, 65 AM. J. POL. SCI. 533, 533-34 (2021).

[71] For example, "[t]he Recidivism Risk Scale is a regression model that has been used in COMPAS since 2000. This regression model was trained to predict new offenses in a probation sample. The system transforms a linear predictor from the regression model to a decile score." Tim Brennan, William Dieterich & Beate Ehret, *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 CRIM. JUST. & BEHAV. 21, 25 (2009). In other words, the COMPAS scores take a continuous variable output and transform it into a decile (10-point) score for ease of comprehension. That

algorithmic prediction as assessment outcomes, which can be distinguished from the ultimate treatment.

C. Extant Anti-SCMBD Policies

At first blush, it may seem that a discussion of SCMBDs is premature — that the issue has no real-world significance today and is therefore not ripe for a thorough policy analysis. In fact, anti-SCMBD sentiment is already finding its way into legal frameworks used by federal agencies.

For example, consider the Federal Trade Commission's 2008 settlement with CompuCredit, a firm that targeted high-risk borrowers for low-limit credit cards.[72] The case was a mundane enforcement action in many ways because CompuCredit charged high fees without sufficient disclosure, and sometimes changed credit limits to fall *below* the borrower's current balance so that CompuCredit could assess even more fees.[73] These alleged acts are bread and butter violations that the FTC routinely targets.[74] But Count III of the complaint concerned a practice that has captured the attention and imagination of observers. That count alleged that CompuCredit changed its borrowers' credit limits when the credit card was used for certain types of transactions including at pawn shops, massage parlors, counseling services, and billiard halls.[75] Although CompuCredit reserved the right to reduce available credit based on "behavioral scoring" in its issuing agreements,[76] the FTC alleged that this did not provide sufficient disclosure for the noted actions that followed.[77]

The FTC nominally treated the behavioral credit limiting as a deception/disclosure problem, but the complaint relies on an implicit assumption that changing credit limits based on the details of a borrower's transactions is alarming and, therefore, must meet

score is subsequently used to make other discrete choices about, e.g., whether to release a detained suspect.

[72] Order for Service of Report and Recommendation at 1, F.T.C. v. CompuCredit Corp., No. 08-CV-1976-BBM-RGV, 2008 WL 8762850 (N.D. Ga. Oct. 8, 2008).

[73] Complaint at 7-8, F.T.C. v. CompuCredit Corp., No. 08-CV-1976-BBM-RGV, 2008 WL 8762850 (N.D. Ga. June 10, 2008).

[74] *See Credit Cards*, FED. TRADE COMM'N, https://www.ftc.gov/news-events/media-resources/consumer-finance/credit-cards (last visited Dec. 29, 2021) [perma.cc/ASD7-6EKA].

[75] Complaint, *supra* note 73, at 34-35.

[76] *Id.* at 26.

[77] Order for Service of Report and Recommendation, *supra* note 72 at 2; Ryan Singel, *Credit Card Firm Cut Limits After Massage Parlor Visits, Feds Allege*, WIRED (June 20, 2008, 12:58 PM), https://www.wired.com/2008/06/credit-card-fir/ [perma.cc/X6UQ-UST7].

heightened standards of notice. Public commentary about the case shows that people have a visceral response to the idea that a single transaction could have a significant impact on credit limits,[78] and this no doubt drove the FTC to require heightened notice, too.

The case makes more sense when considered as an action against SCMBD. To be sure, concerns about behavioral credit-limiting can also be understood using conventional notions of privacy rather than the more specific concern about SCMBDs. After all, one of the Fair Information Practice Principles recognizes a harm when information collected for one purpose (e.g., to generate a credit card bill) is used for another purpose (e.g., to fine-tune credit scores and credit limits).[79] But since the purpose limitation principle has never been adopted by the FTC as a general requirement for fair practices under the FTC Act, and since, as a matter of common practice, the principle is violated all the time in the digital economy, we suspect that this *particular* style of repurposing was objectionable because it taps into an intuitive but unstated anti-SCMBD sentiment.[80] It seems wrong that a little decision, like visiting a billiard hall, could cause a large fluctuation in access to credit.[81]

The Equal Employment Opportunity Commission has developed guidelines that incorporate an anti-SCMBD position even more clearly. The EEOC frequently must make determinations about prima facie cases of discrimination under Title VII, the ADEA, and the ADA in situations where there is no slam dunk evidence of inequality. Because many employers are too small (with an even smaller pool of minority employees in higher ranks) to have rich employee data that could reveal a discriminatory pattern of hiring or compensation, the agency uses assumptions about how a member of a majority group would be treated

[78] Jessica Silver-Greenberg, *Your Lifestyle May Hurt Your Credit*, BLOOMBERG (June 19, 2008), https://www.bloomberg.com/news/articles/2008-06-18/your-lifestyle-may-hurt-your-credit [perma.cc/FZH3-8U6V].

[79] This constitutes a violation of the purpose specification principle, which is a key element of data protection ethics and law. U.S. DEP'T OF HEALTH, EDUC. & WELFARE, RECORDS, COMPUTERS AND THE RIGHTS OF CITIZENS: REPORT OF THE SECRETARY'S ADVISORY COMMITTEE ON AUTOMATED PERSONAL DATA SYSTEMS, at xx-xxi (1973).

[80] The FTC has brought enforcement actions where a business retroactively changed its privacy practices, but those claims are both rare and involve a representation to the consumer about how the information would be used. *See* Complaint at 4-7, *In re* Facebook, Inc., 92 F.T.C. 3184 (2011) (No. 19-cv-2184); Complaint at 3, *In re* Gateway Learning Corp., 42 F.T.C 3047 (2004) (No. C-4120).

[81] Since these credit cards were pitched to risky borrowers, the maximum credit limit was $3,000. Thus, virtually any change in credit limit was likely to be large, at least as a proportion of original credit limit.

if they had the same inputs as the minority complainant.[82] EEOC guidance explains that "the difference in education, experience, training, or ability must correspond to the compensation disparity. Thus, a very slight difference in experience would not justify a significant compensation disparity."[83] In other words, a SCMBD would be evidence of discrimination unless the employer can provide a sufficient, non-vague explanation.

[82] *See* U.S. EQUAL EMP. OPPORTUNITY COMM'N, EEOC-CVG-2001-3, SECTION 10 COMPENSATION DISCRIMINATION (2000) ("Furthermore, the difference in education, experience, training, or ability must correspond to the compensation disparity.").

[83] *Id.*

Example 4[84] in the guidance document illustrates the concept very clearly, so we reproduce it here:

Example 4: Same as Example 3, except A. Smith has more years of experience and a higher average performance rating than A. Jones.

Employees in Protected Class	Salary	Alleged Factors Affecting Salary	Do Proffered Reasons Explain Disparity?	Employees Not in Protected Class	Salary	Alleged Factors Affecting Salary
A. Jones (CP)	$23,000	-3 yrs. exp. -avg. 2 perf. rating	No - A. Jones' pay differential is out of proportion to the difference in explanatory factors.	A. Smith	$31,000	-4 yrs. exp. -avg. 3 perf. rating
				B. Thomas	$34,000	-5 yrs. exp. -avg. 4 perf. rating
				C. Adams	$37,000	-5 yrs. exp. -avg. 5 perf. rating
				D. Buckley	$40,000	-6 yrs. exp. -avg. 5 perf. rating

In this example, the complainant (Jones), who is a member of a protected racial class, is paid $8,000 per year less than the most similar white employee (Smith). The employer explained the difference in salary by reference to the complainant's tenure and average performance

[84] *Id.* at ex. 4.

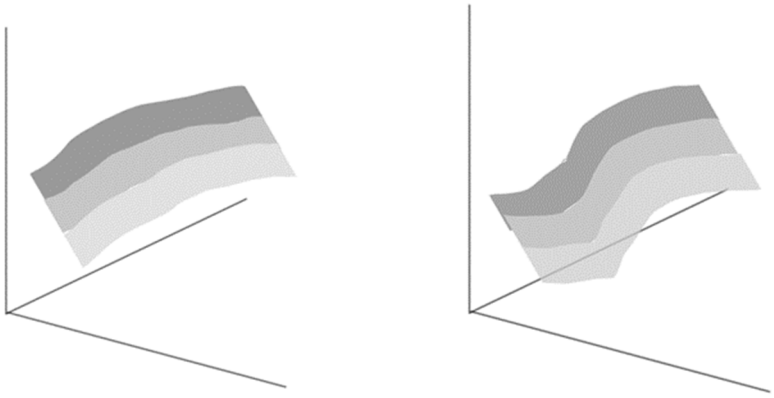
rating, both of which are lower than Smith. Putting aside the problems inherent in relying on subjective performance ratings (which themselves could very well be biased[85]), the EEOC concludes that experience and performance ratings cannot justify the salary differential.[86] The reason is that the data on all the other employees shows a pattern suggesting a difference of one year in experience, or a difference of one point in performance rating, or both, accounted for at most a $3,000 difference in salary among non-minority employees. The gap between the complainant's compensation and the nearest non-minority employee's was unusually large ($8,000 instead of $3,000; note that there was no non-minority employee with equal or lower experience or performance rating metrics). Thus, the EEOC concluded that the deviation is presumptively discriminatory.

Because there are no directly comparable employees, as will frequently be the case with smaller employers, the EEOC has enabled claims to proceed by extrapolating from the available data to estimate what a similar majority race employee would have been paid. That extrapolation uses a presumption of linearity — that is, a rejection of SCMBDs — in compensation functions. To put it visually, a compensation function that looks like the red plane in Figure 1 presents an initial showing of discriminatory employment practices if employees in the lower valley of the pay function are disproportionately or exclusively members of a protected class.

[85] *See* Arin N. Reeves, *Written in Black & White: Exploring Confirmation Bias in Racialized Perceptions of Writing Skills*, *in* NEXTIONS: YELLOW PAPER SERIES (2014), https://nextions.com/wp-content/uploads/2017/05/written-in-black-and-white-yellow-paper-series.pdf [perma.cc/3JAF-9K92]; Joseph M. Stauffer & M. Ronald Buckley, *The Existence and Nature of Racial Bias in Supervisory Ratings*, 90 J. APPLIED PSYCH. 586, 588 (2005).

[86] U.S. EQUAL EMP. OPPORTUNITY COMM'N, *supra* note 82, at ex. 4.

Figure 1. Presumptively valid (left) and invalid (right) compensation functions under EEOC guidance, if employees in the lower compensation area are disproportionately members of a protected class. The three-dimensional illustration represents the EEOC's two variables used to compare justification for salaries: seniority and performance rating.



The employer is still able to justify the cliff-like relationship between experience/performance ratings and salary with some nonspeculative explanation, or if the employer can identify some third factor, the SCMBD may be acceptable. But without such an explanation, a cliff-like shape in a compensation curve that puts minority employees at a disadvantage would need correction and redress.[87] Since we assume that EEOC guidance is well-known to HR professionals, it is reasonable to believe that large firms are already working with anti-SCMBD policies in hiring and compensation.

To be clear, the EEOC is not reflexively against nonlinear relationships that produce SCMBDs. Rather, they embrace skepticism of a SCMBD when it has a disparate impact on protected classes and when they cannot otherwise be explained. This is entirely consistent with our discussion of the interaction between SCMBDs and discrimination as you will see in Part II. But because SCMBD is underdeveloped, there is no overarching theory guiding the EEOC (or any other agency) to know when or why these specific forms of SCMBDs are a problem when others are not. Indeed, the very same EEOC document allows for the sorts of crisp cut-offs and category

[87] The EEOC can have access to this sort of data as part of its broad authority to request employment related data from firms under investigation (although the Trump administration has limited these requirements).

boundaries that will produce SCMBDs by starting its analysis of discriminatory compensation with two binary qualifications: (1) recognizing "minimum objective qualifications," and (2) defining a "pool of comparators" among the firm's employees.[88] Thus, even the EEOC is tolerant of many SCMBDs. A small change in qualifications can justify large differences in compensation when the difference straddles the line defining "minimum objective qualifications," for example. Likewise, the EEOC will allow relatively large differences in compensation between two classes of employees if the classes are not in the same "pool of comparators."[89] Thus, the EEOC is skeptical of SCMBDs that occur *within* a pool of employees but seems to ignore them when they occur between two similar employees who are technically in different pools. We can expect disputes to emerge over the meaning of "minimum qualifications" or "pool of comparators" since those terms define whether SCMBDs are presumptively permissible or not.[90]

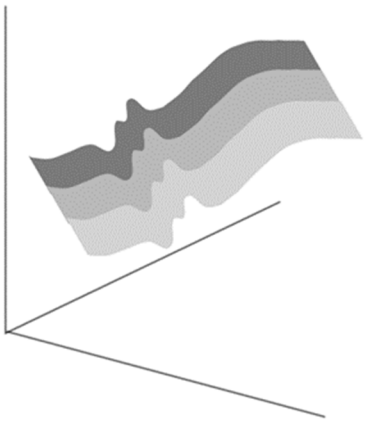
If the EEOC is suspicious of compensation functions that look like the red graph in Figure 1, it should be even more concerned about a function that looks like Figure 2 below. And yet, as we explain in the following Sections, there is reason to believe that modern machine learning systems will increasingly introduce these types of bumps, nodules, and discontinuities.

[88] U.S. EQUAL EMP. OPPORTUNITY COMM'N, *supra* note 82.

[89] *Id.* ("While differences in qualifications, experience, and education ultimately may explain a pay differential, such factors require a pretext or disparate impact analysis to determine whether they are legitimate, and thus should be considered only after the pool of comparators has been determined . . . .").

[90] For example, whether a professor in a history department should or should not be considered in the pool of comparators for a political science professor will determine whether their compensation is subject to the EEOC's SCMBD limits. *See* Bret G. Daniel & Erin B. Edwards, *Employment Law*, 54 U. RICH. L. REV. 103, 108-11 (2019).

Figure 2. An example of an alternative compensation function that would pose even greater SCMBD problems under EEOC guidance.



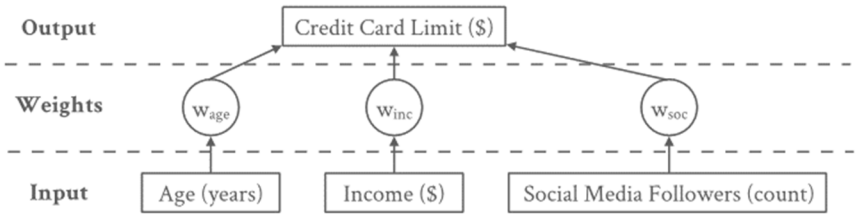
D. Machine Learning Will Introduce More SCMBD

In this Section, we explain five illustrative ways that SCMBD phenomena will emerge from modern machine learning systems: categorical treatments, categorical input variables, model nonlinearities, overfitting, and counterintuitive feature importance.

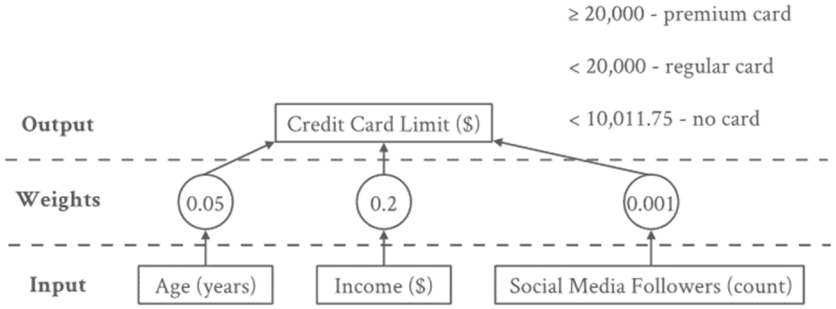
For purposes of this explanation, assume that a bank is developing an algorithm for whether to offer a certain credit card to a consumer. The bank uses conventional credit factors that are generally predictive of a consumer's creditworthiness (age and income), as well as an unconventional input that is loosely correlated with creditworthiness (number of social media followers). Also assume that the algorithm has a simple design: the output is a weighted sum of inputs, where the weights are inferred from a dataset of previously issued credit cards.[91] This type of algorithm is called a perceptron, and it is the building block for more sophisticated neural network machine learning models

[91] For simplicity, we use multistep and linear activation functions with no biases in the perceptron model. We also use the same overall model for both classification and regression tasks. To be precise, the models we propose are variations of perceptrons, which traditionally have a binary output, a step activation function, and optionally a bias term.

(termed "deep learning").[92] The following diagram represents this simple algorithm.



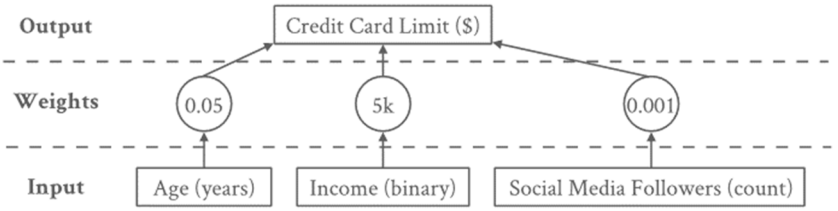
One cause of SCMBD is using binary or categorical treatment for an algorithmic decision-making system. The reason this design creates a risk of SCMBD is that it incorporates one or more output thresholds. If a small change in inputs between two individuals happens to cross a threshold, SCMBD will occur. The diagram below shows the credit card perceptron with categorical output and example weights.



If an applicant is thirty years old, has $50,000 in annual income, and has 10,000 social media followers, the system will (barely) issue a regular credit card. But if the applicant is one year younger, or makes one dollar less per year, of has one fewer follower, then the system will decline the application. A similar boundary condition would exist for an applicant who is forty-six years old, earns $99,683, and has 61,100 followers, as compared to an otherwise similar person who has only 61,099 followers; the former would qualify for a premium card, while the latter would qualify for only a regular card. Although the scores (the assessment dimension) of the two will be very similar, the assessments will place the individuals on opposite sides of a treatment threshold.

[92] STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 676-86, 785-89 (3d. ed. 2010) (describing perceptrons and how they relate to deep learning).

Another potential source of SCMBD is using binary or categorical inputs to an algorithm. Consider the same example as before but replace the continuous income variable with a binary variable for whether income is at least $50,000.



Again, assume an applicant who is thirty years old, has $50,000 in annual income, and has 10,000 social media followers. The algorithm will issue that applicant a card with a $5,011.50 spending limit. But if the applicant makes just one dollar less per year, the spending limit craters to a negligible $11.50.

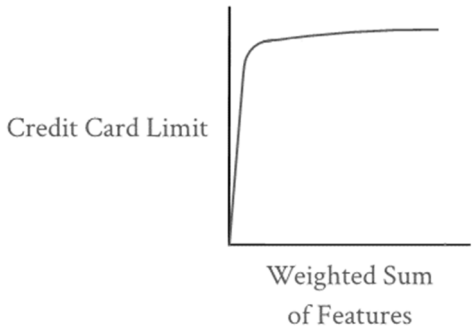
A third way in which machine learning can result in SCMBD comes from model nonlinearities. Our example perceptron uses a simple weighted sum to predict a credit card limit. Modern machine learning systems, however, typically are not so linear.[93] A key insight of deep learning, for instance, is that by connecting a number of perceptron-like networks with nonlinearities it is possible to construct classifiers with complex behavior.

Suppose that the bank modifies its credit card limit perceptron, so that the model additionally applies a steep sigmoid function to the weighted sum of input variables. This type of function is very common in neural network systems (along with step functions and rectifier functions, which are hybrids of linear functions and step functions).[94] Because of the slope of the function, quantitatively small changes in the

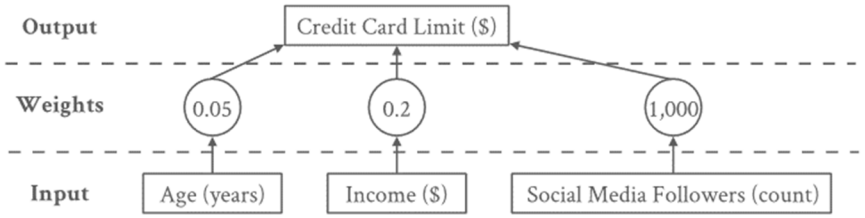
[93] *See, e.g.*, STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 751-54 (4th ed. 2020) (describing nonlinear functions that are commonly used in deep learning systems).

[94] *Id.*

weighted sum of input variables will necessarily lead to quantitatively large changes in the output variable.



A fourth possible cause of SCMBD is overfitting, where a machine learning model becomes sensitive to trends in training data that are not representative of real-world data. Assume, for example, that the bank has just started collecting social media data, and so far only from wealthy clients. The model might infer a strong relationship between a customer's follower count and their approved credit card limit, as in the example below.



In this overfitted model, every single additional social media follower increases the credit card limit by $1,000. A similar overfitting problem could occur if the bank represented social media follower counts as a series of discrete categories (e.g., 400–500 followers, 500–600 followers, etc.). The bank might fail to provide the model with training examples for each of these categories, and as a result, the model might exhibit SCMBD for similarly situated people who cross between categories.

A fifth way in which SCMBD might occur in a machine learning algorithm is due to trends in the real world. As we discuss further below, natural and societal processes often have SCMBD properties. Meanwhile, modern machine learning excels at picking out subtle, nonintuitive interactions from a massive volume of data; the latest

sophisticated models are trained on terabytes of data, include tens of thousands of input variables, and incorporate billions of parameters for encoding these interactions.[95] A model could reflect a strong relationship between a particular input variable and an output variable — and that relationship might exist in the real world — but we cannot explain why the relationship exists.

Imagine, for example, that the bank modifies its credit card algorithm to consider (among many other features) the color of an applicant's car. After training the algorithm, the bank discovers that the weight for red cars is very negative. The bank confirms, in a large-scale regression analysis of its existing customers, that there is a strong inverse correlation between owning a red car and receiving a higher credit card limit. The bank can only speculate about why the relationship exists — maybe a red car reflects genuine irresponsibility and risk-taking, for example. Maybe it reflects historical habits of creditors that discriminated against loan applicants that they regarded as ostentatious even though in fact they are equally creditworthy. Whatever explains it, this is a SCMBD problem: a small change (car color) results in a big difference (a credit card offer).

These five sources of SCMBD are distinct, but they are neither mutually exclusive nor collectively exhaustive. Binary or categorical inputs, for example, might reflect a real-world process that is not continuous. A model's nonlinearities could also contribute to overfitting, for instance. We describe these model behaviors to illustrate how routine machine learning methods can result in SCMBD problems.

While there are not yet a lot of anecdotes about SCMBDs in the wild, so to speak, we can expect that they exist and will increase in number.[96] In many cases, they will be hard for outsiders to uncover unless algorithms are public-facing (such as the Google search bar and query

[95] *See generally* Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jada Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei, *Language Models Are Few-Shot Learners*, ARXIV (July 22, 2020), https://arxiv.org/abs/2005.14165 [https://perma.cc/HK9M-5YKD] (describing the OpenAI GPT-3 model for natural language processing).

[96] Note that the European Union's Ethics Guidelines for Trustworthy AI assumes that SCMBD dynamics are constantly unfolding. HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, ETHICS GUIDELINES FOR TRUSTWORTHY AI 21 (2019) ("Moreover, sometimes small changes in data values might result in dramatic changes in interpretation, leading the system to e.g.[,] confuse a school bus with an ostrich.").

results) or unless there is an external audit. Whether audits for SCMBD should be encouraged or required, however, depends on whether SCMBDs are a form of algorithmic unfairness. We tackle this question next.

II. IS SCMBD BAD?

Having tended to definitions, we turn to why we should care. In this Part, we examine whether the common aversion to SCMBDs is articulable and well-justified. To do this, we map out five arguments that SCMBDs might be undesirable. Three of the arguments (inaccuracy, bias, and strategic behavior) are established and well-recognized forms of unfairness. Two others (disproportionality and hyperselectivity) are not. Thus, probing the instinctive reaction against SCMBDs has paid dividends. It helped us unearth additional considerations to add to the long list of factors that AI ethicists can use to assess the overall fairness of an algorithm.

A. Established Forms of Unfairness

Sometimes SCMBDs raise eyebrows because they are suggestive evidence of a problem we are accustomed to worrying about. First, when a small change seems to have outsize effect, the algorithm may be using, at least temporarily, a spurious correlation to make its predictions. This would undermine the accuracy of decision-making. Second, if the SCMBD causes a disparate impact among groups with protected characteristics, that could signify unfair bias. And third, if the small change that causes a big difference concerns a variable that is easily manipulated by the data subject, the algorithm is likely to induce strategic behavior — gaming by the data subjects who have the information and moral disposition to do so. Thus, SCMBD may seem intuitively to be a nuisance because of the way it interacts with these established forms of unfairness. We briefly consider each.

1. SCMBD and Inaccuracy

Accuracy problems can cause an algorithm to make decisions in ways that are not only inefficient but patently unfair. When algorithms are used by state agencies or important private actors (like employers) to make highly consequential decisions, a needlessly flawed assessment is a moral failing as much as it is a programming one — especially when the errors generated are systematic, substantial, and not self-corrected. Subjection to arbitrary processes is degrading and disempowering. It is

also a violation of equality in the individual (rather than group) sense if like people are treated differently.

SCMBDs might be an indication that the algorithm is inaccurate. Large weights assigned to one or several input values might indicate a model that includes a spurious correlation. This is a well-known problem when machine learning algorithms wind up overfitting to training data. When a large number of correlations are mined, some may be the product of random chance.[97] With an abundance of input variables, a machine learning model may fall victim to the "curse of dimensionality" and rely on intricate variable relationships that do not generalize to accurate predictions.[98] SCMBDs therefore raise a red flag, indicating that the prediction model might be less accurate than it can be.

Examples of spurious correlations abound in the AI literature, and are frequently held up as cautionary tales. One example, possibly urban legend, involves an army-developed image classification system that was trained to recognize images containing a tank.[99] In some stories it's the U.S. army, in others the Red Army.[100] After a training period and what seemed to be tremendous success, accuracy started to plummet.[101] The researchers discovered that the system had been trained using a set of photographs that depicted tanks in the sun and other items in overcast light, so the classification relied heavily on whether there were or were not clouds in the images (or so the story goes).[102]

---

[97] *See generally* NASSIM NICHOLAS TALEB, FOOLED BY RANDOMNESS: THE HIDDEN ROLE OF CHANCE IN LIFE AND IN MARKETS (2005); *Overfitting in Machine Learning: What It Is and How to Prevent It*, ELITE DATA SCI., https://elitedatascience.com/overfitting-in-machine-learning#how-to-detect (last visited Dec. 19, 2021) [https://perma.cc/LN7Y-QRXB].

[98] *See* RUSSELL & NORVIG, *supra* note 93, at 653-55 (describing the problem of overfitting in supervised machine learning); Michel Verleysen & Damien François, *The Curse of Dimensionality in Data Mining and Time Series Prediction*, 3512 LECTURE NOTES COMP. SCI. 758, 758-59 (2005) (noting that the "curse of dimensionality" can cause overfitting in nonlinear models).

[99] BLAY WHITBY, ARTIFICIAL INTELLIGENCE: A BEGINNER'S GUIDE 53 (2012); *The Neural Net Tank Urban Legend*, GWERN, https://www.gwern.net/Tanks (last updated Aug. 14, 2019) [https://perma.cc/S7JF-7EDG].

[100] *Embarrassing Mistakes in Perceptron Research*, WEB STORIES (2011), https://www.webofstories.com/play/marvin.minsky/122 [https://perma.cc/RY88-J2LB] (video interview of Marvin Minsky claiming it was the U.S. Army); yorksranter, *It Was Called a Perceptron for a Reason, Damn It*, THE YORKSHIRE RANTER (Sept. 30, 2017), https://www.harrowell.org.uk/blog/2017/09/30/it-was-called-a-perceptron-for-a-reason-damn-it/ [https://perma.cc/S4UG-9U7Z](claiming it was the Red Army).

[101] *The Neural Net Tank Urban Legend*, *supra* note 99.

[102] *Id.*

An image classifier developed by Xiaolin Wu and Xi Zhang provides a more recent and non-mythical example of overfitting.[103] Their paper claims that an algorithm trained on closely cropped photographs from government IDs can accurately predict criminality.[104] Even putting aside objections to the plausibility, motivation, and potential future uses of such an algorithm (of which there are many), the accuracy that the researchers claim to have achieved is . . . hard to believe. The authors claim that for their best model, only six percent of predicted criminals are false positives.[105] A critique by the lead of one of Google's AI groups pointed out that from the examples published in the Wu and Zhang paper, the non-criminals are wearing collared shirts and the criminals are not.[106] Given that the algorithm was trained on a relatively small set of data (just 1,856 photographs), differences in clothing could be a feature that is seized on by an image classifier. If so, the false positive rate will skyrocket when the classifier is used on images of law-abiding citizens who did not wear collared shirts to their government ID photo shoot.[107]

[103] XIAOLIN WU & XI ZHANG, AUTOMATED INFERENCE ON CRIMINALITY USING FACE IMAGES 1-8 (2016), https://arxiv.org/pdf/1611.04135v1.pdf [https://perma.cc/38LM-VJYP].

[104] *Id.*

[105] *Id.* at 4.

[106] Blaise Agüera y Arcas, Margaret Mitchell & Alexander Todorov, *Physiognomy's New Clothes*, MEDIUM (May 6, 2017), https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a [https://perma.cc/5SYJ-KBYR].

[107] Wu and Zhang's methods for validating their results cannot increase confidence in their results, either, since all they showed is that if they randomly assigned labels to the training dataset, the classifier performs poorly. WU & ZHANG, *supra* note 103, at 4-5. This could very well be due to the lack of a tell like a collared shirt.

Figure 3. Excerpt from Wu & Zhang

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Wu and Zhang, for their part, attribute the success of their algorithm to the classifier's use of facial features including, most importantly, the "upper lip curvature."[108] This is a rather technical way of saying that the labeled criminals are frowning. The great emphasis on the frown during a government photo for use to predict criminal behavior is as much of a SCMBD as a collared shirt, and the correlation could very well be just as spurious. And if it isn't spurious now, it will be soon when criminals learn to smile at their government photo shoot.

There are well-known methods for attempting to detect and fix overfitting problems so long as the developer is knowledgeable and motivated enough to correct for them.[109] As machine learning do-it-yourself applications become available to the general public, inaccurate SCMBDs may increase as the unwary fail to correct for overfitting. Sometimes overfitting might even *serve* a business purpose for firms that

[108] *Id.* at 5.

[109] *See* Benyamin Ghojogh & Mark Crowley, *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial*, ARXIV (May 28, 2019), https://arxiv.org/abs/1905.12787 [https://perma.cc/MR6W-5LER] (describing methods for responding to overfitting in supervised machine learning systems).

are selling an AI version of snake oil, as when Cambridge Analytica claimed to be able to predict political profiles from Facebook data with more precision than they actually could.[110] Thus, to the extent SCMBD overlaps with the problem of algorithm inaccuracy, private and public policies to combat error can leverage the concept of SCMBD to help suss out and reduce flaws.

To date, legal scholars have used due process as a model for challenging inaccurate and arbitrary algorithmic processes.[111] As Aziz Huq explains, challenges to arbitrary algorithms can rely on the due process theories and case law of both the procedural and substantive variety.[112] Similarly, administrative agencies must meet a minimum level of accuracy in their decisions in order to avoid actions that are "arbitrary and capricious."[113] While the constitutional and administrative constraints apply only to government decision-making, the rationale translates readily to legal regimes that restrict how private actors can conduct their hiring, lending, and other affairs insofar that they are carrying out public-like tasks.[114]

By and large, the law does not currently impose a minimum threshold of accuracy for these private decisions even when they are made on random or sentimental bases. In employment, for example, the standard approach to at-will employment is that hiring and retention decisions can be made on any basis so long as they are not made on a specific, prohibited basis (e.g., related to race, age, or a disability that can be accommodated).[115] The reason for law's conservatism (or limited reach) on these matters is not that employment should be untethered from a

[110] "The real-world accuracy of these predictions — when used on new individuals whose data had not been used in the generating of the models — was likely much lower." Letter from the Information Commissioner's Office to Julian Knight MP on Cambridge Analytica Investigation, ICO/O/ED/L/RTL/0181, at 17 (Oct. 2, 2020).

[111] Huq, *supra* note 7, at 1908-09 (referring to the work of Citron as well as Shultz & Crawford).

[112] *Id.* Huq also argues that an algorithm that has a poor fit between its outcome measure and the objective of the decision-making process can also lack "due process" and possibly fail a challenge under the Equal Protection clause. Note 163 begins to develop this notion, by explaining that the Equal Protection Clause could be applied when the governmental actions are not proven to be rational. *Id.* at n.163.

[113] 5 U.S.C. § 706(2)(A).

[114] So one of us has argued. *See* Tal Z. Zarsky, *Correlation Versus Causation in Health-Related Big Data Analysis: The Role of Reason and Regulation*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS 42, 54 (I. Glenn Cohen et al. eds., 2018).

[115] *Protections Against Discrimination and Other Prohibited Practices*, FED. TRADE COMM'N, https://www.ftc.gov/site-information/no-fear-act/protections-against-discrimination [https://perma.cc/5W2G-8WFL] (last visited Feb. 22, 2022) (listing applicable antidiscrimination statutes that constrain employment decisions).

consideration of merit, or even that the definition of "merit" will be elusive and context-dependent (though this may well be true). Rather, it will usually be in a decision-maker's self-interest to make accurate predictions, and the market will do enough to discipline decision-makers who are making inefficient decisions.[116] If self-interest really does provide enough incentive for firms to improve the accuracy of their models, there is little reason to develop legal mandates for minimum levels of accuracy. After all, regulatory intervention is not priceless or error-free.[117] But many market participants and decision-makers (including government decision-makers) might not be under the sufficient competitive pressure that motivates firms to maintain and improve accuracy. If assumptions about accuracy and rational self-interest are misplaced, flawed algorithms will be able to persist, and unfairness will unfold. In some contexts, legal requirements to maintain accuracy may be well-justified, and SCMBD audits can help detect some potential sources of inaccuracy.

However, many SCMBDs will be non-erroneous. That is, they will continue to have predictive validity for reasons that are partially or completely obscure. We must continue to excavate our instincts about SCMBD dynamics in order to understand what problems they pose in the many instances where they reveal a relationship between inputs and outputs that seems to have reasonable predictive power.

2. SCMBD and Discrimination

Perhaps the greatest concern animating exploration and discussion of algorithm fairness revolves around unintended biases. When bias is imbedded in automated algorithms, the algorithms can aggravate (or, at least, fail to improve) existing racial, gender, or class disparities.[118] And it is possible for algorithm bias to emerge even if a machine learning

[116] Because accuracy is often (if not always) in the self-interest of the firms using an algorithm, some scholars engaged in the AI fairness debates treat accuracy or efficiency as distinct from fairness. In their book, *The Ethical Algorithm*, for example, Michael Kearns and Aaron Roth use the term fairness in a way that is definitionally distinct from accuracy. MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN 74-78 (2020). They are concerned primarily with inequitable outcomes or distributions of error, which we discuss below. This sets up an "accuracy versus fairness" rhetorical battle that we think is misleading. By recognizing the importance of accuracy for a just and fair system of resource allocations, the "accuracy versus fairness" debate is properly understood as one specific cross-section of a larger "fairness versus fairness versus fairness . . ." debate.

[117] For a discussion of this balance, see Zarsky, *supra* note 114, at 55.

[118] For a comprehensive and careful analysis of algorithmic disparate impact, see Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 13, at 677-94, 715-23.

algorithm has no direct measure of a subject's race or gender because other input factors that correlate with race and gender can reproduce disparities.[119] Indeed, machine learning ethicists are coming around to recommending that machine learning systems actually *have* access to race and gender so that certain types of checks and corrections can be made automatically.[120]

SCMBD dynamics can be useful tells that an algorithm may have unaccountable disparate impact on vulnerable groups. To be clear, a model that puts great weight on a small change in inputs could affect all subgroups in the same or similar ways. But in cases where some historically disadvantaged subgroups receive worse outcomes as a result of a SCMBD, the SCMBD may be the cause or the symptom of algorithmic bias that keeps discrimination alive.

Algorithms can imbed biases in many ways, including via biased objective functions, biased errors, and biased outcomes.[121] We explain the relationship of each of these to SCMBD.

a. Biased Objective Functions

SCMBDs with a disparate impact could signal that the objective function that is being predicted is not well-tailored to the key outcome that decision-makers should care about. There are multiple points in the algorithm design and training process where human error, past discrimination, and other social factors can bake bias into the algorithm.[122] The most consequential decision of this sort is during the selection of the objective function — that is, the decision about what measurable or coded output variable the algorithm should be forecasting. If an algorithm is trained to optimize the prediction of an output variable that is itself the product of human decision-making (and therefore human error and bias), the algorithm will, by design, become very good not at forecasting the *key variable* (the unobservable characteristic that really matters) but an output that reflects historical human estimates of that key variable. This is one way that biases can be baked into an algorithmic process.

---

[119] *Id.* at 683.

[120] Mayson, *supra* note 5, at 2262.

[121] *See* Harini Suresh & John Guttag, *A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle*, ASS'N FOR COMPUTING MACH.: EQUITY AND ACCESS IN ALGORITHMS, MECHANISMS, AND OPTIMIZATION (Oct. 2021), https://dl.acm.org/doi/fullHtml/10.1145/3465416.3483305 [https://perma.cc/E6JF-P8RD].

[122] Burk, *supra* note 6, at 1160 ("Algorithmic pattern detection and scoring outputs are not found, they are actually constructed by the processes of data harvesting, ingestion, and analysis").

Many of the examples cited in critical works, such as Cathy O'Neil's *Weapons of Math Destruction*, illustrate this problem insofar that it pertains to the machine learning context. When a university trains an algorithm to predict which admissions applicants are most likely to be selected based on past data regarding their incoming grades and the admission office's decision, the machines will replicate whatever biases were held by the human admissions teams that preceded it.[123] When recidivism risk scoring algorithms used in the criminal justice system use subsequent arrest as the outcome to be predicted, the myriad social factors that affect when and where law enforcement make arrests (including the substantial impact of racial prejudice and bias) will be imbedded in the model.[124] And if an algorithm uses historical home prices or search queries to estimate a home value or a query completion (respectively), the algorithm will make its predictions based on how humans *have* valued houses or used a search engine, and not necessarily based on how they *will* or *could* or *should*.[125]

Companies like Zillow and Google can alter their objective function to predict the outcomes that better match collective social goals (and indeed, both companies have done so.[126]) In Zillow's case, recent external research shows that Zillow's "Zestimate" could make more "accurate" predictions of ultimate selling prices by taking into account the racial makeup of a neighborhood, but the company (presumably intentionally) has excluded that factor.[127] In other words, Zillow

[123] Lilah Burke, *The Death and Life of an Admissions Algorithm*, INSIDE HIGHER ED (Dec. 14, 2020), https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd [https://perma.cc/5RN3-KPKG].

[124] *See* Megan T. Stevenson & Christopher Slobogin, *Algorithmic Risk Assessments and the Double-Edged Sword of Youth*, 96 WASH. U. L. REV. 681, 694 (2018).

[125] *See, e.g.*, SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018) (describing how past and present discriminatory behavior are reflected or exacerbated in Google search results).

[126] In Google's case, the company has programmed the autofill generator to avoid showing biased or disparaging results for individuals or certain demographic groups. Issie Lapowsky, *Google Autocomplete Still Makes Vile Suggestions*, WIRED (Feb. 12, 2018, 11:09 AM), https://www.wired.com/story/google-autocomplete-vile-suggestions/ [https://perma.cc/K2DG-TGJU]; Jason Slotkin, *Google Will Block Its Autocomplete Suggestions for Some Election-Related Searches*, NPR (Sept. 11, 2020, 2:34 PM), https://www.npr.org/2020/09/11/911915056/google-will-block-its-autocomplete-suggestions-for-some-election-related-searche [https://perma.cc/7NGR-F74G].

[127] Shuyi Yu, Algorithmic Outputs as Information Source: The Effect of Predictive Algorithms on Home Prices and Racial Biases in the Housing Market 30 (Apr. 12, 2020) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3584896 [https://perma.cc/G27P-KTSN].

estimates not what the house is most likely to sell for, but what the house would be most likely to sell for if the market were indifferent to the racial makeup of its neighborhood. The latter is arguably closer to the unobservable key variable (what a home is really "worth" in some platonic sense) than the actual short-term selling price of the home. Indeed, the same researchers found that Zillow's "Zestimates" had the effect, over time, of *changing* the prices at which homes were sold so that short-term selling prices have started to converge with race-neutral assessments of home value.[128]

But there will always be cases where the available data inevitably leaves the imprint of past human decision-making. In the case of recidivism risk scores, a scoring algorithm would ideally have access not to which arrestees are subsequently re-arrested for some new crime, but which arrestees actually commit a new crime. Yet there is no source for this unbiased output measure. Outside a few narrow contexts where non-law-enforcement surveillance is prevalent, arrests or convictions will be the best measures of an individual's crime commission,[129] imperfect as they are. But this does not relieve the data steward of responsibility. To the contrary, a data steward who knows that their outcome measure is a noisy, human-influenced approximation of the key variable of interest should understand that they must manage this flaw.

With these intuitions in mind, let us return to the SCMBD dynamics that might result from a biased objective function. Consider, for example, the potential SCMBD created by reliance on the variable "Age at First Arrest" in recidivism risk scoring. This variable is often given significant weight in risk scoring[130], and could conceivably contain discontinuities where, for example, arrestees who were first arrested as teenagers are predicted to be much more likely to be re-arrested than arrestees whose first arrest occurred in their early twenties. In other words, a difference in merely one year (or less) in the age at first arrest would lead to a substantial difference in the outcome for the arrested individual. A recidivism risk score that contains this SCMBD has a potential problem of uneven detection or enforcement in two places — the input (small differences in age of first arrest) *and* the output (big differences in the probability of re-arrest) in any region where residents who committed a crime were more likely to be arrested both as

---

[128] *Id.*

[129] *Measuring Recidivism*, NAT'L INST. OF JUST. (Feb. 20, 2008), https://nij.ojp.gov/topics/articles/measuring-recidivism [https://perma.cc/MF73-6LD4].

[130] Cynthia Rudin, Caroline Wang & Beau Coker, *The Age of Secrecy and Unfairness in Recidivism Prediction*, 2.1 HARV. DATA SCI. REV., Mar. 31, 2020, at 27, https://hdsr.mitpress.mit.edu/pub/7z10o269/release/4 [https://perma.cc/B2RD-NY5N].

teenagers and today. Since heavily patrolled regions in U.S. cities are disproportionately minority[131], the lopsidedness in both inputs and outputs will have a racial bias.

If the criminal justice system were able to magically fix the output variable so that we could measure the true outcome of interest — whether a person actually commits another crime — and not just whether the person was arrested for one, then the weight assigned to age at first arrest would be reduced. That factor would surely still have some value, but it would not correlate as strongly with subsequent criminal behavior as it does to subsequent arrest.

Age at first arrest may also be less accurate and more error-prone for black residents who live in neighborhoods or attend schools that have greater police presence and juvenile law enforcement.[132] We discuss the problems with biased error next. But bias in the errors won't even be detectable if both the output and input variables are similarly biased. Hence, discovery of a SCMBD relating to recidivism risk should put users on alert that the gap between observable outputs used to train an algorithm and the true outcome of interest may reveal little more than the excessive weight that humans had previously placed on some factor, and this can have racially salient implications.

b. Biased Errors

The next and most conspicuous form of bias comes from unequal error across subgroups. For example, a Propublica critique of COMPAS recidivism risk scores concluded the scores were biased because black arrestees who did *not* reoffend were assigned high risk scores that would have led to pre-trial detention almost twice as often as white arrestees who did not reoffend.[133] This inequality in false-positive rates is one form of bias. Inequality in false-negative rates can also cause bias. In the case of COMPAS scores, the same study found that white arrestees who

131 *See* Andrew Gelman, Jeffrey Fagan & Alex Kiss, *An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias*, 102 J. AM. STAT. ASSOC. 813, 814 (2007).

132 *See* Rashida Richardson, Jason M. Schultz & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 15, 18, 20-21 (2019) (explaining how over-policing and police misconduct directed at minority communities could bias predictions of criminal activity).

133 Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/FU46-YNQH].

re-offended had the benefits that come with wrongly being labeled low-risk more often than black arrestees (again by a factor of two).[134]

SCMBDs can help identify and isolate some of the causes of biased error. Consider, for example, novel credit scoring algorithms that use detailed online and purchasing behavior to establish credit scores. Suppose a data steward discovered a SCMBD related to purchases from particular types of stores (like pawn shops) or particular types of behaviors (like scrolling too fast through terms and conditions).[135] If these input variables are valid but less so for minority or female loan applicants, heavy reliance on them will cause a disproportionate amount of error to fall on those groups. These problems would not emerge if the algorithm had access to race and gender or a close proxy because machine learning would self-correct and give a different weight to the SCMBD inputs for those groups. But in the great many contexts where race, sex, or close proxies (like zip code) are removed specifically for the purpose of eliminating discriminatory treatment, the disparate impact of the variables that are left cannot be controlled.[136] For that reason, Sandy Mayson has argued that removal of race and close proxies from decision-making algorithms is a mistake, even if the motivations are good.[137] But introducing race and gender presents its own risks of misuse or public distrust. Removing SCMBDs that have a disparate impact might be a better option because their removal could alleviate some of the problems of biased error without the complexity that comes from allowing race or gender to be used by a model.

c. Bias Without Error (and Satisfying Explanation)

Finally, we might be concerned about disparate impacts across groups irrespective of error. Indeed, some computer scientists have recommended programming machine learning algorithms to detect and automatically correct for disparate impacts even if the model is less

[134] For other types of scoring and treatment decisions that result in variable rather than binary responses (e.g., errors in the predictions of grades or test scores), the analogs to "false positive" and "false negative" error amounts to positive or negative deviations between the predicted value and the true value. Again, as with binary outcomes, bias in errors can be measured using only positive error, only negative error, or some weighted combination of both.

[135] Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 164 (2016).

[136] *See* Mayson, *supra* note 5, at 2243-48.

[137] *See id.* at 2262.

accurate as a result.[138] They are inspired by the EEOC's "four-fifths" rule that treats deviations in outcomes between groups as presumptive indicators of discriminatory disparate impact.[139] But this is controversial because automatically correcting for disparate impacts in all circumstances would interfere with other forms of fairness. In addition to making the decision model more inaccurate, it can generate group differences in error rates, pitting one form of bias against another.[140] There is more to unfairness than lack of statistical parity in the outcomes of protected groups.[141]

But automatic correction of disparate impacts makes more sense and is less controversial when a SCMBD is the source of disparity. After all, the premise of a SCMBD is that some factor has an inexplicably large role in the outcome, so there cannot be a rationale for the disparate impact it causes (or at least finding such a rationale would be challenging). That is, to use the legal terms of art, there is no articulable business necessity (in the case of Title VII litigation) or individualize suspicion (in the case of police searches or seizures).

For instance, consider the example of floor-protecting pads and credit scoring that we discussed earlier as a potential SCMBD. It could very well be that the purchase of floor-protecting pads has a racial dimension — that their use is part of an established custom for some groups and not others and might not be stocked or prominently displayed in stores that serve other minority communities. If so, placing heavy weight on the purchase of furniture pads would cause a disparate impact to minorities. If the furniture pads had only a small impact on credit scoring, there may not be a great problem. After all, the effect of that

---

[138] Dwork et al., *supra* note 14, at 214 (offering a protocol that ensures the demographics of the set of individuals receiving a classification are the same as the demographics of the underlying population (within a preset margin) while treating similar individuals as similarly as possible. The authors call this approach "fair affirmative action"). *See generally* Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Krisitina Lerman & Aram Galstyan, *A Survey on Bias and Fairness in Machine Learning*, 54 ACM COMP. SURVS., July 2021, at 11-25, https://dl.acm.org/doi/pdf/10.1145/3457607 [https://perma.cc/24FS-46PM].

[139] 29 C.F.R. § 1607.4 (2021).

[140] *See* KEARNS & ROTH, *supra* note 116, at 84-89 (discussing how corrections generate conflicts between different definitions of fairness as well as the challenge of achieving fairness between subgroups and offering some solutions).

[141] For a skeptical take on automatic removal of disparate impacts, see (and listen to) Stewart Baker, *Ethical Algorithms*, VOLOKH CONSPIRACY (Dec. 5, 2019, 6:09 PM), https://reason.com/2019/12/05/ethical-algorithms/ [https://perma.cc/ZFF3-E668] ("I have long suspected that much of the fuss over bias in machine learning is a way of smuggling racial and gender quotas and other academic social values into the algorithmic outputs.").

one factor would be small, and one can come up with a rationalization for including the factor. As we suggested in Part I,[142] floor-protecting pads plausibly correlate with some amount of caution and care of the purchasers. But when the purchase of the pads has a large impact on outcomes, the disparate impact is more troubling. The emphasis on this purchase, despite the algorithm's access to a large number of other purchases and factors that should correlate with caution and care, has no satisfying explanation. The elimination of the SCMBD therefore improves equality at the cost not of accuracy, but of *inexplicable* accuracy. When a disparate impact is concerned, a model's lack of interpretability or explainability should have extra salience.[143]

Of course, SCMBD dynamics in a machine learning algorithm will not necessarily exacerbate race and gender gaps. In fact, they could go the other way by adding weight to features or combinations of features that accurately identify minority data subjects for better scores. This is one reason that our policy recommendations in Part IV advocate SCMBD audits and additional testing rather than automatic correction or removal.

3. SCMBD and Strategic Behavior

So far, we have discussed algorithmic decision-making as though the subjects of the decision-making process will passively accept the ranking and outputs without changing the inputs through behavioral modification. Of course, it is most likely that many people *will* alter their behavior in response to the existence and nature of a decision-making algorithm.[144] Some responses will correspond to real, internalized differences in behavior, but others will be a form of "gaming" — that is, superficial changes in behavior in order to improve short-term outcomes.[145]

---

[142] *See supra* note 47 and related text.

[143] That is, there is no articulable business justification for an inexplicable relationship between the variables.

[144] FINN BRUNTON & HELEN NISSENBAUM, OBFUSCATION: A USER'S GUIDE FOR PRIVACY AND PROTEST 8-41 (MIT Press 2015) (where the authors provide a long list of instances in which individuals engage in active obfuscation measures to interfere with surveillance efforts); Gary T. Marx, *A Tack in the Shoe: Neutralizing and Resisting the New Surveillance*, 59 J. SOC. ISSUES 369, 374-84 (2003).

[145] Jane Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1, 25 (2018).

In the algorithmic context, "adversarial machine learning" ("AML") has emerged as a field of research and practice, exploring the ability to engage in such potential intentional manipulations.

SCMBD dynamics leave an algorithm vulnerable to strategic behavior. As long as the input variable is under the control of the data subject, SCMBDs mark parts of the model where a little bit of effort by the data subject can yield a big payoff in terms of treatment. For example, if furniture pads are used to indicate better credit risk, those with access to the right sort of information will purchase the pads prior to applying for a loan even if they have no interest in using them. SCMBD therefore enables gaming, and gaming often leads to other forms of unfairness. First, gaming will degrade the accuracy of the model as more and more people engage in strategic behavior, so whatever predictive validity a SCMBD may have at T_0 can be lost before the algorithm is corrected in T_1 (and as we explained above,[146] inaccuracy in this context quickly transforms to unfairness as well). Gaming can also cause inequalities because those with access to information about the algorithm and those with an appetite to engage in gaming are not a random subset of the population.[147] A gameable decision-making system is therefore inefficient and inequitable.

The gameability of a system causes other types of waste and psychic costs as well. If a system *can* be manipulated by changing behaviors, individuals will feel some amount of pressure to constantly monitor their choices and behaviors to optimize how they will be treated by an algorithm downstream. The anxiety, self-censorship, and timidity often attributed to surveillance would, arguably, become acute when an individual can constantly change their conduct without too much cost at any given moment, especially when such changes might have substantial implications. Thus, to the extent SCMBDs increase gameability (and the implicit burden to self-monitor and exploit opportunities for gaming), they deserve suspicion.

The connection between SCMBD and gaming may seem at first to be a trivial aspect of decisional fairness, but it has a lot in common with critiques of laws that create harsh regulatory cut-offs. Lee Fennel demonstrates how the use of categories can incentivize regulated actors to manipulate their treatment by swapping into a different group.[148] Adam Kolber has described gaming of state real estate tax laws that set higher rates for million-dollar mansions. (Unsurprisingly, there are an abundance of transactions in those states involving homes sold at *just*

[146] *See* discussion *supra* Part II.A.1.

[147] See discussion in Bambauer & Zarsky, *supra* note 145, at 11-12, 29-32.

[148] *See* Lee Anne Fennell, *Sizing Up Categories*, 22 THEORETICAL INQUIRIES LAW 1, 3 (2021) (noting that this is also merely a factor to be balanced against the disadvantages of other regulatory rules).

under $1 million.)[149] Accountants routinely advise clients to plan their finances so that they fall within specific tax brackets or so that they can take advantage of certain tax incentives, and they help calibrate their clients' actions at the end of the fiscal year if they are getting too close to the boundaries of one category or another.[150] Of course, these phenomena are not limited to legal cutoffs, either. Boxers aim for the highest weight possible in their class, flirting with disqualification during the final pre-fight weigh-in. To avoid accidentally sliding into the next weight category, they work out intensively in a steam room prior to weighing in order to lose water weight.[151] In all these contexts, unfairness and inefficiency follow. These are acceptable parts of a process or scheme that would be too difficult to administer without categories and cut-off rules, but big data algorithms do not have identical problems of practical administrability. Thus, if gameable SCMBDs can be avoided, they should.

To summarize, SCMBD dynamics indicate potential problems within existing conceptions of algorithmic fairness by undermining accuracy, increasing inequity, or by creating a toehold for gaming. SCMBD also implicates some other forms of fairness that are unique to it. We discuss these next.

B. SCMBD as a Distinct Form of Unfairness

Our natural aversion to SCMBD dynamics may be purely instrumental. That is, it may stem from the suspicion that the SCMBD is a biproduct or traveling companion of one of the forms of algorithmic unfairness that is already well-recognized in the field. But another possibility, which we explore here, is that SCMBD is a manifestation and direct proof of some other form of unfairness that has not been sufficiently developed in the Fairness, Accountability, Transparency, and Ethics literature. One candidate is an ethic of proportionality that is broken by SCMBDs, and another is a preference for parsimony and chance over hyperselectivity. We explain each here.

---

149 Kolber, *supra* note 54, at 684.

150 *See, e.g.*, *5 Ways to Avoid Bumping Your Income into a Higher Tax Bracket*, TAXACT BLOG, https://blog.taxact.com/avoid-higher-tax-bracket/ [https://perma.cc/R32M-TNTP] (addressing advice as to how a higher tax bracket could be avoided).

151 *See, e.g.*, *'Not Far from Death': How Fighters Are Risking Their Lives to Make Weight*, BBC (Mar. 14, 2017), https://www.bbc.co.uk/bbcthree/article/a4372439-aef4-4f18-bd9a-d914e3f37f2a [https://perma.cc/D4WR-SWD8] (detailing the fighters' (at times dangerous) practice of striving to refrain from moving into a higher weight category).

1. Disproportionality

Proportionality is a stable fixture in law and collective values. A serious crime should receive a harsher punishment than a minor crime for both deterrence reasons and retributive ones.[152] Liability and damage awards should be modulated by the level of care for the same reasons. Therefore, cliffs and harsh cutoffs in treatment created by SCMBDs violate a collective sense of proportionality.[153]

To understand whether an aversion to SCMBDs is well-justified on proportionality grounds, it will be necessary to understand the fundamental values that proportionality serve. Let us start with a consequentialist function. Utilitarian goals like deterrence work by ensuring that there is greater incentive for people to avoid doing worse things.[154] But SCMBDs (at least, the ones that are interesting) really *do* have a statistically valid relationship with outcomes. If we are mostly interested in improving outcomes and occasionally encouraging changes in behavior, we might want to retain SCMBDs in a model even if we don't understand the relationship between inputs and outputs. Thus, when we raise potential problems with SCMBDs and proportionality, we are really homing in on deontological principles of equality and desert.

First, consider equality. In its simplest form, it requires that similar individuals are treated similarly.[155] By definition, none of the individuals treated differently as a result of SCMBDs are similar, though. There is a small but meaningful difference between them. Equality, however, goes well beyond the simple notion of the unequal treatment of equals. Philosophers as far back as Aristotle have discussed the proportionality aspects of equality.[156] "Proportional Equality" calls for treating all the individuals in proportion to relevant characteristics.[157] It requires scalar proportionality that increases outcomes in lock step

[152] *See* Solem v. Helm, 463 U.S. 277, 299 (1983).

[153] Kolber, *supra* note 54, at 674.

[154] *See* Gary S. Becker, *Crime and Punishment: An Economic Approach*, *in* ESSAYS IN THE ECONOMICS OF CRIME AND PUNISHMENT 1, 9-14 (Gary S. Becker & William M. Landes eds., 1974); MICHAEL CAVADINO, JAMES DIGNAN, GEORGE MAIR & JAMIE BENNETT, THE PENAL SYSTEM: AN INTRODUCTION 37 (6th ed. 2020).

[155] ARISTOTLE, NICOMACHEAN ETHICS 39 (G.P. Goold ed., H. Rackham trans., Harvard Univ. Press rev. ed. 1934) (c. 384 B.C.E.).

[156] Stefan Gosepath, *Equality*, STAN. ENCYCLOPEDIA PHIL. (Apr. 26, 2021), https://plato.stanford.edu/entries/equality/#ProEqu [https://perma.cc/T6XX-RKUV]; *see also* Brian M. McCall, *Demystifying Unconscionability: A Historical and Empirical Analysis*, 65 VILL. L. REV. 773, 778 (2020).

[157] LAW AND MORALITY: READINGS IN LEGAL PHILOSOPHY 736 (David Dyzenhaus, Sophia Reibetanz Moreau & Arthur Ripstein eds., 3d ed. 2007).

with the relevance of each factor.[158] Scalar proportionality is violated by SCMBDs when the *relevance* of a small change is, or should be considered, small (which by supposition, it is).[159]

As with disparate impacts, the disproportionality problem with SCMBD has a close connection to the opaqueness of big data algorithms. If an input needs to have an understandable relationship to a prediction in order to satisfy the test for proportionality, SCMBDs will fail every time. The disproportionality problem of SCMBDs can help ethicists and lawmakers understand how to set practical limits on mandates for algorithm transparency by defining an area where transparency gets the biggest bang for the buck. Even if demands for explainable algorithms are generally unwise, they may be valuable when an algorithm is using a SCMBD in its modeling.

This form of disproportionality is compatible with the Equal Protection clause as it relates to individuals' enjoyment of fundamental rights. The Supreme Court's decisions on voter qualifications are illustrative. In *Kramer v. Union Free School District No. 15*, for example, the Court held that states could not disqualify residents from school board elections on the basis of not owning property or not having children at home.[160] In *Dunn v. Blumstein*, the Court rejected long residency requirements for voter eligibility.[161] The crux of both cases was that a state was burdening a fundamental right (i.e., imposing a big difference) and could not adequately justify why it distinguished among potential voters based on seemingly trivial or irrelevant distinctions.[162] Constitutional due process safeguards protect individuals from irrational or illogical treatment even with respect to privileges that are less sacrosanct than fundamental rights.[163] Some rationale must be supplied. As Richard Fallon explains, this idea "is captured by perhaps the most persistently recurring theme in due process cases: government

---

158 We are adopting the "scalar" modifier used in "scalar consequentialism." *See* Alastair Norcross, *The Scalar Approach to Utilitarianism*, *in* THE BLACKWELL GUIDE TO MILL'S *UTILITARIANISM* 217(Henry R. West ed., 2006).

159 For a review of this discussion in the work of Thomas Aquinas, see BENJAMIN JOHNSON & RICHARD JORDAN, WHY SHOULD LIKE CASES BE DECIDED ALIKE? A FORMAL MODEL OF ARISTOTELIAN JUSTICE 27-28 (2017), https://scholar.princeton.edu/sites/default/files/benjohnson/files/like_cases.pdf [https://perma.cc/8VBV-H9JM].

160 Kramer v. Union Free Sch. Dist. No. 15, 395 U.S. 621, 632 (1969).

161 Dunn v. Blumenstein, 405 U.S. 330, 360 (1972).

162 *See id.*; *Kramer*, 395 U.S. at 632.

163 *See, e.g.*, Armour v. City of Indianapolis, 566 U.S. 673, 680 (2012) (discussing how tax measures that make distinctions among tax-payers will trigger rational basis review, requiring some sort of rationale).

must not be arbitrary,"[164] and SCMBD provides at least some initial proof that arbitrariness is in play.

Two important caveats are in order before SCMBDs are declared unfairly disproportional (or at least thought to be so unless found otherwise). First, SCMBDs are proportional in the sense of ensuring that individuals with predictively better outcomes are treated better and individuals with predictively worse outcomes are treated worse. Arguably, SCMBDs provides a sort of proportionality, just not an even, linear proportionality. This begs the question whether proportionality requires a linear relationship between input values and outcomes.[165] Aristotle himself left few clues about his beliefs on the topic, but legal analyses of criminal punishment do not seem to require it. Interpretations of the Eighth Amendment restriction on cruel and unusual punishments and the Equal Protection Clause, for example, seem to be satisfied as long as less-bad behavior receives less-bad punishment (of any degree and any sort).[166] American jurisprudence might not be the best source for enduring ethics, but as the examples noted throughout this Article demonstrate, many highly consequential decisions related to hiring, tax, public health, and resource allocation also tolerate discontinuities in treatment in order to promote other goals such as administrative efficiency. Some of these factors (like administrative costs) have little relevance to machine learning algorithms that can be constrained to create strictly linear relationships without significant expense, but the larger point still stands: the expectation of linearity in proportionality is an altogether different requirement that demands additional moral justification (although some of the justifications provided above might seem to advocate a strong preference for linearity).[167]

---

[164] Richard H. Fallon, Jr., *Some Confusions About Due Process, Judicial Review, and Constitutional Remedies*, 93 COLUM. L. REV. 309, 322-23 (1993).

[165] Some scholars do believe Aristotle believed a nonlinear function describing the relation between inputs and outputs would prove problematic. *See* Johnson & Jordan, *supra* note 159, at 28. Note, however, that even linear functions might be very steep and generate a sense of unfairness. More importantly, it is unclear what Aristotle's position would be regarding such a scenario.

[166] *See* Jane Bambauer & Andrea Roth, *From Damage Caps to Decarceration: Extending Tort Law Safeguards to Criminal Sentencing*, 101 B.U. L. REV. 1667, 1698-702 (2021).

[167] *See* Re'em Segev, *Continuity in Morality and Law*, 22 THEORETICAL INQUIRIES LAW 45, 52-68 (2021) (explaining that structuring an argument that justice, fairness and equality require proportionality must be premised on the recognition of "scalar consequentialism" which he defines and explains).

A second caveat is that SCMBDs only fail *apparent* proportionality tests. They seem disproportional only because they currently leverage insights or patterns which, if replicable and nonspurious, are probably real but beyond human explanation. Yet if a pattern is stable and can reliably predict an important outcome, the failure to use it to achieve various social goals may be the greater moral failing. This argument shows greatest strength when incorporating a SCMBD might provide great social benefits, and limited unfairness-based concerns. To take an example, we are wise to exploit patterns that link certain biological or genetic markers to health problems even if we don't fully understand them. If life-preserving therapies are held up until the theoretical work catches up out of deference to proportionality, the costs in life and health will be severe.[168]

More fundamentally, perhaps a conception of proportionality that depends on human notions of relevance engages in a naturalism fallacy — a wrong and counterproductive assumption that human knowledge is superior to machine knowledge purely by virtue of being human. Aziz Huq has documented the ways in which machine sense-making diverges from human thought, and has warned against giving human systems too much credit in the process.[169] Arguably, as long as we are confident that a SCMBD is enduring and nonspurious, proportionality might require *humans* to recalibrate how relevant a factor is rather than requiring machines to do so. That said, as the field of descriptive ethics attests, we ignore at our peril broadly- and strongly-held instincts about fairness, even if those instincts are on shaky ground theoretically.[170]

---

[168] Indeed, findings regarding the benefits of hygiene in hospitals have preceded a full understanding of the possible transmission of illness through contact. For one popular reference to this issue, see Rebecca Davis, *The Doctor Who Championed Hand-Washing and Briefly Saved Lives*, NPR (Jan. 12, 2015, 3:22 AM ET), https://www.npr.org/sections/health-shots/2015/01/12/375663920/the-doctor-who-championed-hand-washing-and-saved-women-s-lives [https://perma.cc/D6G4-HRH]. Today, some diets like plant-based or Mediterranean diets seem to confer anti-cancer properties and can be used as part of a treatment plan even though the reasons for the influence are unclear. *See* Esther Molina-Montes, Elena Salamanca-Fernández, Belén Garcia-Villanova & Maria José Sánchez, *The Impact of Plant-Based Dietary Patterns on Cancer-Related Outcomes: A Rapid Review and Meta-Analysis*, NUTRIENTS, July 6, 2020, at 1, 23.

[169] Huq, *supra* note 7, at 1898-99; (comparing racial bias in human systems versus algorithms); Aziz Huq, *A Right to a Human Decision,* 106 VA. L. REV. 611, 640-46 (describing opacity of human minds versus machine learning processes).

[170] NORA HÄMÄLÄINEN, DESCRIPTIVE ETHICS: WHAT DOES MORAL PHILOSOPHY KNOW ABOUT MORALITY? 4 (1st ed. 2016).

Beyond equality, a proportionality-based aversion to SCMBD taps into the notion of desert.[171] The desert principle justifies an individual's entitlement to an output based on his or her actions or character.[172] If an individual, for instance, worked many additional hours or studied several years for an advanced degree, she *deserves* to receive higher pay. If an individual worked to protect her health, he might be entitled to a better rate in his health insurance. Conversely, retributional theory would indicate that greater and more severe crimes merit a proportionally more extensive punishment.[173]

A desert-based argument may be more limited than the equality one because it would seem to be inapplicable when the inputs (the "small changes") concern a factor outside of the subject's control. Small differences in age or environmental factors do not implicate desert since growing older usually does not indicate a substantial achievement, and nobody controls their environment. Therefore, it is difficult to argue that a person has earned better credit or insurance rates or even a shorter prison sentence based on wholly external factors (such as whether one lived in a warmer or cooler climate zone).[174]

But within the set of factors that *are* within the subject's control, desert can play an instructive role. When one spouse is offered a much lower credit limit because she has spent a little bit less (or holds fewer credit cards), she can cry foul based on proportionality of desert. She may argue that her payment record is the same as her spouse's, so she doesn't deserve much worse credit terms. Or consider two employees whose salaries are over twenty percent different based entirely on a five percent difference in work productivity such as billable hours.[175] If this is a SCMBD situation (it may not be)[176] the lack of balance between work and pay violates general notions of desert and suggests that their pay needs to be better calibrated.[177] Unlike the proportional equality

[171] *See* John Kleinig, *The Concept of Desert*, 8 AM. PHIL. Q. 71, 71-74, 76 (1971).

[172] *See* Ronen Perry & Tal Z. Zarsky, *Queues in Law*, 99 IOWA L. REV. 1595, 1614 (2014).

[173] Kolber, *supra* note 54, at 670 (referring to the work of Husak, who derives the principle of such proportionality from the broader notion of desert).

[174] *See* Kleinig, *supra* note 171, at 74.

[175] Under Locke's theory, entitlement comes from labor, which is why this hypothetical provides the clearest fit. *See* Perry & Zarsky, *supra* note 172, at 1617.

[176] Much would depend on where the two appear on the distribution of billable hours; if they are close to one end or the other of a normal distribution, it could be inferred that an additional five percent of working time is extremely difficult and personally costly for those who do it, thereby suggesting the difference is not "small" or justifying the difference in pay from a desert perspective. *See supra* Part I.B.

[177] *See* Perry & Zarsky, *supra* note 172, at 1617.

considerations discussed above, discussions of desert in moral philosophy expect and even require linearity (or "geometric proportionality" as it is described by Aristotle).[178] In other words, the fact that Al worked five percent more hours than Bob means that Al *deserves* compensation that is five percent more than Bob — no less, no more.

The "desert"-based justification also has shortcomings, even beyond the restriction in scope to inputs that are under the effort or control of the subject. Outside of criminal justice (where the negative reflection of desert — retribution — plays a large role), desert enjoys only very limited application in the law. It is typically embraced only when it coincides with utilitarian theories about optimal incentive structures.[179] Some scholars (including John Rawls) reject its use in theories of distributive justice.[180] And other commentators limit desert arguments to instances in which resources are plentiful rather than scarce.[181]

Nevertheless, all the problems with proportionality, whether of the equality or desert variety, can be managed by recognizing proportionality as a qualified interest — one that should be balanced against other fairness values like accuracy and nondiscrimination.

2. Hyperselectivity

A second reason to treat SCMBD as a problem in its own right concerns an aversion to processes (whether mechanical or human-driven) that attempt to judge people with too fine a measure. We may justifiably prefer a system that uses a few important factors to sort people into different treatments but otherwise treat everyone the same. To explain this logic, it's worth looking afresh at the decision to selectively discriminate among individuals in the first place.[182]

Decisions about how to allocate scarce resources or penalties can be made in only two ways: by pooling potential recipients and distributing the resource using a neutral factor such as queues or lotteries, or by

---

[178] Peter Celello, *Desert*, INTERNET ENCYCLOPEDIA OF PHIL., https://iep.utm.edu/desert/#SSH1cii (last visited Dec. 29, 2021) [https://perma.cc/3GBD-HHGR].

[179] For a discussion of its application in copyright law, see Lior Zemer, *The Making of a New Copyright Lockean*, 29 HARV. J.L. & PUB. POL'Y 891, 891-947 (2006).

[180] Celello, *supra* note 178 (referring to the work of Rawls).

[181] *Id.*

[182] Throughout this Section, we use the term "discriminate" to mean separating people and treating them differently on *any* basis. We do not use the term to mean altering treatment on the basis of race, sex, or some other protected status.

discriminating between them using one or more factors.[183] The diversity visa lottery, for example, is a pooling system (at least with respect to immigrants from one particular country) because it awards visas by randomly selecting a set number of visa applicants from a particular country.[184] The Alaska Permanent Fund is another example because it evenly distributes the pool of dividends from the state's extraction of natural resources across all eligible Alaska residents.[185] Pooling schemes are designed to treat all subjects in the pool the same without assessing the merits or risks associated with any person in the pool.[186]

A discriminating system, by contrast, attempts to reduce the role of random chance and flat distributions by allocating resources according to some key characteristic — merit, need, moral turpitude (in the case of criminal punishments), or some other motivating factor.[187] When a decision-maker is discriminating between subjects to allocate resources based on a key characteristic, a decisionmaker *must* use an algorithm — some set of rules to weight proxies that the decisionmaker believes to be well-correlated with the key characteristic. This is so whether the decision-maker is a machine or a human.

A preliminary analysis that often gets lost in the discussion of AI is whether a resource really should be distributed on a discriminating basis instead of pooling the resource and treating every prospective recipient exactly the same. To be sure, the answer will often be, "Yes, we need to discriminate." Pooling across the full range of heterogeneous individuals will often distort or completely undermine the goal of the program. For example, social services or economic stimulus is often most valuable when it is targeted to those who need it. Hiring for a high-skill position will require the employer to assess which candidates are well-qualified and will help the firm the most. In these cases, resource allocation calls for discriminating between subjects in a system that requires merit, need, or some other key factor to drive the allocation of the resource. But sometimes, explicitly asking whether the resource can be pooled or shared randomly can reveal an opportunity to dispense

183 Bambauer & Zarsky, *supra* note 145, at 6-8; *see* James C. Cooper, *Separation Anxiety*, 21 VA. J.L. & TECH. 1, 3-4 (2017) (discussing the tension between separation and pooling in the privacy context).

184 *Diversity Visa Program: Selection of Applicants*, U.S. DEP'T OF STATE — BUREAU OF CONSULAR AFFS., https://travel.state.gov/content/travel/en/us-visas/immigrate/diversity-visa-program-entry/diversity-visa-submit-entry1/diversity-visa-selection-of-applicants.html (last visited Dec. 29, 2021) [https://perma.cc/8TGD-9E8L].

185 *History*, ALASKA PERMANENT FUND CORP., https://apfc.org/fund-education/ (last visited Dec. 29, 2021) [https://perma.cc/XS5A-35GP].

186 Cooper, *supra* note 183, at 3.

187 *See* Perry & Zarsky, *supra* note 172, at 1621.

with selection systems that are not worth the administration and potential resentment that they cause.

Even within discriminating systems, additional questions can be asked: once a set of important factors is used to separate and discriminate between individuals, is it worth it to add additional factors to adjust and fine-tune how each individual is treated? Once a core set of factors helps divide individuals into mostly-homogenous subclasses, is it worth the effort and the risks of resentment, opaqueness, etc. to *further* differentiate between the individuals? Or would the mission of the program be better served by a satisficing rule that switches to pooling within the mostly-homogenous subclasses? Even though we know *ex ante* that the subclass could be further separated and rearranged into still more accurate strata, a mixed approach of using some important variables to separate and then switching to pooling may better serve the decision-maker and subject alike.

This inquiry relates to the concept of parsimony, and SCMBD is a direct indicator of its lack. Parsimony is the quality of preferring a simpler model or positive theory over a more complex one so long as it performs about as well in its predictions.[188] In the philosophy of science, parsimony is valued in part because it is more testable, with fewer variables that theoretically explain real world outcomes.[189] Since it is more testable, that means it is more easily falsifiable as well, and thus more amenable to correction if it is later proven wrong. Thus, parsimony is preferred in part to avoid overfitting and spurious models. Parsimony is desirable for the purposes of explainability and related theories of desert, too. Parsimony is also valuable in dynamic processes where we actually want a decision-making process to induce individuals to change their behavior. A system that uses factors related to education, effort, or avoidance of criminal behavior will help motivate individuals to become better educated, put in more effort, and refrain from criminal conduct. But if the decision-making system also uses one thousand other variables, some of which have confusing relationships to the ultimate treatment of the individuals, the salutary incentives will be diminished.[190]

While explainability and behavioral incentives might not be overriding concerns for a particular decision as compared to accuracy

[188] Parsimony is synonymous with Occam's razor. *See Simplicity*, STAN. ENCYCLOPEDIA OF PHIL. (Dec. 20, 2016), https://plato.stanford.edu/entries/simplicity/#QuaPar [https://perma.cc/M5EE-3UTW].

[189] *Id.*

[190] *See* Annie Liang & Erik Madsen, Data and Incentives, Proceedings of the 21st ACM Conf. on Economics and Computation 41 (July 13–17, 2020).

and other fairness considerations, its purpose and value can increase with the increased complexity of an algorithm. At the margin, it may not be worth the modest gains in accuracy to introduce a SCMBD. In other words, even if a SCMBD factor helps improve accuracy a little bit, it might not be worth it to eke out that small improvement in the short-run if the added complexity causes problems in trust and motivation. Much like theories of harm from surveillance or from the repurposing of data initially collected for a different purpose, an algorithm that uses a SCMBD can make people more cautious and guarded in their actions because they never know when some small decision (e.g., to get a massage) will cause a large change in their status (e.g., lost access to credit). Even if it is socially desirable to harness the power of big data algorithms to look for correlations from lots of different inputs, society may be better off settling for a system that is selective but not *hyperselective,* on any given input.[191]

This hypersensitivity claim is distinctively different than the other "fairness"-based arguments stated thus far. While the former are normative claims premised on analytical arguments, the latter refers to "positive fairness" — a public perception of unfairness that might further lead to additional unwanted outcomes (such as caution and discomfort).[192] Conduct that runs against positive fairness might generate inefficiencies (such as those noted above — caution and discomfort) and might also serve as a proxy for normative unfairness given its reflection of public preferences.[193] However, positive fairness claims are best premised on behavioral studies — studies we hope to contribute in the future.

Thus, there are both intrinsic and instrumental reasons to hold SCMBDs in disfavor. However, these critiques of SCMBD must be kept in context. As the next Part explains, the world is full of SCMBD dynamics whether we use automated decision-making or not, and there are some contexts where the cure for SCMBD is worse than the disease.

III. IS SCMBD NOT SO BAD?

Distrust of SCMBDs is fueled by assumptions — assumptions that input factors have roughly linear relationships with outcomes, and

[191] At the risk of hyperbole, complex decision-making models could even contribute to the sort of societal complexity that makes a civilization vulnerable to collapse. JOSEPH A. TAINTER, THE COLLAPSE OF COMPLEX SOCIETIES 91-126 (1988).

[192] *See* Perry & Zarsky, *supra* note 172, at 1603-04 (for the distinction between positive and normative fairness).

[193] *Id.*

assumptions that we know, approximately, what the scale of various effects should be. A decision to override a SCMBD that has predictive validity is a commitment to limit the most powerful potential of big data and machine learning — namely, to correct the preexisting mental models that humans use to predict how things will work out. If SCMBDs are treated as presumptively bad, and especially if they are automatically and unreflectively smoothed out, we may lose an opportunity to discover a latent relationship that is important and true and to dislodge an erroneous heuristic.[194]

In this Part, we make the case that surprising SCMBDs can be cautiously embraced as some of the many phenomena (like quantum physics) that can be validated and put into service even though they are mysterious and difficult to intuitively grasp. We also compare and contrast SCMBDs to similarly lumpy treatment by the law in order to import and export insights about decision-making under real world conditions. We conclude with an argument that in some cases, it may be more valuable for society to correct the real-life dynamics that wind up getting picked up as signals from a machine learning algorithm rather than to program the algorithm to pretend that a SCMBD-like relationship between inputs and outputs does not exist.

A. Life is Lumpy

There are many natural, social, and legal phenomena that cause discontinuities and lumpiness in outcomes. Consider how small changes make a big difference in the natural world: it is full of asymptotes where differences that initially cause small changes suddenly cause dramatic

[194] Erroneous heuristics have a home in many common psychological tendencies. One such tendency is to believe that big events are *always* preceded by big causes (while such events might have been a result of many small changes, or even mere chance). These tendencies have been termed by some as the "proportionality bias" and are mentioned as one of the reasons for the attractiveness of conspiracy theories. Nsikan Akpan, *How to Keep Conspiracy Theories From Ruining Your Time with Family*, PBS NEWSHOUR (Dec. 5, 2019, 11:26 AM EST), https://www.pbs.org/newshour/science/how-to-keep-conspiracy-theories-from-ruining-your-thanksgiving [https://perma.cc/MU94-N85N] (discussing the work of Robert Brotherton on this issue); *see also* Thea Buckley, *Why Do Some People Believe in Conspiracy Theories?*, SCI. AM. (July 1, 2015), https://www.scientificamerican.com/article/why-do-some-people-believe-in-conspiracy-theories/ [https://perma.cc/7UX5-GZ6N]. Other scholars in the field of operational and managerial studies have argued that people fall into a "linear bias" by believing the world is governed by simple linear relationships while often that is not the case. *See* Bart de Langhe, Stefano Puntoni & Richard Larrick, *Linear Thinking in a Nonlinear World*, HARV. BUS. REV., May–June 2017, at 4 ("Our brain wants to make simple straight lines.").

ones. Pharmaceutical drug dosages can have a chaotic nonlinear relationship to efficacy and side effects.[195] Bread dough has a nonlinear relationship to baking time. These natural SCMBDs are jarring because humans experience the world through mostly Newtonian laws of physics. Apply a force and the ball will accelerate. Apply double the force, the ball will doubly accelerate. Of course, we do also interact with everyday physical phenomena that feature exponential growth[196] and functions with steep cliffs.[197] For a mundane example, consider phase transitions when ice becomes water or water becomes steam.[198] Here, at a distinctive point, a slight change in one parameter (such as temperature) leads to a dramatic change in outcome. But the examples of SCMBD (like the melting of ice) that are a routine part of life get internalized without significantly disrupting human expectations of linearity.[199] And humans are extremely bad at appreciating the qualities of exponential growth as they time and again fall for variations of the "pyramid scheme" and fail to foresee the spread of pandemics, leading to what some refer to as the "exponential growth bias."[200]

It requires a sort of mental discipline to observe the world without expectations and accept its nonlinear qualities.[201] Even when the nonlinearities are well understood, causally, by the well-educated (for example, exponential growth in biological systems), individuals who are not accustomed to thinking about nonlinear phenomena will be caught

---

[195] Aristides Dokoumetzidis, Athanassios Iliadis & Panos Macheras, *Nonlinear Dynamics and Chaos Theory: Concepts and Applications Relevant to Pharmacodynamics*, 18 PHARM. RSCH. 415, 424 (2001) (discussing how the notions of the sensitivity from the initial conditions and the qualitatively different behavior for different, even slightly, values of the control parameters, surely play an important role and must be taken into account in modeling since their presence is suggested by experiments).

[196] *See* Eyal Zamir & Doron Teichman, Mathematics, Psychology, and Law: The Legal Ramifications of the Exponential Growth Bias 6 (Mar. 14, 2021) (unpublished manuscript), https://ssrn.com/abstract=3804329 [https://perma.cc/529R-RZ29].

[197] *See* Frank Wilczek, *When Small Changes Make a Big Difference*, WALL ST. J. (Feb. 6, 2019, 11:18 AM ET), https://www.wsj.com/articles/when-small-changes-make-a-big-difference-11549469927 [https://perma.cc/U39N-DAQS] (discussing phase transition, as well as sharp transition in the creation of magnetic forces and fields).

[198] For a similar usage of this example, see Adam J. Kolber, *Smoothing Vague Laws*, *in* VAGUENESS AND LAW: PHILOSOPHICAL AND LEGAL PERSPECTIVES 275, 283 (Geert Keil & Ralf Poscher eds., Oxford Univ. Press 2016).

[199] Langhe et al., *supra* note 194, at 4.

[200] Zamir & Teichman, *supra* note 197, at 2-3, 42.

[201] *Id.* at 4 (explaining that perhaps it is best to replace intuitive judgements with computer-driven decision support systems given the tendency to fall for this bias).

off-guard[202] (as when a small number of COVID cases swiftly overwhelms a community).[203] But there is also a vast terrain of natural and social phenomena that appear to be random (even to experts) yet may actually be deterministically caused by very small differences in initial conditions. Indeed, this is the focus of chaos theory, the branch of mathematics that studies "when the present determines the future, but the approximate present does not approximately determine the future."[204] That entire branch of study is an acknowledgment that what seems like random happenstance is often the result of small changes of input.

Social life is full of SCMBDs, too. All social tipping points (like critical mass or network effects) are examples of SCMBDs in social behavior.[205] These are situations where small incremental changes in, for example, the number of adopters of a new social media platform leads to small changes in additional adoption until a certain point, at which the next small increase causes a huge difference in adoption rates. The spread of viral videos and patterns in democratic participation also have socially-driven SCMBD tipping points.[206] Indeed, Malcolm Gladwell's book on tipping points, which popularized their discussion, is subtitled: "How Little Things Can Make a Big Difference."

Some of the SCMBDs that emerge in big data algorithms are the products of the laws of physics and social dynamics, but some of them are no doubt caused by the laws of, well, *laws*. Regulation frequently creates artificial cliffs that set cut-off rules and other strict categories. The boundaries of these rules subsequently cause dislocations in the patterns of human behavior and circumstances. Therefore, when a small change pushes the individual to a different class or category, it can make a big difference. This is what Lee Fennell has colorfully dubbed "lumpiness," which is featured in many social constructs, including the law[207] and we mine the legal literature on lumpiness in the next Subpart to see what we can learn about lumpiness in machine algorithms. For

[202] Moreover, attempts to "debias" and thus counter this effect are often unsuccessful. *Id.* at 29.

[203] *Id.* at 15.

[204] *Chaos Theory*, WIKIPEDIA, https://en.wikipedia.org/wiki/chaos_theory (last visited Jan. 5, 2021) [https://perma.cc/9NNT-NECF].

[205] THOMAS C. SCHELLING, MICROMOTIVES AND MACROBEHAVIOR 91-96 (2006).

[206] *See* M. Lynne Markus, *Toward a "Critical Mass" Theory of Interactive Media*, 14 COMMC'N RSCH. 491, 501 (1987).

[207] LEE ANNE FENNELL, SLICES AND LUMPS 7 (2019).

our purposes right now, we simply want to acknowledge that the law itself may be the cause of a machine algorithm's SCMBD.[208]

To demonstrate, consider the threshold for receiving federal Medicare health insurance. Subject to some exceptions, the threshold is crossed at age sixty-five.[209] A fintech entity creating a credit matrix for older borrowers with a lower socio-economic status will be very sensitive to ages that are close to sixty-five, even if the importance of age for the rest of the range is relatively small and linear. It is foreseeable and quite reasonable that the algorithm will provide substantially *lower* scores for those just under sixty-five, given the possible financial risk of a medical crisis without sufficient health insurance coverage. Thus, the lumpiness of Medicare coverage will have a secondary effect on financial risk scoring. But Medicare rules could have a tertiary effect as well. Consider how the same fintech company might score a person who is in their early forties. For *that* set of data subjects, one correlate of risk might be the age of the subject's parents, again as a result of the Medicare age cut off. A person whose parent is sixty-three might be at greater risk of suddenly deciding to quit working for a while to care for an ailing parent than a person whose parent is sixty-six (and has access to rehabilitation or long-term care facilities). This would be a tertiary result of the Medicare age cutoff, and it would be harder to intuitively understand than the secondary effects. At some point, a ripple effect of the age cutoff will create patterns that are detectable by machine but inexplicable to observers who don't naturally think about the impact of Medicare rules on, e.g., the credit risks of forty-year-olds.

Human perception and cognition (as well as those of other animals) are to a great extent non-linear, too, because of the way neurons and other network structures in the brain operate.[210] Thus, models for human vision and other cognition have started to change to adopt what we now know about non-linear perception.[211] Even aesthetics have

---

[208] Kolber, *supra* note 54, at 676 (explaining that laws operating on the basis of other laws that are bumpy are bumpy themselves).

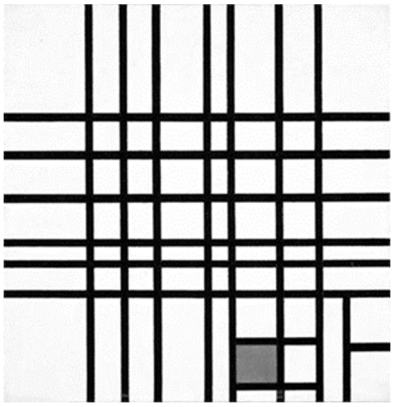
[209] Dena Bunis, *Medicare Eligibility: Do You Qualify?*, AARP (Nov. 15, 2021), https://www.aarp.org/health/medicare-insurance/info-04-2011/medicare-eligibility.html [https://perma.cc/L5UC-BPZ7].

[210] Peter Neri, *Nonlinear Characterization of a Simple Process in Human Vision*, 9 J. VISION 1, 1 (2009). *See generally* ROBERT B. PINTER, NONLINEAR VISION: DETERMINATION OF NEURAL RECEPTIVE FIELDS, FUNCTION, AND NETWORKS (1992) (discussing the "emphasis on nonlinear aspects of vision, from human perception to eye cells of the fly").

[211] Anna Kutschireiter, Simone Carlo Surace, Henning Sprekeler & Jean-Pascal Pfister, *Nonlinear Bayesian Filtering and Learning: A Neuronal Dynamics for Perception*, 7 SCI. REPS. 8722, 8722 (2017) ("The robust estimation of dynamical hidden features, such as the position of prey, based on sensory inputs is one of the hallmarks of

SCMBDs. Indeed, the most famous paintings of Piet Mondrian are his studies of "dynamic equilibrium" where lines and shapes are balanced in a way such that, at least to his eye, a small change to the placement of any of them would cause significant imbalance and cause the painting to lose its elegance.[212]

Figure 4. Piet Mondrian's Composition with Blue



Yet, while we have solid evidence that SCMBD phenomena routinely occur in human biological and social life, we remain blind to them when we rely solely on the current state of knowledge about cause and effect. We are, of course, all the more blind if we rely solely on our own intuitions and lived experiences. These nonlinear dynamics must be teased out through experiments so that we can first know that they occur. Later, if we are lucky, we may be able to alter our causal theories to explain them.

Internet firms have successfully used rapid and dynamic systems of randomized controlled trials (which are referred to as "A/B testing" in the industry) in order to find and then use seemingly minor adjustments that yield big changes in effect.[213] A/B experiments in the Internet

perception. This dynamical estimation can be rigorously formulated by nonlinear Bayesian filtering theory."); Michael J. Richardson, Alexandra Paxton & Nikita A. Kuznetsov, *Nonlinear Methods for Understanding Complex Dynamical Phenomena in Psychological Science*, PSYCH. SCI. AGENDA (Feb. 2017), https://www.apa.org/science/about/psa/2017/02/dynamical-phenomena [https://perma.cc/C7WA-EAAH] ("In many instances, the probabilistic determinism of complex human behavior can only be understood and explained using nonlinear methods of analysis and modeling.").

[212] Abstraction, *1939-42: Piet Mondrian, Dutch*, KIMBELL ART MUSEUM, https://kimbellart.org/collection/ap-199405 [https://perma.cc/J6QX-W3RK].

[213] Jane R. Bambauer, *All Life Is an Experiment. (Sometimes It Is a Controlled Experiment.)*, 47 LOY. U. CHI. L.J. 487, 494 (2015); Michelle N. Meyer, *Two Cheers for*

economy are often maligned as exploitative to the extent that their insights are used to do something that is at least believed to be against the best interests of their users and customers.[214] But when the practice of perpetual A/B testing to find and use SCMBDs is transferred to another context, for example, to fine-tune and customize medical care, the practice that seems risky in one context may be exciting and desirable in the other.[215]

Computers running machine learning algorithms are not necessarily predisposed to expect or seek out linear relationships, so they have an advantage in identifying previously overlooked nonlinear relations.[216] Treating SCMBDs as a problem could be a step in the wrong direction, as these are precisely the insights that innovative algorithmic processes can help us discover and harness.[217]

B. What We Can Learn from Lumpy/Bumpy Laws

Law itself is a frequent source of SCMBDs. In legal systems, small changes that make a big difference can be seen in rules that rely on cutoffs ("bumps") or stark categories (or "lumps"). The Medicare age rule and Three Strikes laws are examples of cutoffs, and laws defining first, second, or third degree homicides are examples of stark category rules.[218] The SCMBD quality of legal rules has received critical attention

Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation, 13 COLO. TECH. L.J. 273, 277 (2015).

214 ZUBOFF, *supra* note 23, at 301.

215 Ron Kohavi, Diane Tang, Ya Xu, Lars G. Hemkens & John P.A. Ioannidis, *Online Randomized Controlled Experiments at Scale: Lessons and Extensions to Medicine*, 21(1) TRIALS 150, 156 (2020), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7007661/ [https://perma.cc/BL8A-75VU] ("Even tiny changes should ideally undergo continuous and repeated evaluations in randomized experiments and learning from their results may be indispensable also for healthcare improvement.").

216 Dinesh Bacham & Janet Zhao, *Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling*, 9 MOODY'S ANALYTICS RISK PERSPS. - MANAGING DISRUPTION, July 2017, at 28, https://www.moodysanalytics.com/-/media/article/2017/risk-perspectives-managing-disruption.pdf [https://perma.cc/8U9E-ZKML]; *see* Zamir & Teichman, *supra* note 197, at 4.

217 *See* Tal Z. Zarsky, *The Privacy – Innovation Conundrum*, 19 LEWIS & CLARK L. REV. 115, 161 (2015).

218 John Clark, James Austin & D. Alan Henry, *'Three Strikes and You're Out': A Review of State Legislation*, NAT'L INST. JUST. RSCH. BRIEF, Sept. 1997, at 1, https://www.ncjrs.gov/pdffiles/165369.pdf [https://perma.cc/BEU4-76M4]; *Three Strikes Basics*, STAN. L. SCH.: THREE STRIKES PROJECT, https://law.stanford.edu/stanford-justice-advocacy-project/three-strikes-basics/ (last visited Dec. 27, 2021) [https://perma.cc/2T3J-VNAW]. Another example which might come to mind is tax brackets, which we chose not to address because the higher tax bracket almost always

in legal scholarship. We can mine that literature to understand how the pros and cons might translate to automated big data decision-making. After all, legal rules are called upon to differentiate and pass judgment on people using various inputs, just as scoring and decision-making algorithms do.

Legal decisions are typically made by first assessing the person or situation on some fine-grained scale ("input"), and then applying a treatment rule to the assessment to decide what to do with them ("output"). For example, a finder of fact will first assess the tort plaintiff's evidence of causation on a probability continuum, and then decide whether they have reached the threshold for preponderance of the evidence to prove the element thus finding liability. Fortunately, this is similar to how firms typically use algorithmic scoring systems, too. Even when treatment of individuals falls into a few crude categories, the algorithm will first assign a fine-grained score and then slot them into the treatment groups according to cut-off or quota rules.[219] So far so good: this means the comparison between automated algorithms and legal outcomes is apt.

Both the assessment and treatment (input/output) parts of the legal process play a role in legal SCMBDs, albeit of two different sorts. For the assessment stage, legal literature on the familiar rules versus standards debate is most relevant. The decision to use a rule or a standard will determine whether the fact-finder will rigidly apply just a few, clear factors or will instead incorporate a holistic assessment of a wide range of inputs. The benefit of rules, of course, is that they provide clear notice and constrain discretion in a way that achieves consistency.[220] And they are often easier to administer since fewer factors have to be considered, let alone measured. But standards are, or at least have the potential to be, more accurate reflections of the quality that is intended to be inferred.

Whether rules or standards are used for assessment, the law still governs the relationship between inputs and legal outputs in a manner that is that is either smooth or bumpy — bumpy in the sense that it breaks people into a small number of treatment options (or lumps) with substantial "cliffs" between them. In terms of outputs, consider legal judgments which use binary treatment — guilty or innocent, for example. Leo Katz argues that "either/or" qualities make the law

applies to the marginal income, thus avoiding several of the problematic dynamics here described.

[219] See discussion of outputs *supra* Parts I.B and I.D.

[220] Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557, 608-09 (1992).

"perverse."[221] He uses the deep bench of doctrines where courts award litigants either all or nothing to illustrate the problems. For instance, in the torts context, a defendant can be found to have either acted reasonably (and pay nothing) or negligently (and pay all).[222] Katz notes the curious persistence of these doctrines despite the fact that people have strong, visceral responses against these sorts of legal cliffs and would prefer to smooth them out.[223] Over time, courts have introduced smoothing doctrines like comparative fault and the loss of chance doctrine to provide continuity where the law was once binary,.[224] Yet Katz nonetheless acknowledges that there are circumstances for which the law simply must include either/or features, making some of the SCMBDs of law (or its "perversions," to use Katz's terminology) unavoidable.

In several influential articles, Adam Kolber conceptualizes the harsh cut-offs in law somewhat differently. Kolber, like Katz, recognizes that the outputs of the law are often "bumpy"[225] as when federal court jurisdiction was denied in a case because the claim was a mere penny(!) short of the statutory threshold.[226] Focusing less on what people intuitively judge as fair and relying more on accuracy and institutional capacity, Kolber calls for greater "smoothing" of legal outcomes.[227] This would be achieved by assuring that the impact of a law is gradual and proportional to the factors that determine its application. Similarly, Lee Fennell has pointed out that the law is often "lumpy": passing a specific threshold causes a substantial change in legal status.[228] Her demonstrative examples include doctrines related to governmental takings that provide rights-holders with full compensation only after crossing a specific (and very high) threshold of loss.[229] Like Kolber, Fennell calls for eliminating discontinuity in the law when gradual shifts can be used instead.[230] And while she, too, recognizes that gradual responses are often infeasible given high administrative costs, her work

[221] *See generally* LEO KATZ, WHY THE LAW IS SO PERVERSE (2011) (promoting the notion that the either/or qualities of the law create philosophical dilemmas referred to as "perverse").

[222] *Id.* at 145.

[223] *Id.* at 146 (referring to the work of Zerubavel).

[224] *Id.* at 145.

[225] Kolber, *supra* note 54, at 655.

[226] *Id.* at 662-63.

[227] *Id.* at 658.

[228] FENNELL, *supra* note 207, at 4.

[229] *Id.* at 218 (discussing, among others, Lucas).

[230] *Id.* at 4.

is closest to our own because she hypothesizes that law could make use of greater data-driven personalization to cheaply create more graduated legal responses and outputs.[231]

Overall, the scholars addressing the noted themes recognize that gradual regulatory responses (or outputs) would be more accurate (that is, they would better match the treatment to the assessment). However, a smooth transition between outputs (which resemble differences in inputs) would come with high administrative costs.[232] This is often unavoidable, as when deciding which individual to hire, or deciding whether a court does or does not have jurisdiction over an individual. They also recognize that there is social utility in stark categories (lumps) and cutoffs (bumps) because with clear rules in place, and with high enough stakes to demand attention, the subjects of laws creating this reality will know the consequences of their future actions and can plan accordingly. Also, bumpy (and lumpy) as the application of a law may be, the general public can at least be assured that the relation between inputs and outputs was subject to a public process of review and *ex post* democratic accountability. And in drastic cases, individuals subjected to a harsh cut-off might be able to appeal to equitable doctrines of leniency and mercy that are common features of the law.[233] For these reasons, society often begrudgingly accepts the mixed effects of bumpy/lumpy laws.

The noted literature shares some commonalities with the broader discourse on rules versus standards, but not as many as it might seem. The rules versus standards debate concerns the role of decision-maker discretion — specifically, the flexibility to consider different inputs, or to consider the *same* inputs differently. The most natural connection between the two discussions would link lumpiness to rules and smoothness to standards since one of the goals and putative advantages of standards is to permit the decision-maker to fine-tune treatment and avoid inaccurate misclassifications. Since rules limit discretion by providing specific guidance — bright line rules — *ex ante*, the known flaw of rules is that there will error at the boundaries of the rule.[234] Since the goal of a standard is to permit the bending of the rule to meet the specific facts, standards seem to be a familiar smoothing agent.

---

[231] *Id.* at 10, 191. *See generally* Ariel Porat & Lior Jacob Strahilevitz, *Personalizing Default Rules and Disclosure with Big Data*, 112 MICH. L. REV. 1417 (2014) (promoting the notion of personalized law).

[232] FENNELL, *supra* note 207, at 10.

[233] Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245, 1285 (2016).

[234] For some key references to this discussion, see Kaplow, *supra* note 220, at 557; Pierre Schlag, *Rules and Standards*, 33 UCLA L. REV. 379, 379 (1985).

However, this mapping is not right in practice or even in theory. As Adam Kolber explains, the two debates capture different aspects of the law.[235] There can be smooth rules and lumpy standards. Rules can be smooth if they are rigid but complex enough to take many relevant factors into account when striving to calibrate inputs to outputs. Indeed, any big data algorithm that lacks SCMBDs will be a smooth set of rigid rules. Even in the small data world, rules that have detailed instructions like some of the more elaborate criminal sentencing guidelines or the payouts used for workers' compensation schedules provide examples of smooth rules.

Conversely, standards can be lumpy if decision-makers use their discretion to create cliff-like judgments. This can occur if the decision-makers are using their discretion to accurately capture real SCMBD dynamics (for the instance by promoting the normative notion that there is a vast difference between risking a serious injury and risking death, or between building a bridge that is tall enough to let the tallest ships through and one that is just one foot shorter.). It can also occur if they over-rely on a specific (and even minor) factor, either consciously or unconsciously. But also, standards will create cliffs any time they are used to sort people, no matter how holistically and carefully, into discrete categories that are treated very differently. Consider how judges might use their discretion to establish a standard for reasonable care. No matter how they use their discretion, the winner-take-all nature of a negligence claim will still create lumpiness — situations where two defendants whose conduct were very similar nevertheless straddle opposite sides of the fault line.

Thus, both rules and standards can generate SCMBDs in the context of "fact inputs-legal outputs." However, the "rules versus standards" debate is very valuable in a couple respects, and is potentially more relevant to the sort of algorithmic SCMBDs we have been discussing because automated algorithms are used primarily to *assess* rather than to design treatment groups. That is, the number and type of treatment options is either predetermined or subject to human discretion when a firm decides to use a machine algorithm. The algorithm is trusted to predict (assess) the individuals, and then to slot them into treatment categories based on pre-programmed rules.[236]

Comparing algorithmic SCMBDs to the "rules versus standards" debate is illuminating because the two realms are surprisingly

[235] Kolber, *supra* note 54, at 667 ("The smooth-bumpy distinction simply captures a different feature of law than does the rule-standard distinction.").

[236] See our discussion *supra* Part I.D.

incompatible. SCMBDs that come from machine learning have the rigid consistency of rules and the accuracy of standards. They are not the product of inter-judge inconsistency where different decision-makers are actually applying different models; to the contrary, at any given time, a machine learning model will be the same for each person who is judged by it. And yet, because the model can take advantage of rules learned from a wide array of different inputs, machine learning algorithms have the advantages that are meant to be captured in standards — allowance for processing of any number of relevant criteria.

But while machine learning algorithms may be more consistent and more accurate all at once, they do not explain themselves, and therefore cannot provide the sort of notice that can help guide behavior and make law predictable, which is one of the key virtues of legal rules. The fact that standards are often adopted despite their lack of clear notice is a useful challenge for AI ethicists. Because law often opts for murky but better-tailored standards over clear rules, it provides an important counterexample to any claim that transparency and explainability are necessary conditions for justice.[237] But the insight runs the other direction, too: when rules are preferred over standards in law because notice and clarity really *are* deemed to be critical to the fairness of a legal outcome, that same logic should apply to any automated algorithm that takes over the decision-making function. The model used by such an algorithm would have to be parsimonious, explainable, and fairly static. If those really are the requirements, there is little reason to switch to a machine decision-maker (aside from the costs of human staffing). This last assertion could be somewhat mitigated when acknowledging that the process of humans deciding on the basis of rules also includes a substantial amount of vagueness and uncertainty, given the hidden nature of the inner workings of the mind. Thus, machine-driven decisions are not that different.

So, to synthesize and summarize the "lumpy/bumpy laws" and "rules versus standards" literature, strict and highly consequential cutoffs in the law lead almost inevitably to arbitrary (inaccurate) outcomes at the boundaries, and they intuitively make jurors and observers uncomfortable. Their drawbacks sound in inaccuracy and incompatibility with human intuition. But these drawbacks of legal SCMBDs are frequently outweighed by concerns of practical

[237] For an example linking explainability to justice in the context of promoting governance, see Brennan-Marquez, *supra* note 11, at 1295 ["explanatory standards serve governance values by eliciting information about official conduct—a precondition of democratic and administrative pushback"].

administrability and by the (relative) benefits of notice that clear and simple rules can bring by setting expectations and guiding good behavior.[238] Moreover, the open democratic process that creates legal SCMBDs provides motivation for lawmakers to avoid unjustified SCMBDs and a mechanism for the public to demand change (smoothing).

Now the payoff: we can see how SCMBDs in automated decision-making (and especially in opaque machine learning algorithms) differ from the SCMBDs that appear in legal rules. As with legal SCMBDs, algorithmic SCMBDs are intuitively disfavored. But that's where the similarities end. Legal SCMBDs cause inaccuracy because they create cliffs based on a small set of simple rules that cannot capture the complexity of objective that the decision-maker is trying to achieve (e.g., measuring desert or need).[239] Big data SCMBDs, by contrast, help improve a decision-maker's accuracy. Or at least, this is so for the SCMBDs that are the most interesting — the ones that have cleared a performance standard and seem to be statistically valid. Thus, when it comes to accuracy alone, SCMBDs in big data are more acceptable than SCMBDs in the law.

On the other hand, the benefits of legal cut-offs and category lumpiness are *not* present with algorithm SCMBDs. Legal cut-offs and stark categories provide procedural benefits of parsimony, clear notice, and behavior guidance. Most big data algorithmic decisions have none of these features.[240] They are the result of complex rather than parsimonious models, they are usually latent (absent some legal requirement or custom of explaining an automated decision[241]), and they, therefore, cannot be useful for guiding subjects' future behavior. And even if SCMBDs were revealed to subjects, there is no guarantee

---

[238] KATZ, *supra* note 222, at 144; Carol M. Rose, *Crystals and Mud in Property*, 40 STAN. L. REV. 577, 577 (1988) ("[T]heir great advantage, or so it is commonly thought, is that they signal to all of us, in a clear and distinct language, precisely what our obligations are and how we may take care of our interests.").

[239] Alternatively, when simple rules in law are more standard-like in practice, and permit myriad factors to be considered, then they have the problems of standards — less notice, and more discretion and arbitrariness from human factors.

[240] Crawford & Schultz, *supra* note 16, at 122-23 (discussing the lack of notice and "due process" in general in algorithmic decisions).

[241] Given the complexity of machine learning algorithms, meaningful after-the-fact review is extremely challenging to design, too, even if or when it is required. Huq, *supra* note 7, at 7; Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1094-99 (2018) [hereinafter *The Intuitive Appeal of Explainable Machines*] (discussing the challenge of overcoming the inscrutability and non-intuitiveness of automated decision-making).

that the subject would actually be able to make the small change that yields a big difference.[242] Some small changes are costly (like moving a few blocks away) and some are impossible (like being younger). Thus, the benefits of legal SCMBDs are drawbacks of big data SCMBDs, and vice versa.

Another possible difference between algorithm SCMBDs and legal rule lumpiness and bumpiness is the nature of boundary conditions. A typical legal rule involves a limited number of inputs and a simple combination of those inputs, such that only a (relatively) small number of SCMBD cliffs are possible. In a complex machine learning system, by contrast, there can be a large number of SCMBD discontinuities. These differences may be particularly problematic for algorithm accountability because individuals may be affected by distinct SCMBD decision boundaries and may not even be aware of the nature of the differences between their outcomes. Therefore, at least in some cases, legal SCMBDs present an issue of lesser concern.

So, big data-driven SCMBDs will require a new analysis of the tradeoffs between procedural interests (notice, explainability, and parsimony) on one hand and accuracy on the other. Note that legal cut-offs, even when premised on or derived from clear rules share some of the problematic attributes with the big data ones. The actual decision at the end of the day is made by a human decider who might be considering numerous factors to establish whether a specific threshold was crossed. For that reason, the benefits lost when shifting to algorithm-driven decisions are not as substantial since their parsimony and explainability is already compromised.[243]

When assessing these tradeoffs, we should avoid relying on heuristics and implicit assumptions that may have historically held true but are broken by big data algorithms. First, procedural rights are at least partly a tool to improve accuracy. In a low information environment, procedural protections that allow individuals to challenge the propriety of the way they are treated is one of the best ways to achieve accurate outcomes. In a high information environment, accuracy and procedural interests are less interconnected, maybe wholly unrelated, and possibly even negatively correlated in some cases.[244] In other words, legal and

---

[242] *See* Barocas et al., *supra* note 56, at 84 (discussing changes that are costly or impossible).

[243] We thank Adam Kolber for this insight.

[244] Tracy Tullis, *How Game Theory Helped Improve New York City's High School Application Process*, N.Y. TIMES (Dec. 5, 2014), https://www.nytimes.com/2014/12/07/nyregion/how-game-theory-helped-improve-new-york-city-high-school-application-process.html [https://perma.cc/AU52-UNHB].

cultural traditions probably had to rely on procedural rights both for intrinsic purposes (because they are good in themselves to give notice, voice, and legitimacy to subjects) *and* instrumental purposes (because those same rights tended to produce more accurate outcomes). However, the logic of procedural rights doesn't carry over as well to AI decision-making if black box methods are verifiably more accurate. The internal balance between these elements has thus changed.

Second, in considering the balance between procedural interests and accuracy, distributional fairness adds a third dimension. Transparent and parsimonious models of decision-making might have better effects on how error is distributed across different subpopulations, but it might not. This third dimension of fairness will have context-dependent effects on the tradeoffs between procedural fairness and accuracy. Any assumptions that increased procedure is better for vulnerable minority populations should be examined.

One last distinction between SCMBDs in law and SCMBDs in big data algorithms is worth reflecting on: SCMBDs in big data can be smoothed without any administrability problems (as opposed to the need to apply new forms of laws or regulations in the legal context). Indeed, one of the great advantages of automated decision-making is that it can implement *any* set of instructions, including SCMBD removal, while minimizing the impact on accuracy and other decisional goals. If SCMBDs are objectionable *per se*, firms can smooth them out in a way that preserves other objectives with virtually no additional cost or hassle.[245] Whether they should do so is, of course, harder to say.

C. No YOU Smooth Out

SCMBDs offer an opportunity to learn. They reveal a striking relationship in the complex knot of factors that affect our social and economic lives. A SCMBD could unearth some unintended ripple effects of lumpy laws, or could highlight industry practices and social norms that have surprisingly large consequences. When that happens, they might motivate an effort to smooth out those predicate practices and to thereby improve the fairness of life itself (and not just the algorithm). For example, a SCMBD audit that uncovers a discontinuity between the age of a credit applicants' parents and the predicted risk might be mysterious initially but could eventually uncover the ripple effects that Medicare cut-off have on the financial health of family members.[246]

[245] *Constrained Optimization*, WIKIPEDIA, https://en.wikipedia.org/wiki/Constrained_optimization (last accessed Jan. 5, 2022) [https://perma.cc/K38D-S5G5].

[246] *See supra* text accompanying notes 204–205.

A SCMBD could also inform decision-makers that their own decision categories are too lumpy. For example, consider ProPublica's analysis of COMPAS recidivism risk scores.[247] The bias in false positive error that the journalists uncovered was premised on an assumption about how arrestees are treated. COMPAS produced recidivism risk scores on a 1–10 scale, but the ProPublica researchers assumed arrestees would be treated in roughly binary categories — either as a risky individual (score above 3) or not (score of 1–3).[248] If courts treat scores above 3 as "risky" and deny pretrial release as a result, then the factors that go into the COMPAS scores (including age, age at first arrest, and criminal record) are bound to have SCMBD effects close to the boundaries between 3 and 4. There will be some small difference in age, or in the severity of previous charges, or in some other factor that causes a big difference in treatment (from release to detention.) If instead the judge uses the scores to create steadily graduated responses (e.g., release, release with monitoring, release with bail of lesser or greater amounts, and no release), the incidence of SCMBDs would be reduced (and the racial bias in error would, too.) The discovery of a SCMBD can instruct decision-makers to modify their ultimate treatment of the data subjects by avoiding all-or-nothing decisions, to the extent possible. Indeed, the report itself quotes a professional who trains judges about how to use the risk scores: "These risk factors don't tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be."[249]

To harness this potential and to better understand the tradeoffs between SCMBDs and competing notions of fairness, a decision-maker using an automated process would be well advised to detect SCMBDs but not, necessarily, to correct for them. The goodness or badness of SCMBDs will require some reflection on their advantages and shortcomings in context.[250] Thus, the first step in optimal AI policy design will encourage the initiation of SCMBD auditing.

IV. THE VALUE AND LIMITS OF SCMBD AUDITS

The task of defining and implementing ethics in Artificial Intelligence will keep technology policy experts busy for a very long time to come.

---

[247] Angwin et al., *supra* note 133.

[248] *Id.*

[249] *Id.*

[250] Kolber, *supra* note 54, at 688 (explaining, in the broader context of "cut off" rules, that a theory of descriptive and normative law needs to provide a response as to whether a relationship needs to be smooth or bumpy).

SCMBD is just one of a number of potential sources of unfairness. However, it might be a particularly useful portal into the larger enterprise of fairness because SCMBDs are easy to find and will surely capture the public's attention.

In this final Part, we explain why it would be good policy (as a matter of industry self-regulation or as part of a future public mandate for AI impact assessments) to run an audit for SCMBD. [251] After all, even if readers were not convinced in Part II that SCMBDs are frequently troubling, they are bound to create scandal. The general public has a strong, intuitive aversion to SCMBD dynamics. An audit that explores how the SCMBD relates to various theories of fairness will redound to the benefit of both the company and an alarmed public, even if it ultimately concludes that a SCMBD is well-justified. First, we situate the SCMBD audits within the wide range of proposals for machine learning audits. We then explain how they could be performed and end with a brief discussion of the hardest part of the process: deciding what to do *after* a SCMBD has been detected.

A. The Purpose of the SCMBD Audit

Broadly speaking, there are two types of algorithm audits: process-focused audits that seek to explain a specific algorithmic decision about an individual, and outcome-focused audits that review algorithms against fairness criteria.[252] Both types of audits could be used to address SCMBD concerns.

---

[251] Anecdotally, we have heard from some machine learning implementers that in their opinion it is already a practice of good data science to seek out and eliminate SCMBDS whenever possible using smoothing or "regularization" techniques.

Interviews of programmers and data scientists at Palantir with Tal Zarsky in New York City, N.Y. (Dec. 19, 2019) (on file with authors). Regularization controls for steep slopes and excessive fluctuations in an effort to avoid overfitting the training data. *See* Megha Mishra, *Regularization: An Important Concept in Machine Learning*, TOWARDS DATA SCI. (May 26, 2018), https://towardsdatascience.com/regularization-an-important-concept-in-machine-learning-5891628907ea [https://perma.cc/WW7F-W3Z3].

[252] The process-focused audits that we describe in this Section are part of a broad literature on interpreting and explaining machine learning systems. *See generally* CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING: A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE (2019), https://christophm.github.io/interpretable-ml-book/ [https://perma.cc/YF7J-N485] (providing an overview of explainable machine learning models and methods for interpreting other models); Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE ACCESS 52138 (2018) (surveying research literature on explainability and interpretability in machine learning); Nadia Burkart & Marco F. Huber, *A Survey on the Explainability of Supervised Machine Learning*, 70 J. A.I. RSCH. 245 (2021) (summarizing explainable machine learning models and a diverse range of methods for interpreting

Process-focused audits attempt to address "black box" concerns about opacity by providing regulators or the subjects of automated scoring systems with information about the latent workings of the algorithm.[253] For scholars who consider inscrutable algorithms as a breach of fairness or due process *per se*, an audit that provides some meaningful explanation about a decision is very important (if not mandatory).[254]

There are now several regulatory mandates requiring firms to let subjects peer inside the black box, to some extent. Most prominently, the EU's General Data Protection Regulation requires that data subjects be provided with the "logic" of automated decision-making systems associated with serious consequences,[255] and the GDPR also guarantees that data subjects can challenge certain algorithmic determinations.[256] The EU is also forwarding a proposal to provide similar rights for many forms of AI-driven analytics in general.[257] The "process-focused" form of auditing is also reflected in sector-specific U.S. consumer protection law. For example, in the context of consumer credit, the Fair Credit Reporting Act requires credit bureaus to disclose "key factors" for adverse credit scoring decisions[258] and guarantees consumers a right to submit a "written request for . . . reasons" after adverse credit decisions.[259] Similarly, the Equal Credit Opportunity Act and

other models); Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti & Dino Pedreschi, *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUTING SURVS. 1 (2018) (surveying and comparing scholarship proposing methods for interpreting machine learning models). Outcome-focused audits are situated in scholarship on identifying forms of group bias in machine learning systems. *See generally* SOLON BAROCAS, MORITZ HARDT & ARVIND NARAYANAN, FAIRNESS AND MACHINE LEARNING: LIMITATIONS AND OPPORTUNITIES (2019) (for an overall discussion of the difficulties of achieving fairness in machine learning, including a discussion of the various biases and transparency).

253 PASQUALE, *supra* note 12, at 142.

254 *See* Kaminski, *supra* note 11, at 190-93 (demonstrating an extensive review of the literature on this point).

255 General Data Protection Regulation 2016/679, art. 15(1)(h) O.J. (L 119).

256 *Id.* art. 22(3); *see also* Kaminski, *supra* note 11, at 204; Wachter et al., *supra* note 4, at 850-51 ("Generally, the idea is to create a simple human-understandable approximation of a decision-making algorithm.").

257 Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, arts. 13, 52, COM (2021) 206 final (Apr. 21, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 [https://perma.cc/KC9X-9E8E] (providing transparency rights both in general and for "certain AI systems").

258 15 U.S.C. § 1681g(f)(1).

259 15 U.S.C. § 1681m(b)(1).

Regulation B require a creditor to furnish the "principal reasons" motivating an adverse decision.[260]

A "process-focused" approach aspires for fairness through transparency. Information about the algorithm empowers users *ex post* with a better understanding of how they were judged, which can lead to formal and informal challenges. The information can also be used by the data subject to make informed decisions about future conduct to improve their chances of success, thus increasing personal autonomy.[261] Transparency also serves a deterrent function *ex ante*, of course, to ensure that a firm would not use prohibited factors, a close proxy for a protected class, or any other factor that would trouble data subjects or authorities. But process-focused audits have drawbacks, too. The sort of disclosure that is comprehensive and most valuable for understanding an algorithm may not be feasible for some machine learning and artificial intelligence applications because they use complex and ever-changing decision models that frustrate explanation.[262] And if the regulatory requirements mandate simplification to enable suitable transparency, this might compromise the system's overall precision and efficacy. Even in a modest, more limited form, process audits can reveal trade secrets, increase gaming, or inhibit accuracy and innovation.[263]

Outcome-focused audits examine how a scoring or decision-making algorithm will affect society without necessarily requiring an explanation of the model and its inner workings .[264] They are part of an impact assessment firms carry out prior to introducing an automated decision-making tool (and throughout its use as well) rather than

---

[260] 15 U.S.C. § 1691; 12 C.F.R. § 1002.9 (2021); *see* Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 241, at 1099-1108.

[261] Barocas et al., *supra* note 56.

[262] Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1040 (2017) (discussing the general arguments against transparency in this context). However, there are many forms of transparency that *are* possible, even with deep learning algorithms. *See* Selbst & Barocas, *The Intuitive Appeal of Explainable Machines*, *supra* note 241, 1109-115.

[263] *See* Bambauer & Zarsky, *The Algorithmic Game*, *supra* note 145, at 28; Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PENN. L. REV. 633, 638-39 (2017); Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI*, 2019 COLUM. BUS. L. REV. 494, 591-610.

[264] The EU's Ethics Guidelines for Trustworthy AI, *supra* note 96, at 29, also address the importance of such forms of audits. "The outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users)." *See also* Chander, *supra* note 262, at 1043 (noting work by several researchers proposing methods for measuring bias without full transparency).

public-facing disclosure tools.[265] For example, there are auditing methods in use or in development that identify race and gender biases of the sorts we described in Part II.[266] SCMBD audits fit well in this model because they shed light on systemic rather than individualistic unfairness.[267] An individual who discovers that their treatment was affected by a SCMBD would not be able to know how this particular feature of the predictive model relates to accuracy or distributional fairness whereas an outcome-focused audit can do follow-up exploratory analyses of these sorts. Moreover, a SCMBD audit is fairly easy to do (as we explain next) and therefore, might be a sensible procedure early in an impact assessment.

B. How to Audit for SCMBDs

Tools that already exist for examining algorithms can be easily adapted to address SCMBD concerns.[268] While a detailed review of methods for scrutinizing algorithm behavior is beyond the scope of this work, we briefly describe several approaches that could surface possible SCMBD problems.[269]

In the most straightforward scenario, an algorithmic decision-making system uses an interpretable model, where an analyst can directly inspect the model and understand its behavior. The perceptron algorithm for making credit card recommendations in Part I.D is a good example: merely examining feature weights is sufficient to understand what the model has learned and how it will behave. A large weight for an intuitively unimportant feature, such as the color of a person's car, is a warning sign about possible SCMBD.

The trend in modern machine learning is toward models that are not readily interpretable, such as deep learning approaches.[270] Computer scientists have responded by developing tools to explain how machine

---

[265] Indeed, such impact assessments are required under several existing and proposed regimes. *See* Yifat Nahmias & Maayan Perel, *The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations*, 58 HARV. J. LEGIS. 145, 159-62 (2021).

[266] Dwork et al., *supra* note 14, at 226.

[267] *See, e.g.*, Chander, *supra* note 262, at 1044 (noting the reviewing of inputs and outputs by a third party as a possible form of algorithmic auditing).

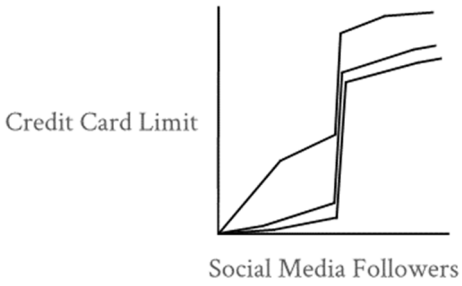
[268] Some scholars have addressed the usage of transparency measures comparing inputs to outputs. *See* Lehr & Ohm, *supra* note 51, at 709-10.

[269] *See* Kaminski, *supra* note 254.

[270] Deep learning models rely on complex functions that can elude human intuition and increasingly operate at a scale that exceeds individual capacity for understanding. *See* Burkart & Huber, *supra* note 252, at 1-4; Guidotti et al., *supra* note 252, at 9.

learning systems behave, even when the models themselves are not interpretable.[271]

Individual conditional expectation ("ICE") plots are a particularly simple and widely adopted tool for understanding machine learning models.[272] The concept is straightforward: visualize how an algorithm responds to changes in an input variable, by plotting examples of algorithm output when the input variable changes. An ICE plot consists of a series of curves, each the result of running the algorithm on an example datapoint (e.g., from a training dataset) while varying the input variable of interest. The following ICE plot is an (admittedly contrived) example based on the credit card scoring algorithm in Part I.D.



Notice the spike in the ICE plot for how the algorithm responds to changes in the credit card applicant's social media following. That jump in the ICE curves suggests a possible SCMBD issue: a quantitatively small change in an input variable (social media followers) is consistently resulting in a quantitatively large difference in algorithm output (the recommended credit card limit).

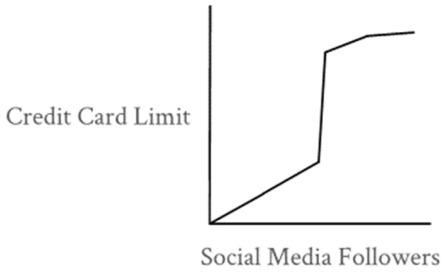
Partial dependence plots ("PDPs") are conceptually very similar to ICE plots (and, in fact, are the predecessor to ICE plots).[273] Generating a PDP is the same process as generating an ICE plot, except instead of sketching a curve for each example datapoint, the PDP has one curve representing the average algorithm output. Intuitively, a PDP asks: how

[271] *See* Guidotti et al., *supra* note 252, at 9-13 (describing the general problem of explaining "black box" machine learning systems).

[272] Alex Goldstein, Adam Kapelner, Justin Bleich & Emil Pitkin, *Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation*, 24 J. COMPUTATIONAL & GRAPHICAL STAT. 44, 47-51 (2015).

[273] Jerome H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, 29 ANNALS STATISTICS 1189, 1219-23 (2001).

would the model respond, averaged across the dataset, to changes in the value of a specific input variable?



For SCMBD purposes, interpreting a PDP is much the same as analyzing ICE plots. A nonintuitive spike — like in the example above — is a warning about a possible SCMBD issue.

Counterfactual explanations ("CFEs") are another valuable tool for surfacing possible SCMBD problems.[274] CFEs operate on individual datapoints and ask, roughly: what is the smallest hypothetical set of feature changes that would result in the algorithm producing a different output? The original work on CFEs was intended to empower users, in the context of the GDPR's right to understand algorithmic decision-making (although not necessarily meeting the law's specific requirements).[275] But CFEs could easily be repurposed to identify

[274] *See generally* Susanne Dandl, Christoph Molnar, Martin Binder & Bernd Bischl, *Multi-Objective Counterfactual Explanations*, 12269 LECTURE NOTES COMP. SCI. 448 (2020) (proposing a version of CFE that generates a diverse set of counterfactuals by reframing CFE as an optimization problem); Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam & Payel Das, *Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives*, 32 CONF. NEURAL INFO. PROCESSING SYS. 590 (2018) (proposing a CFE method that identifies the minimal features to obtain a classification result when present and to change a classification result when absent); Amir-Hossein Karimi, Gilles Barthe, Borja Balle & Isabel Valera, *Model-Agnostic Counterfactual Explanations for Consequential Decisions*, 108 PROC. MACH. LEARNING RSCH. 895 (2020) (proposing a CFE method that generalizes across model and data types and generates a diverse set of counterfactuals); Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard & Marcin Detyniecki, *Comparison-Based Inverse Classification for Interpretability in Machine Learning*, 853 COMMN'CS COMPUT. & INFO. SCI. 100 (2018) (proposing a CFE method that identifies a minimal change needed to alter a classification result); Ramaravind K. Mothilal, Amit Sharma & Chenhao Tan, *Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations*, ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 607 (2020) (proposing a C method that generates a diverse and feasible set of counterfactual explanations); Wachter et al., *supra* note 4, at 844-45 (defining what counterfactuals are).

[275] *See* Wachter et al., *supra* note 4, at 843-44.

possible SCMBDs: a counterfactual where a small change makes a big difference is, by definition, a SCMBD. CFEs could identify SCMBD risks for individuals, or for datasets (by testing each individual in the dataset), or for a space of hypothetical individuals (by generating a simulated dataset that is representative of possible individuals).

These three approaches are far from the only ways to examine opaque algorithms for SCMBD problems. Other tools exist that explain algorithms by building small interpretable machine learning models around individual datapoints, for example, or attempting to convert a complex model to a set of simple rules.[276] There are subtle and challenging tradeoffs between these approaches. ICE plots and PDPs, for instance, do not account for correlations among features. CFE methods involve value-laden assumptions and can produce very different output depending on those assumptions.[277]

We do not endorse any particular approach for identifying SCMBDs in algorithmic decision-making systems. Rather, our goal is to show that existing tools are up to the task. As large organizations increasingly carry out ethical reviews for algorithms, detecting SCMBDs should be a goal alongside detecting transparency and bias issues.

C. SCMBD Detected. Now What?

If a decision-making algorithm is using a SCMBD in its model, the firm will be best served by learning more about the SCMBD, so long as the inquiry doesn't add significant cost. A firm could decide to smooth out SCMBDs automatically or as a default if further investigation can't be performed immediately. But in many cases, it would be a service to the firm and to the community to explore the SCMBD and understand how it relates to accuracy and distributive goals.

For both accuracy and distributional effects, it is fairly easy to measure the cost of removing a SCMBD by simulating its removal with the training or feedback data that produced the SCMBD in the first place. If the SCMBD were smoothed, how would error change? And if the firm knows the gender, race, or other protected characteristics of subjects, how would that change in error be distributed?[278] Do the

---

[276] *See* Burkart & Huber, *supra* note 252, at 9-11, 25-46 (surveying methods for machine learning explainability).

[277] *See* Barocas et al., *supra* note 56, at 80.

[278] If the firm doesn't have demographic data on their subjects, an estimate can be made using aggregate data available elsewhere. For example, a SCMBD related to an input for whether the data subject uses a Mac or a PC would have predictable impact on race based on available data on Mac adoption by race.

patterns in error reveal anything about the causal source of the SCMBD? (That is, can we discover preexisting human or social factors that affect the inputs or outputs and that could themselves be changed?)

Armed with this information, the firm can make a fully conscious choice about how to balance competing values in accuracy, antidiscrimination, gaming, proportionality, and parsimony. *That's* when matters actually get hard. It requires a firm or its auditors to adopt a unified and mostly-coherent theory of what is fair and unfair. In a small data world, technical infeasibility allowed for a greater amount of pluralism in corporate or political governance because the options were sparse. With big data and machine learning, decision-makers cannot engage in polite indecision. Moreover, there is no escape from the ethical debate because even a decision to withdraw from automated decision-making and revert to human systems will lead to unfairness and inefficiencies and thus attract criticism.[279] Thus, while automated decision-making does introduce some new problems that are unique to machine learning (including the proliferation of SCMBDs), its greatest problems are not new. Instead, new technologies are unearthing old, festering problems.

CONCLUSION

When a small change in behavior or characteristics causes a large difference in how a person is treated, it will strike many people as wrong. SCMBD dynamics are likely to increase in number and salience as machine learning is introduced in more contexts. Thus, SCMBDs deserve a prominent place in discussions about fair and ethical AI.

This Article has provided solid theoretical footing for our intuitive reactions against these SCMBDs. Some explanations are innate (goals of proportionality and parsimony) and some are instrumental (goals of accuracy and non-discrimination.) Yet none of these are slam-dunk justifications for eliminating SCMBDs altogether. Our discussion leads to two insights with applicability beyond the law and policy of AI. First, a preference for proportionality is less defensible than it may seem, particularly when there is good evidence that a SCMBD relationship is accurate. There may be other natural, social, or even legal phenomena that cause disproportionality, and it may not make sense to saddle

[279] This is a form of the "compared to what?" meta-critique of machine learning criticism. *See* Jane Bambauer, *Other People's Papers*, 94 TEX. L. REV. 205, 256-57 (2015); Tal Z. Zarsky, *Governmental Data Mining and Its Alternatives*, 116 PENN. ST. L. REV. 285, 289 (2011) (applying a methodology of examining alternatives as a means to study data mining strategies).

downstream actors with the job of smoothing out and correcting for these dynamics.

Second, using SCMBD as a lens to view the larger discourse on ethics during rapid technological innovation, we find that the *pace* of change is not really the source of rancor. Rather, it's technology's tendency to make the impossible possible, and the invisible visible, that will cause strife between stakeholders with different ethical priorities. In addition, technology renders existing cutoffs, "cliffs". "lumps" and "bumps" invisible thus requiring extra rigor in attempts to expose them. We should therefore expect AI applications to produce a lot of light *and* heat in technology policy discourse in the coming years.