What is the right to individualized judgment?

Renée Jorgensen Bolinger[1]

Legal license to treat an individual in certain ways---subjecting them to special scrutiny, detaining them, ruling that they are liable to penalties---frequently depends on our degree of confidence, given the available evidence, that such treatment would be appropriate. It is increasingly possible for law-enforcement agencies to leverage large datasets to build statistical models of the correlations between traits of interest (e.g. criminal offending) and a wide array of other variables. Suppose such a model predicts that it is highly likely that Alex will reoffend. Would Alex have a moral or justice-based complaint against this being offered as evidence that Alex poses a danger to others?

Some scholars (and courts) have said yes: relying on statistical evidence violates Alex's right to be "treated as an individual". Others protest that this 'right' is unintelligible, and does not justify eschewing the use of evidence that might increase the accuracy of our determinations. The central question for this paper is what this 'right to individualization' is a right *to*, exactly. What interest or important choice does it protect? My hope is that by getting clearer on this point, we can make some progress toward determining whether relying on algorithms at various points in the criminal justice system would run afoul of it. After exploring a few different ways to fill in the foundation and content of the right, I ultimately suggest that it is best understood as protecting agents' entitlement to a fair opportunity to avoid clashes with the law. If so, then the right does not preclude the use of statistical methods per se, but rather requires that it be possible for citizens to anticipate which variables will be used as predictors. Furthermore, it condemns reliance on various indexes of distributive injustice, or unchosen properties, as evidence of (actual or probable) law-breaking.

1. Background

Algorithmic tools have made their way into law-enforcement at various stages, as enthusiasm for data-driven policing and sentencing swells. To name just a few examples: PredPol and HunchLab are used by dozens of police departments to predict hotspots for property crime, assault, and auto theft; in Chicago, the 'Strategic Subjects Initiative' (SSI) leverages past arrest and crime reporting data to identify and rank those most likely to be involved (either as perpetrator or victim) in future violent crime, in order to guide the allocation of police resources; and until recently, New Orleans PD relied on Palantir's

[1] Preliminary workshop draft, comments are welcome! (*bolinger@princeton.edu*). **Please do not cite without permission.** My thanks to participants at the 2019 workshop on the *Democratic Implications of Algorithmic Decision-Making* at Princeton University, and the 2020 Radcliffe Institute Workshop on *The Ethics of Technology at Work and in Public Institutions* at Harvard University, for helpful discussion of an earlier version of this paper.

Gotham data analysis program, and the LAPD used the 'Los Angeles Strategic Extraction and Restoration' (LASER) program the same way.[2] Post-conviction, the Correctional Offender Management for Profiling Alternative Sanctions (COMPAS) is one of several tools used as a risk-assessment measure for sentencing and probation decisions, though it is increasingly used to guide pre-trial decisions about detention and bail.[3]

Tools of this kind trade in statistical and actuarial inference. Their core models are built from inputting a number of variables about each entry in a historical database, and running regressions to model the strength of correlations.[4] Then they can assign scores to new cases based on the strength of correlations between a selected set of variables about the subject and groups with high, medium, or low offending rates. *Static variables* (e.g. age at first arrest, criminal history, etc.) encode elements of a subject's file which do not change over time, while *dynamic variables* (e.g. years since last offense, employment status, drug use) may change over time and can reflect the subject's current behavior. The former perform better as predictors of non-violent offending, and models relying exclusively on static variables outperform tools that are also sensitive to dynamic variables.[5] However, it's worth noting that since offending is measured by *arrest* (or in some cases *conviction)* as a proxy, rather than directly observed, some proportion of these tools' accuracy is just their ability to predict arrest patterns, which are subject to enforcement bias.

---

[2] LAPD suspended their use of LASER after significant public protest. NOPD suspended their contract with Palantir in early 2018, after public backlash at the secrecy of the initial arrangement and terms.

[3] Other well-known fourth-generation risk-assessment tools include Ohio Risk Assessment System (ORAS), the Correctional Assessment and Intervention System (CAIS), and the Level of Service/Case Management Inventory (LS/CMI). These tools are the newest in a long string of measures which assign risk scores to subjects by analysing actuarial information concerning factors like arrest history, educational and employment history, age, gender, and other demographic information. Of these, COMPAS is the only tool explicitly designed to trace changes in a subject's risk score over time. See Herrschaft 2014, '*Evaluating the Reliability and Validity of the Correctional Offender management Profiling for Alternative Sanctions (COMPAS) Tool: Implications for Community Corrections Policy*' Dissertation, Rutgers University; Kehl, Guo, and Kessler 2017, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing' *Responsive Communities Initiative*, Berkman Klein Center for Internet & Society, Harvard Law School.

[4] The different prediction tools vary with respect to which types of data they draw on. PredPol, SSI, and LASER employ a mix of arrest, crime reporting, and conviction data, raising worries that enforcement bias and differing levels of confidence in the police distort the dataset in ways that compromise the fairness of the algorithms. See Richardson, Schultz, and Crawford, 2019, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice' *NYU Law Review* (94):192.

[5] Herrschaft 2014 (supra note 3); Dressel & Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* 2018. Dressell & Farid also found that untrained subjects who were given case files, and asked to make a prediction (without being given any particular instruction as to *how*, outperformed COMPAS both in accuracy (as measured by AUC) and fairness (more even distributions of false negative errors) across demographic groups. But see Degiorgio, & DiDonato 2013, 'Predicting probationer rates of reincarceration using dynamic factors from the Substance Abuse Questionnaire-Adult Probation III (SAQ-Adult Probation III)', *American Journal of Criminal Justice*, 39, 94-108, for findings that *adding* dynamic factors to static demographic models in fact improves the fit of a model predicting probation revocation for substance abuse.

There are some obvious ethical challenges (bias in error rates, data looping, redundant encoding, etc.) that have already been the subject of significant academic and media attention. There are also a handful of epistemic concerns about the probative value of actuarial and statistical inferences, including worries that these systems base predictions on spurious correlations, are lacking in explanatory value, and allow a fixed datapoint in a person's past (like age at first offense) too much weight in predicting whether they will reoffend. I will set all of these aside in order to focus on a third type of concern: whether decisions informed by these methods run afoul of a moral or political duty to treat the subjects of our inferences as *individuals*, reflected in the legal requirement that suspicion be 'individualized'.[6]

Leaving aside programs which focus on predicting *locations* of crimes rather than individuals, the basis for assessments offered by these algorithmic tools is, ultimately, observations about the behavior of groups of *other* people with properties similar to the immediate subject. The reasoning structure is an actuarial inference: it moves from the conjunction of *the subject has feature G* and *the relative frequency of feature F among others with G is* x to confidence of approximately x in *the subject has feature F*. Paradigmatically 'specific' inference, by contrast, takes the observation *the subject has feature G* to provide direct (albeit not certain), probabilistic support for the conclusion *the subject has feature F*.[7] If the right to individualized suspicion excludes reliance on actuarial evidence, as some argue, does it also rule out the use of the algorithmic tools mentioned above?[8] Rather than trace the constitutional grounds or legal interpretation of this right, I want to explore its political grounds: what interest does it protect, and is that interest threatened by relying on the outputs of algorithmic methods as evidence to reach the required degree of credence in determinations of probable cause, guilt, or sentencing?

---

[6] For overviews of the legal interpretation of this right, see Andrew E. Taslitz, Myself Alone: Individualizing Justice through Psychological Character Evidence, 52 *Maryland Law Review*, 1, 3-30 (1993); David A. Harris, Particularized Suspicion, Categorical Judgments: Supreme Court Rhetoric Versus Lower Court Reality Under *Terry v. Ohio*, 72 *St. John's Law Review,* 975 (1998).

[7] Hence, though eyewitness testimony is not entailing evidence of guilt, it falls on the 'specific evidence' side of the division: we infer based on the fact that *the subject closely matches the eyewitness description* that probably *the subject was the person seen by the eyewitness*, with no intermediate step invoking the relative frequencies among a class of other people. I discuss and diagnose differences in the justificatory force of 'individualized' and 'statistical' evidence in other work (see Bolinger, 'Explaining the Justificatory Asymmetry between Statistical and Individualized Evidence', in Hoskins & Robson*, eds., The Social Epistemology of Legal Trials,* forthcoming).

[8] A quite different claim, also introduced as the 'right to an individualized decision', is concerned with the use of algorithms not as applied to *groups of people* in order to predict the *fittingness* of some specific treatment or verdict, but to actually *decide cases* on the total evidence, where the learning data is all pre-existing caselaw (for instance), and the inputs are the facts of a given new case. These applications present very different problems, and for simplicity I will set them aside. For an informative discussion, see Reuben Binns, 'Human Judgement in Algorithmic Loops', *Regulation & Governance* (forthcoming; on file with author).

2. Interpreting the Right to Individualization

Several of the theorists who argue that relying on statistical or actuarial evidence violates the requirement that suspicion be individualized gloss the requirement as arising from a moral duty tied in some way to the values of autonomy and respect. For instance, Duff (1998) anchors his argument against using actuarial evidence to predict whether a subject will reoffend in the claim that "[t]o respect the defendant as a responsible citizen, we must treat him and judge him as an autonomous agent, who determines his own actions in the light of his own values or commitments. His membership of this actuarial group is part of the context of that self-determination; and as observers, we might think it very likely that he will have determined himself as a criminal." Nevertheless "respect for autonomy, and the 'presumption of harmlessness' which follows from it, forbids us to ascribe criminal dangerousness to anyone, unless and until by his own criminal conduct he constitutes himself as having such a character."[9] A number of philosophers and political theorists have argued that a similar entitlement also holds outside the legal domain: that in general, respecting others' moral autonomy prohibits basing our moral appraisals of their character on statistical evidence.[10]

Judges articulate the purpose and value of requiring that evidence be individualized somewhat differently. In an opinion rejecting the use of actuarial evidence for sentencing in *United States v. Shonubi*, Judge Newman emphasized that the 'specific evidence' requirement is only satisfied by "evidence that points specifically to [behavior] for which the defendant is responsible."[11] Justice Stevens, meanwhile, highlighted the rule's protective function for establishing probable cause in his dissent in *Samson v. California*: "The requirement of individualized suspicion, in all its iterations, is the shield the Framers selected to guard against the evils of arbitrary action, caprice, and harassment."[12]

The variety of characterizations given, and grounds offered, for the right to individualization leaves it ambiguous which treatments transgress it. There are a few

---

[9] R.A. Duff, 'Dangerousness and Citizenship', in A. J. Ashworth and M. Wasik (eds.), *Fundamentals of Sentencing Theory* (Oxford University Press) (1998), at p. 155-6.

[10] See, e.g., A. Walen, `A Unified Theory of Detention, with Application to Preventive Detention for Suspected Terrorists' (2011) *70 Maryland L Rev* 871; L. Buchak, 'Norms for Credence and Norms for Belief' (2014); S. Moss, *Probabilistic Knowledge* (Oxford University Press, 2018). Some maintain that we wrong others whenever we use statistics to draw inferences that diminish or would lead us to act against the subject's interest. See e.g. D. Wasserman, 'The Morality of Statistical Proof and the Risk of Mistaken Liability' (1992) 13 *Cardozo Law Review* 935; R. Basu, "The Specter of Normative Conflict: Does Fairness Require Inaccuracy?" in *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*, eds. Erin Beeghly and Alex Madva (New York: Routledge, forthcoming).

[11] Judge Newman, *United States v. Shonubi* (103 F.3d 1085 2d Cir. 1997), at 1089-1090. He also wrote, "The statistical and economic analyses relate to drug trafficking generally and not to Shonubi specifically." at 1091.

[12] *Samson v. California*, 547 U.S. 843, 860 (2006), Stevens, dissenting.

different candidates that are worth working through. A simple reading of Judge Newman's reasoning in *Shonubi* suggests an interpretation that contrasts individualization with generalized treatment; something like

- *A right that high-stakes (or morally charged) decisions be personalized, rather than being subjected to 'one-size-fits-all justice.'*

As Harcourt (2007) stresses, though, relying on statistical or actuarial data allows sentencing determinations to be highly *tailored* to the individual.[13] We can expect this to be at least equally true of the judgments about probable cause or reasonable suspicion made using algorithms, given the very large databases and high number of personalizing variables these methods allow us to take into consideration. But this form of individualization does not seem to address the motivating concern, and does not capture the connection to the individual's *responsibility* that Judge Newman stressed. Personalized treatment also, as Lippert-Rasmussen (2011) points out, is isn't always in our interest, and is in some tension with other intuitive principles of justice, such as the generality and equal application of law, and the fair social distribution of various burdens.[14]

If we take the connection to individual responsibility to be of central importance, we could (echoing Duff) interpret the right as

- *A right to be respected as a presumptively law-abiding citizen, unless and until one defeats this presumption through one's own action and behavior.*

This is reminiscent of Dworkin (1977)'s claim that detaining a person based on actuarial prediction, however accurate, is unjust "because that denies his claim to equal respect as an individual."[15] It is also consistent with the way that Walen (2011) articulates the duty of the state to respect the autonomy of its citizens: "A state must normally accord its autonomous and accountable citizens this presumption [that they are law-abiding] as a matter of basic respect for their autonomous moral agency."[16] And it is consonant with suggestions by Armour (1994), Duff (1998) and Moss (2018) that taking statistical generalizations as reason to conclude that an individual is *probably* dangerous run afoul of the individual's

[13] B. Harcourt, *Against Prediction: Profiling, Policing, and Punishment in an Actuarial Age* (Chicago, IL: University of Chicago Press, 2007)

[14] K. Lippert-Rasmussen, "We are all Different": Statistical Discrimination and the Right to be Treated as an Individual', *Journal of Ethics* (2011) 15:47–59.

[15] R. Dworkin, *Taking Rights Seriously* (1977):13.

[16] A. Walen, 'A Punitive Precondition for Preventative Detention: Lost Status as a Foundation for a Lost Immunity', 48 *San Diego Law Review* 1229 (2011).

moral entitlements.[17] So interpreted, the right seems to require approaching each individual as a novel case, *without* a background expectation that our knowledge of other cases will give us reliable guidance concerning them.

Something about this resonates: the fundamental orientation of law-enforcement toward the members of the political community it serves should not be expressive of suspicion and disrespect. Being viewed as *probably law-abiding* orients police to respect and protect; being viewed as *probably lawbreaking* activates a very different script. There are reasons to doubt that this difference in default orientation is in fact directly or even primarily responsive to the *evidence* whether one is probably lawbreaking, rather than stereotypes or group-based bias.[18] Still, it seems true that citizens are entitled to something like a presumption of 'harmlessness' or good intent; or at the very least, that something like *appropriate civic respect* is central to the right to individualization. But this alone won't explain what is objectionable about the use of actuarial data in *Shonubi*. In sentencing contexts generally, guilt has *already* been established; the presumption has already been defeated by admissible evidence. If the right amounts to nothing more than that we should treat agents as law-abiding until we have adequate particularized evidence that they aren't, then there is no conflict at all between it and the use of algorithmic risk scores in making sentencing determinations. While one could simply accept this conclusion, I am loathe to cede this ground so quickly; it seems instead that the 'presumption of law-abidingness' does not exhaust the obligations grounded in civic respect. What else might it entail? Perhaps

- *A right to not be subject to extra burdens simply on account of one's social identity, or group memberships.*

This is the interpretation naturally suggested by Colyvan et. al (2001), and rejected as unrealistically idealistic by Tillers (2005).[19] The thought that equality of standing, or respect for autonomy, entitles individuals to be free from *extra* scrutiny or being disproportionately subject to the burdens of enforcement is also the central animating idea behind Bambauer (2015)'s explication of why statistical evidence should not be used to establish probable

[17] J. Armour, 'Race Ipsa Loquitur: Of Reasonable Racists, Intelligent Bayesians, and Involuntary Negrophobes', Stanford Law Review, Vol. 46, No. 4 (1994), pp. 781-816; Moss, *Probabilistic Knowledge;* Duff (1998) 'Dangerousness and Citizenship'.

[18] Analysis of transcripts of traffic stops in Oakland, CA found that police officers speak significantly less respectfully to black than to white community members, even after controlling for officer race, infraction severity, stop location, and stop outcome. (Voigt, Camp, Prabhakaran, Hamilton, Hetey, Griffiths, Jurgens, Jurafsky, and Eberhardt, 'Racial disparities in police language', *Proceedings of the National Academy of Sciences*, 2017, 114 (25) 6521-6526.)

[19] M. Colyvan, H. Regan and S. Person, 'Is it a Crime to Belong to a Reference Class?' *The Journal of Political Philosophy* (2001) 9: 2, pp. 168-181; P. Tillers, `If wishes were horses: discursive comments on attempts to prevent individuals from being unfairly burdened by their reference classes', *Law, Probability and Risk* (2005) 4, 33−49.

cause.[20] Something like it is also the centerpiece of Barbara Underwood's (1979) explanation of why racial membership and other protected categories are an inappropriate base for statistical prediction. Underwood pushes the idea further, however, grounding protection against these—and other unalterable features—in concern for autonomy: "Of all the factors that might be used for predictive purposes, those beyond the individual's control present the greatest threat to individual autonomy. Use of such factors in a statistical prediction device is particularly undesirable if the device is to be used in a context in which autonomy is highly valued."[21]

All three of these themes—respect, autonomy, and the need to avoid compounding disadvantage—seem central to understanding the right to an individualized decision.[22] But there is a thoroughly different, more procedural gloss, well-worth considering, which holds that the right is actually a proxy for the right *to an explanation* for the state's treatment:[23]

- *A right to an explanation for the State's exercise of coercive powers.*

As Vredenberg (ms) compellingly argues, the value of explanations of this kind is instrumental: it is a prerequisite for agents' ability to act on the political system, holding it accountable and forming the rules which characterize the basic structure of society.[24] The right so understood requires more than that the subject be given *some* true account or other of why a legal decision concerning her has been made; the explanation offered must equip her to act intentionally to hold the decision-making body accountable. It therefore must bear some relation to the *actual* decision-making procedure employed. This is compatible with the use of algorithms, so long as we make public which properties are being used as variables, and roughly how the predictions are arrived at. In understanding the moral core of the right to individualization as a right to the information necessary to form and reform the legal policies, this gloss aligns closely with Justice Stevens' comment that the right is a shield against arbitrary uses of state power.

Each of these candidate interpretations highlights something of value in the intuitive right to individualization, but, I think, also leaves out something important. The values highlighted—connection to responsible agency, respect for autonomy, concern with compounding disadvantage and constraining the exercise of coercive power—are also

---

[20] J. Bambauer, 'Hassle', *Michigan Law Review* (2015) 113(4) pp. 461-511.

[21] B. Underwood, 'Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment' *Yale Law Journal (*1979) *88:*1408, at p. 1436.

[22] Add ref: Hellman, 'Sex, Causation, and Algorithms', working paper.

[23] This interpretation is implicit in the 'explanationist' strand of the legal literature on statistical evidence, which contend that statistical evidence should be inadmissible in trials because it is inadequately *explanatory*, or supplies probabilistic support without raising the *plausibility* of the hypothesis that the defendant is guilty.

[24] K. Verdenberg, 'The Right to an Explanation', *working paper (*on file with author).

central to the value of the rule of law.[25] So, I propose that we understand the right to an individualized legal decision as an entailment of the requirements that law be public, clear, and prospective; in short:

- *A right to fair opportunity to avoid hostile encounters with law enforcement, entailed by the publicity of law.*[26]

The provisions that law be public and clear secure citizens' ability to know *what* the law forbids; the requirement that law be prospective (rather than retroactive) secures their ability to intentionally act to avoid violating the law. These, together with the other five requirements Fuller (1969) articulates, are meant capture necessary conditions on the law's ability to structure citizens' relationships to each other and the state in a way that expresses respect for their autonomy and equality as agents, and takes lexical priority over considerations of administrative efficiency.[27]

The virtue of the rule of law centers on the ability of law to shape citizens' practical reason and ground reliable expectations; holding them to standards which they had fair opportunity to meet. This is especially important because falling afoul of the law has drastic consequences for an individual's freedom and her ability to pursue her life projects, as well as for her participation in the political community. Such violations expose a person to the coercive power of the state in various forms, ranging from asset forfeiture, to being deprived of liberty and stripped of civic rights. Even when these costs are kept proportionate to the harm the subject imposed on the community, people have a legitimate complaint if they were not afforded fair opportunity to avoid such exposure.

But *actual* law-breaking isn't the only activity that puts agents at the mercy of the state. Given the uncertainty under which law enforcement and criminal justice systems must operate, whatever factors are treated as grounds for reasonable suspicion or evidence of law-breaking also expose subjects to the coercive power of the state. Even if the grounds for suspicion are *personalized*, agents only have fair opportunity to avoid hostile treatment by law enforcement if it is also public and prospective what will be treated as evidence of criminality. This construal of the right encompasses several of the central

---

[25] L. Fuller, *The Morality of Law* (New Haven: Yale University Press, 1969), p. 106. Fuller gives eight requirements for the rule of law: law must be (1) general, (2) publicly accessible, (3) prospective rather than retrospective, (4) clear, (5) non-contradictory, (6) possible to satisfy, (7) stable, and (8) there must be congruence between what the law requires and what is enforced.

[26] Importantly, I do not mean to imply that the interest exists *only* in the domain of legal decisions. The relationships of respect and answerability that law formalizes may extend to informal, interpersonal interactions, and so plausibly the interest protected by a formal right to an individualized decision may persist in as a moral claim in informal contexts. My thanks to Deborah Hellman for discussion on this point.

[27] My analysis in this section draws heavily on Fuller's articulation of the value of the rule of law, particularly as developed and defended by C. Murphy, `Lon Fuller and the Moral Value of the Rule of Law', *Law and Philosophy*, Vol. 24, No. 3 (May, 2005), pp. 239-262.

themes from above. It requires that agents of the state default to respectful engagement, absent specific evidence of law-breaking, and that disproportionate burden or suspicion should be responsive to the individual's *responsible action*. It also grounds the right ultimately in the preconditions for the laws to be fair.

3. Is the use of algorithmic tools consistent with the right to individualization?

Understood this way, the right to an individualized decision does not reduce to due process, or the ability to challenge the evidence on which one is sentenced. Rather, it aims to ensure that criminal justice procedures are consistent with the conditions on the legitimacy of the state's exercise of coercive power over its citizens, and are expressive of respect for the citizens' autonomy.[28] The right does not necessarily rule out using statistical models or algorithms to guide legal decision-making, but it does require that models be constructed in a way that affords subjects a fair opportunity to avoid receiving a high risk-score (or coming under suspicion).

What does it take to secure 'fair opportunity' of this kind? Minimally, it must be in-principle possible for an individual of any permissible social identity to avoid coming under suspicion. Consequently, a high risk-score can't be tied to unchangeable identity-tracking properties (e.g. race, gender, build). But mere in-principle avoidability is not sufficient; the properties used to indicate criminality must also be features that the law-abiding agents can *in fact* act to avoid. So, more controversially: they shouldn't be tied to things that subjects have little real chance of escaping, like residence in high-crime neighborhoods, poor educational background, or family environment.[29]

The rule of law expresses respect for subjects' autonomy only when it enables citizens to anticipate how the state will act, and what it requires them to do. So, in addition to the *brute ability* to avoid properties that would lead to having a high risk-score, subjects must also be able to *act intentionally to avoid* them—which means they need to able to know

[28] Consequently, the grounds offered in *Loomis v. Wisconsin* are, on this view, not the ones which explain what is problematic about using the risk-scores given by COMPAS to make sentencing determinations.

[29] I am primarily focused on a notion of 'fair opportunity' that is non-comparative, demanding simply a normatively sufficient chance of avoidance. But a comparative conception of fairness is also relevant here, requiring that a subject have *not substantially worse chances* of avoidance than others in the political community (I am indebted to comments from Chad Lee-Stronach for this point). While both notions are difficult to make precise, the latter intersects with a dilemma arising from antecedent distributive injustice: children who grow up in concentrated urban poverty do not have prospects of avoiding criminality comparable (or even close) to those with different social starting positions. (For a discussion of this dilemma, see especially B. Ewing, 'Recent Work on Punishment and Criminogenic Disadvantage' *Law and Philosophy* (2018) 37: 29–68; Howard, 'Moral Subversion and Structural Entrapment', *Journal of Political Philosophy* 24(1) (2016): 24–46; H. Kim, 'Entrapment, Culpability, and Legitimacy' *Law and Philosophy* (forthcoming); Shelby, 'Justice, Deviance, and the Dark Ghetto', *Philosophy & Public Affairs* 35(2) (2007): 126–160; and Watson, 'A Moral Predicament in the Criminal Law', *Inquiry* 58(2) (2015): 168–188.)

which variables are used, and roughly how. Note that it does *not* require full avoidance of any use of statistics for predictive purposes. But it does require publicity not only of what the *laws* are, but also of what will be used to predict or establish probable law-breaking, so that agents can act intentionally to avoid falling under suspicion. This sort of transparency is also important for auditing error rates, and enabling external agents to hold law-enforcement accountable.

4. Two worries about algorithmic transparency

We might have two different concerns about ensuring this level of transparency. First, a concern that the predictive algorithms could simply be *too complex to understand, let alone explain;* a particularly salient worry for applications of unsupervised deep machine learning. As it happens, it is unlikely that either high-dimensional models or deep learning methods will be necessary—or even much help—for optimizing the algorithms' predictive accuracy in the context of criminal law. Though proprietary protections prevent direct confirmation, by best guesses the algorithm used by COMPAS is not much more complex than the second-generation risk-measures developed in the 1970s: basically a simple linear predictor.[30] This is unsurprising: while great advances have recently been made in recognition tasks, machine-intelligence has yet to stably out-perform simple rules at predicting social outcomes, and consistently plateaus around 65-70% accuracy.[31]

Furthermore, criminal law is an area where it's especially treacherous to use extant databases (e.g. requests for service, crime reports, arrests, or convictions) as training data. The recorded data is invisibly shaped by administrative discretion, as well as upstream structural injustices that artificially forced overlap between communities of color and criminogenic conditions (especially underfunded schools and depressed economic conditions). Machine learning run on this data will pick up and reflect correlations that are causally spurious but genuinely "there" in the data, projecting these traces of past injustice forward.[32]

Bracketing these concerns about training data, and *even if* the predictions made were highly accurate, the right to individualization as I have interpreted it may more directly preclude certain applications of machine-learning in developing the algorithms. A learning

---

[30] Dressel and Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* 4, eaao5580 (2018):3. "Despite using only 7 features as input, a standard linear predictor yields similar results to COMPAS's predictor with 137 features. We can reasonably conclude that COMPAS is using nothing more sophisticated than a linear predictor or its equivalent."

[31] A. Narayanan, 'How to recognize AI snake oil' (2019 unpublished manuscript); M. Yang, S. C. Wong, J. Coid, The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychol. Bull.* 136, 740–767 (2010).

[32] For more thorough articulation and three detailed case studies of dirty data being used to train the models for predictive policing software, see Richardson, Schultz, and Crawford.

method which bases the risk prediction on correlations that *emerge* between very large numbers of variables and the outcomes is backward-looking and opaque. Insofar as it finds unexpected or surprising relationships, and bases new verdicts on these, it tends toward retroactivity, imbuing properties that had been considered harmless with criminal significance after the fact. If we cannot anticipate which properties will yield a high risk-score, then we cannot satisfy the requirement to be *prospective*.

The second concern arising from requiring transparency is that it will make the algorithms vulnerable to strategic gaming.[33] First, let's get clear on the assumptions behind this objection. There are very particular conditions under which strategic gaming is problematic: the proxy criteria only weakly or contingently correlate with the target criteria, the proxy properties are within subjects' deliberate control (they are alterable), the tradeoff costs of gaming the proxy criteria are low, and moreover this can be done affecting the subject's true eligibility with respect to the target criteria.[34] If any one of these conditions is not met, then either a subject's attempt to game the proxy will *also* change how they fare with respect to the target, or the difficulty or costs involved in strategic gaming will offset the incentive.[35] There is also a potential social upside to enabling strategic behavior: it allows people who are over-represented in the false-positives to avoid the proxy criteria, or contest the decision, and so lower the relevant error rate. When the costs of a false-positive are high, the error-correcting tendencies of publicity can outweigh the costs of strategic gaming.

There may be many administrative decisions for which it is permissible to use secret proxies satisfying all these conditions, but I contend that the administration of criminal law is not one of them. The decision to use a given property as a proxy for criminality imposes significant costs—at *least* high risk of 'hassle factor' (the costs associated with being subjected to extra scrutiny), at worst high risk of unwarranted punishment—on the bearers of the proxy properties. With few exceptions (possibly for tax evasion and fraud), attaching these costs to secret proxies cannot be made consistent with the core values of public law.

If the secret proxy criteria are only weakly connected with criminality, then the state cannot justify its selection by appeal to the harm principle, necessity, or the decision's having been ratified by a democratic decision-making process. If, in addition, reliance on

---

[33] For much of the following discussion, I am indebted to immensely helpful conversations with Katie Creel.

[34] I've drawn these conditions for problematic strategic gaming from Cofone & Strandberg, 'Strategic Games and Algorithmic Transparency' (conference draft, on file with author).

[35] For instance, LSAT scores are an oft-used proxy for the facility of reasoning needed for success in law school (the target criteria). But they are also robustly connected to the target, such that students who strategically aim only to improve their LSATs—enrolling in test-prep courses and practicing critical reasoning skills—thereby also make themselves better candidates with respect to the target criteria.

the proxy concentrates false-positives disproportionately on an already disadvantaged subpopulation, members of that group have a dual complaint against secrecy of the proxy: one grounded in autonomy, and one in political equality.[36] If we assume instead that the proxy is robustly connected to criminality, then publicity is net-beneficial: it incentivizes those for whom the proxy correctly indicates criminal behavior to avoid crime-adjacent activity, or to take up behaviors that make it less likely they will commit a crime; and it equips those for whom the proxy was misleading to avoid the proxy, thus reducing the false-positive error rate.

It is precisely this incentivizing effect that prompted Underwood to comment that, when selecting factors to use as predictors, "[r]espect for autonomy thus counsels not only against the use of uncontrollable factors, but also against the use of those controllable factors that involve behavior generally regarded as private and protected against official interference."[37] Law enforcement is fundamentally different in its orientation than some other applications of predictive algorithms. The law does not—must not—aim to detect 'social cancers' even before they manifest. It instead functions to set expectations for behavior, using the coercive apparatus only to hold agents accountable to these very expectations; the impulse toward secrecy must be resisted, and predictions based only on factors that are sufficiently avoidable and not core to the agents' valuable exercise of autonomy in their private lives.

5. Upshots

On the interpretation that I have offered, when legal decisions are made in ways that do not afford subjects a fair opportunity to avoid hostile encounters with law enforcement, this constitutes a failure of the rule of law. Ensuring that citizens have such fair opportunities strongly constrains which variables can be used as a basis for suspicion or prediction. It rules out reliance on static factors outside the deliberate control of subjects--like age, gender, race--as well as a number of indexes of disadvantage (zip code or neighborhood, income level, previous police contact, number of acquaintances with police contacts or arrest records, education level). The former are in-principle unavoidable. Using the latter to justify the imposition of yet more costs on the residents, this time in the form of increased risk of suffering unjustified state coercion, is patently unfair.

---

[36] It's possible to leverage this point to suggest that while secrecy is inappropriate for criteria of suspicion in street crimes, it is more permissible for the enforcement of white-collar crime, since the populations bearing the extra burden of enforcement might be expected to either not be a stable subgroup of the population, or not one that is already disadvantaged.

[37] B. Underwood, 'Law and the Crystal Ball' (1979), at p. 1438.

While risk-assessment or crime-prediction algorithms *could* in principle be designed to be independent of these variables, it is at best unclear what evidential or predictive value they would have. Of the extant risk-prediction tools, those that conditionalize on static variables alone outperform those that *also* incorporate dynamic variables; we can expect that both would outperform prediction based only on the small subset of dynamic variables that are not ruled out by the considerations just raised. So while the use of an algorithm to estimate risk of offending may be consistent with the moral interests protected by a 'right to individualized evidence', it is unclear whether such permissible predictors will have significant evidential value.