THE ETHICAL USE OF PREDICTIVE ALGORITHMS

Katrina Geddes

INTRODUCTION

Northern life insurers in the 1880s (largely White, male, and monied) struggled to visualize a postbellum world in which African Americans enjoyed the same longevity as Whites, so they projected historical Black mortality rates into the future, and paid lower insurance benefits to Black policyholders as a result. Metropolitan, for example, would only grant two-thirds of the standard benefit to beneficiaries of Black policyholders, even if they paid the same premiums as Whites.[1] Insurers justified this discrimination on the basis of higher Black mortality rates, even as they chose not to penalize Southern White policyholders (with similarly high mortality rates) in the name of post-war reconciliation.[2] Framing higher Black mortality as a "statistical fact" also suggested its *inevitability* and *inalterability*, thereby obscuring the ways in which the historical actions of Whites had actively produced these mortality outcomes through slavery and racial oppression.

In 1884, Massachusetts state representative Julius C. Chappelle proposed legislation to prohibit insurers from discriminating on the basis of race. Chappelle was not persuaded by historical data, or claims of business necessity. He argued in favor of a forward-looking approach that emphasized *future possibilities*, including the possibility that African Americans could, one day, live as long as Whites, especially if they lived on equal terms.[3] Chappelle viewed the Civil War as irreparably severing the past from the present and the future, giving postbellum hope to new opportunities for African Americans. For this reason, Chappelle argued, historic Black mortality rates should *not* be used to insure a future that might look very different. The Massachusetts legislature passed Chappelle's anti-discrimination bill, and it was signed by Governor George Dexter Robinson.[4]

Today, the statistical discrimination normalized by insurers over the course of the twentieth century has assumed a new, algorithmic form.[5] Risk assessment tools use statistical correlations within population data to "predict" the likelihood of an individual

---

[1] DAN BOUK, HOW OUR DAYS BECAME NUMBERED: RISK AND THE RISE OF THE STATISTICAL INDIVIDUAL 34 (2015).
[2] *Id.* at 37.
[3] *Id.* at 41.
[4] *Id.* at 44.
[5] Rodrigo Ochigame, *The Long History of Algorithmic Fairness*, PHENOMENAL WORLD (Jan. 30, 2020), https://phenomenalworld.org/analysis/long-history-algorithmic-fairness.

defaulting on future payments,[6] engaging in criminal activity,[7] or requiring hospital care.[8] Despite the efficiency gains associated with their use, predictive algorithms also cause harm along two intersecting dimensions: relational (individual vs. collective), and temporal (backward vs. forward-looking). The *backward-looking* approach focuses on how historical data is assembled and constructed to provide a specific narrative about the past (and related predictions about the future). Data about individual lives is sanded down to their common threads, and mined for statistical correlations. This commensuration of individuals along a statistical distribution *erases* important qualitative differences between individuals through its construction of homogeneity. In criminal sentencing, for example, algorithmic predictions of recidivism will draw on population-level data to guide decision-making about individual defendants, without considering *qualitative* contextual data about their character, history, or circumstances that reflect their divergence from their statistical peers. Because the risk assessment tool will treat the defendant as a statistical average, this de-individualization may harm her sense of dignity, autonomy, and singularity.[9] Over time, as other individuals experience similar treatment, public trust in institutions may diminish. People who feel unseen or unheard by algorithm-assisted decision-making may be less likely to cooperate with institutions that rely on such decision-making, thereby frustrating their institutional effectiveness.

In contrast, the *forward-looking* approach examines how the future is constructed to resemble the past in ways that foreclose future potentiality to certain "high-risk" groups. Individual defendants who share statistical features with historical recidivists, for example, are expected to recidivate in the future. Risk assessment tools cannot conceive potential futures that diverge from the historical data on which they have been trained, so these futures will be foreclosed to the individual, as they are foreclosed to the algorithm. Tools like COMPAS construct a specific temporal relation between past, present, and future, in which criminal activity inevitably recurs throughout, thereby lending the algorithm its

---

[6] Mark Kear, *Playing the credit score game: algorithms, 'positive' data and the personification of financial objects*, 46 ECON. SOC'Y 346, 346-68 (2017), https://www.tandfonline.com/doi/abs/10.1080/03085147.2017.1412642.

[7] *See, e.g.*, Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109 (2017), https://openscholarship.wustl.edu/cgi/viewcontent.cgi?article=6306&context=law_lawreview.

[8] Ziad Obermeyer et al., *Dissecting racial bias in an algorithm used to manage the health of populations*, SCI., Oct. 25 2019, at 447, 447-53., https://science.sciencemag.org/content/366/6464/447.

[9] *See, e.g.*, *Loomis v. Wisconsin*, 881 N.W.2d 749 (Wis. 2016), https://caselaw.findlaw.com/wi-supreme-court/1742124.html.

preemptive power.[10] Only by transforming a future uncertainty (the recurrence of criminal activity) into an inevitability, does COMPAS render it actionable in the present. Other possibilities for the future (e.g. non-criminality) are foreclosed by the algorithm. As a result, a potential future in which the defendant is free, and does not recidivate, is physically foreclosed by the defendant's continued incarceration (on the algorithmic assumption of future recidivism).

Naturally, this foreclosure of future potentiality harms the affected individual, whose life prospects are altered by these algorithmic assumptions. But this foreclosure of future potentiality also causes *collective* harm, as it affects more and more individuals who share the same statistical features. Risk assessment tools which *over*-predict recidivism for Black defendants, and *under*-predict recidivism for White defendants,[11] for example, split the future into two racially distinct times: a White time that is futurally open (that enjoys the benefit of the doubt, and access to new opportunities), and a non-White time that is futurally closed (constrained by the risks reflected in historical data). This unequal distribution of future potentiality locks communities of color into cycles of incarceration and poverty, thereby entrenching existing inequalities, and frustrating opportunities for change. This asymmetric distribution of future potentiality reflects what Charles Mills calls the "racialization of time," or the transfer of time from one set of lives to another.[12]

The unethical uses of predictive algorithms can be grouped under each of these temporal frames, bound together by a theory of temporal governmentality, or the governance of conduct through the mode of time. Differentiating between these temporal frames will allow us to develop unique remedies for the unique harms generated by each frame. The backward-looking approach is more compelling in circumstances where the legitimacy of decision-making depends on its level of individuation. In contrast, the forward-looking approach finds particular purchase in settings where the potentiality of the future is highly valued, and its foreclosure is sorely felt. Predictive algorithms cannot conceive of a future that diverges from the historical data on which they have been trained. This might be described as their *futural* inaccuracy, or their inability to perceive unscanned horizons. To the extent that future rights and interests are

---

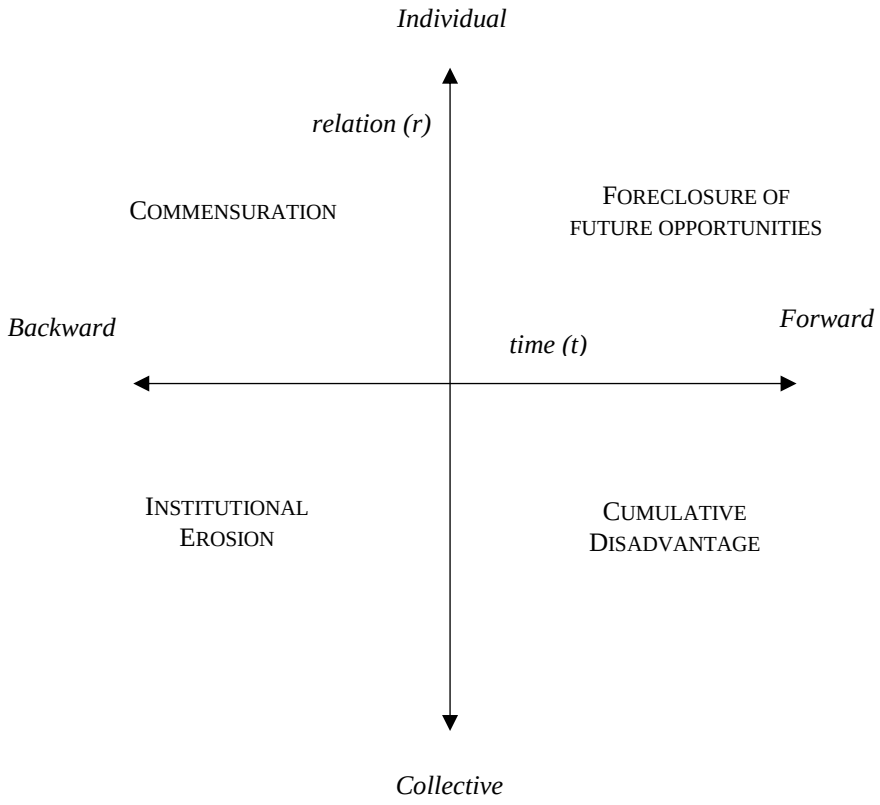
[10] *See, e.g.*, Julia Angwin et al., *Machine Bias*, PROPUBLICA, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[11] *Id.*

[12] See Charles W. Mills, *White time: The chronic injustice of ideal theory*, 11 DU BOIS REV. 27, 27-42, https://www.scholars.northwestern.edu/en/publications/white-time-the-chronic-injustice-of-ideal-theory.

determined on the basis of these limited predictions, opportunities for social change will be similarly constrained.

In other words, the moral permissibility of deploying a specific predictive algorithm is likely to be informed, not only by a conceptualization of the harm(s) caused under either or both of the aforementioned temporal frames, but also by a consideration of any utility gains associated with its use, and the relative importance of individualized treatment (or an accommodation of future potentiality) to the perceived legitimacy of the relevant decision. Each of these factors will influence the moral permissibility of specific predictive algorithms.



THE FORWARD-LOOKING APPROACH

As aforementioned, the ethical harms generated by the use of predictive algorithms can be charted along two temporal axes: looking *backward* at the historical data that was constructed to build the predictive model, or looking *forward* at the futures that are foreclosed by the algorithm's predictions. This Section focuses on the harms associated with the second temporal frame, including the frustration of opportunities for social change, and the entrenchment

of existing inequalities. Predictive algorithms foreclose opportunities for social mobility by binding us to predetermined data pathways, informed by historical inequalities. I evaluate these harms through the racialization of time, and show that the potentiality of the future is a social good that should be preserved.

Earlier this year, British high school students who were unable to sit their final exams due to coronavirus restrictions received "predicted" grades, algorithmically derived from historical grades received by students at their school, and student rankings based on teacher estimated grades. Students at state schools disproportionately received worse-than-expected grades, relative to students at private schools. A high-performing student at a historically low-performing school could not receive a higher grade than had previously been achieved at her school, regardless of her individual performance.[13] As a result, many students lost their place at universities where their entrance offer had been conditional upon the receipt of a certain grade.

Intuitively, this outcome offends our sense of basic fairness, and our belief that education should support social mobility. At the *individual* level, the algorithm fails to capture each student's unique academic ability and effort, including the preparations they had made for the final exam. This offends our sense of individual agency, and self-determination. At the *collective* level, the algorithm's preferential treatment of students from high-income schools reduces the capacity of education to support social mobility. Since a student's final grade determines her entry into university (and tertiary education is highly correlated with improvements in socio-economic status), a strong final grade could materially alter the prospects of a student from a low-income background. Instead, the algorithm *entrenched* existing inequality by foreclosing to low-income students the possibility of a future that looked different from their past. In this sense, the collective harm wrought by the algorithm is the reproduction of inequality, and the lost opportunity for social change.

The public outcry over predicted grades reflects the dissonance between the algorithm's disparate impact on low-income students, and popular perceptions of education as the great social equalizer. It demonstrates also that the forward-looking approach, by focusing on the futures foreclosed by predictive algorithms, finds particular purchase in settings where the potentiality of the future is highly valued, and its loss is sorely felt. Education is one such

[13] *See, e.g.*, Melissa Fai et al., *Lessons in 'Ethics by Design' from Britain's A Level Algorithm*, GILBERT + TOBIN (Sept. 11, 2020), https://www.gtlaw.com.au/insights/lessons-ethics-design-britains-level-algorithm.

setting, as is criminal justice. Risk assessment tools which overpredict recidivism for Black defendants, and underpredict recidivism for White defendants, lock communities of color into cycles of incarceration and deprivation, further eroding their opportunities for rehabilitation and social change. The forward-looking approach locates the wrongness of algorithmic discrimination in this *foreclosure* of future opportunities, and the entrenchment of existing inequality.

THE RACIALIZATION OF TIME

The ubiquity of risk assessment tools reflects a reluctance to afford certain categories of people the possibility of a future that looks different from their past. Statistically-evidenced propensities treat an individual "as if his present conduct could be inferred from his past conduct; as if he were determined rather than free."[14] This cybernetic loop of datafied activity becomes a self-fulfilling prophecy: the "system watches what you do; it fits you into a pattern; the pattern is then fed back to you in the form of options set by the pattern; the options reinforce the patterns; the cycle begins again."[15] Behavioral patterns identified by algorithms open or close access to resources, and these structural constraints reinforce existing patterns of behavior. Conditioning access to opportunities on the basis of historic data produces a "digital caste system, structured by existing racial inequities."[16] Society is stratified into two groups: on the one hand, those whose access to resources is determined by data-based inferences fed through automated systems, and on the other, the privileged few who are able to shed their data shadows, and enjoy an undetermined future. As we have discussed, COMPAS' asymmetric distribution of classification errors (false positives for Blacks, and false negatives for Whites) splits "the future into two racially distinct times—a (white) time that is futurally open and a (non-white) time that is futurally closed."[17]

Charles Mills argues that the racialization of *space* (through segregation, redlining, and gentrification) has been widely documented, but the racialization of *time* has received substantially less attention.[18] He describes two ways in which time is racialized:

---

[14] David T. Wasserman, *The Morality of Statistical Proof and the Risk of Mistaken Liability*, 13 CARDOZO L. REV. 935, 952 (1991), https://heinonline.org/HOL/LandingPage?handle=hein.journals/cdozo13&div=48.

[15] LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999).

[16] RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE 10 (2019).

[17] Bonnie Sheehey, *Algorithmic Paranoia: The Temporal Governmentality of Predictive Policing*, 21 ETHICS AND INFO. TECH. 49, 49-58 (2019), https://link.springer.com/article/10.1007/s10676-018-9489-x.

[18] Mills, *supra* note 12.

first, in the construction of historical narratives, and secondly, in the management and division of scarce temporal resources. First, the construction of historical narratives involves selective "amnesias, excisions, and forgettings" that produce pasts that are "usable" in the sense that they promote a specific group identity, and trajectory.[19] The White settler state, for example, set the "historical chronometer" at zero in order to erase the indigenous history that preceded it, and to justify the theft of land, and the genocide of native peoples.[20]

Secondly, the scarce resource of time itself – scarce to all mortal peoples – is unevenly accessed and distributed. Mills describes how regimes of "temporal exploitation" *transfer* time "from one set of lives to another," through the segregation of communities guarded by temporal, as well as spatial, walls.[21] For White colonial settlers, for example, political freedom was their birthright. For the African colonies seeking independence, however, freedom would have to wait "a little longer." Imperialism structured time to move quickly "for the extraction of [colonial] capital, resources, and surplus value," but move slowly for the educational and political development of colonized populations.[22] Meanwhile, in the Jim Crow South, racial inequalities of temporality manifested in "unequal temporal access to institutions, goods, services, resources, power, and knowledge."[23] Legalized apartheid imposed a "disjunctive time structure" in which African Americans waited for nearly everything, whether it was a "colored" seat on a segregated bus, or secondhand textbooks passed down from White students years after their content had become obsolete.[24]

Today, racially differentiated access to temporal resources is still apparent in the mass incarceration of young Black men; in the chronically underfunded health care systems that serve communities of color; and in the delayed federal response to Hurricanes Katrina and Maria. It is also apparent, I argue, in the way that predictive algorithms unequally construct future temporality, creating a third dimension to the racialization of time. In overpredicting recidivism for Black defendants, and underpredicting recidivism for Whites, risk assessment tools like COMPAS display a willingness to accommodate a rupture with the criminal past – but only for Whites.

---

[19] *Id.* at 30.
[20] *Id.* at 31.
[21] *Id.* at 28.
[22] Michael Hanchard, *Afro-Modernity: Temporality, Politics, and the African Diaspora*, 11 PUB. CULTURE 245, 245-68(1999), https://read.dukeupress.edu/public-culture/article-abstract/11/1/245/49888/Afro-Modernity-Temporality-Politics-and-the.
[23] *Id.*
[24] *Id.*

For Black defendants, in contrast, the criminal past continues *inevitably*, and algorithmic predictions reflect this assumption.[25] As a result, a potential future world in which a "high-risk" defendant is free, and does not recidivate, is foreclosed to the Black defendant.[26]

THE UTILITY OF INDETERMINACY

We have grown accustomed, over time, to accept incarceration as an appropriate punishment for past crimes, largely due to what we hope is their deterrent effect, namely, that the cost of crime (incarceration) will discourage their future commission. What appears to have generated less consensus, however, is the ethical permissibility of incarcerating individuals for expected *future* criminality. Incarcerating an individual for time *t* in order physically to incapacitate them from anticipated future crime seems unethical for at least five reasons. Intrinsically, it contradicts the presumption of innocence (a fundamental tenet of the criminal justice system), and it denies future indeterminacy to those individuals, arrogantly assuming that their futures can be entirely predicted from the past. Incarceration for future crime also seems *contingently* unethical if it affects certain social groups unequally (disparate impact); if the social costs of such incarceration outweigh the social benefits; and if proponents of such incarceration actively sustain the factual premise for its alleged utility. In short, it seems possible to marshal a variety of legal, economic, moral, and philosophical arguments against the incarceration of individuals in the expectation of future criminality.

THE PHILOSOPHICAL ARGUMENT

An algorithm is not *capable* of imagination; it cannot visualize a horizon beyond the historical data on which it was trained. Therefore, the range of options it "predicts" for the future will necessarily be constrained by the permutations of the past. If a discrete outcome is not observed in the training set, it will not appear within the algorithm's predictions. The infinite potentiality of the future – all of the unprecedented events that could occur – are inaccessible. So, when we rely on algorithms to "predict" the future, we deny ourselves that indeterminacy. A risk assessment tool trained on historical recidivism data will never "predict" a future in which "high-risk" defendants successfully re-integrate in marginalized communities with the help of significant public sector

[25] *See, e.g.*, JURGEN HABERMAS, THE PHILOSOPHICAL DISCOURSES OF MODERNITY: TWELVE LECTURES (1987).

[26] *See, e.g.*, JOHANNES FABIAN, TIME AND THE OTHER: HOW ANTHROPOLOGY MAKES ITS OBJECT (1983).

investments in education, housing, and employment. However, a community organizer might. Humans, unlike algorithms, are capable of imagination; Julius C. Chappelle envisioned a postbellum world in which Blacks lived as long as Whites. It is this capacity to imagine social transformation that should guide our decision-making about the future, not the limited viewfinder of an algorithm.

The manner in which predictive algorithms *erase* the indeterminacy of the future (by "predicting" a narrow range of future outcomes drawn from historical data) is similar to the manner in which automated decision-making eliminates the indeterminacy of legal text. Laurence Diver writes about the "hermeneutic gap" between the textual representation of a legal norm, its interpretation, and its real-world instantiation.[27] A legal text, on its own, cannot compel specific behavior; it must first be read, interpreted, and applied to a specific factual scenario. In this sense, the text affords a temporal delay in which a legal subject can evaluate what the law requires of her. This temporal space facilitates not only contemplation, but also *contestation*; because the law is not self-applying, it yields multiple, competing interpretations of what is required in any given scenario. This affordance of interpretive deliberation, Diver argues, is critical for legal contestation within a democracy. And we have worked hard to get here – to raise population literacy levels, to make texts widely available, to forget the dark days of scribal culture where a handful of texts were jealously guarded and interpreted by religious elites to the illiterate masses.[28] The affordance of interpretability, as Jeremy Waldron has discussed elsewhere, respects legal subjects as rational actors with dignity, agency, and capacity for reasoned action.

Code, in contrast, strips language of its nuance and ambiguity,[29] removing capacity for the kind of moral deliberation that is consistent with freedom of thought, and other forms of respect for autonomy.[30] The immediacy of code *collapses* the hermeneutic gap between a legal text and its instantiation. The application of code is predetermined, and immediately executed (if *x*, then *y*) without consideration of unincorporated conditions.[31] The

---

[27] Laurence Diver, *Computational legalism and the affordance of delay in law*, 1 J. CROSS-DISCIPLINARY RES. COMPUTATIONAL L. (2020), https://journalcrcl.org/crcl/article/view/3/6.
[28] *Id.* at 3.
[29] Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1, 3 (2019), https://www.gwlr.org/wp-content/uploads/2019/01/87-Geo.-Wash.-L.-Rev.-1.pdf.
[30] Seana Valentine Shiffrin, *Inducing Moral Deliberation: On the Occasional Virtues of Fog*, 123 HARVARD L. REV. 1214, 1218 (2010), https://www.jstor.org/stable/40648485?seq=1.
[31] Diver, *supra* note 27.

immutability of code forecloses alternative conduct so that it actively *constitutes*, rather than guides, legal reality.[32] It strives toward perfect execution, rather than flexible interpretation.[33] Automated legal decision-making, then, eliminates the spaces for deliberation and contestation that legal text traditionally affords. The question becomes one of systems design: where, within our labyrinthine legal system, is the affordance of delay important, and where can it be dispensed with? How can we preserve "contingent, evaluative, hermeneutic spaces" within computational architectures that achieve our twin goals of efficiency, and legitimacy?

Just as automated decision-making removes spaces for normative deliberation, I argue that predictive algorithms erase the indeterminacy of the future. Rather than allowing the future to remain open, and unknowable, predictive algorithms construct "future outcomes" from a narrow set of historical events. As the allocation of future resources (including time, and economic opportunity) is increasingly guided by this narrow view of possible futures, the space for moral imagination shrinks. Our vision of the future is not shaped by what *could* be, or what *should* be, but by what *has* been, and with what confidence interval. Just as reflexive reliance on computational legalism risks collapsing the deliberative spaces that legality was built upon,[34] disproportionate faith in algorithmic predictions risks shrinking the imaginative range of future-oriented legal decision-making.

[32] *Id.*
[33] *Id.*
[34] *Id.*