Algorithms and the Individual in Criminal Law

*Renée Jorgensen Bolinger, Princeton University (*bolinger@princeton.edu*)*

Draft as of Jan, 2021

1 Background

Legal license to treat an individual in certain ways—subjecting them to special scrutiny, detaining them, or ruling that they are liable to penalties—often depends on our having a sufficiently high degree of confidence, given the available evidence, that such treatment would be appropriate. Sometimes when making these determinations, an individual police officer or judge relies exclusively on their own observations and evaluation of the available information; sometimes they consult or defer to the judgment of an expert. Given the high-stakes nature of these decisions, someone attentive to vulnerability of human decision-making to cognitive biases of all sorts—not to mention the ways prejudice can shape and distort legal determinations—might hope to improve these decisions by outsourcing a bit: introducing the hard numbers, comforting uniformity, and impartiality of algorithmic predictions based on large datasets.[1] Rather than relying on a single agent's personal assessments of whether a particular suspect is likely to reoffend, for example, we might hope to leverage historical trends in re-arrest or re-conviction data to yield some objective measure of the risk, insulated from individual irrationality or animus.

This, in its most optimistic frame, was the driving promise and aim of the risk assessment tools first developed to aid parole decisions in Chicago in 1933. The original model leveraged factors like marital status, behavioral infractions within the detention facility, prior arrest record, and work history to sort inmates into nine rough cohorts, and furnished parole boards with the relative frequency of re-arrest among past members of an offender's cohort, as an indicator of the probability that inmate would reoffend.[2] Since then statistical 'Risk and Needs Assessment' (RNA) tools have been refined and proliferated; they are now used in the majority of jurisdictions in the United States to guide pre-trial decisions relating to whether (and how high)

[1] Suggestions of this kind are made in (J. Kleinberg, S. Ludwig, et al. 2020), and (Taslitz 2010).

[2] (Burgess 1936-7).

to set bail, as well as post-trial determinations concerning whether to divert a defendant from incarceration and when to consider them for parole.[3]

Though automated, most of these tools employ straightforward statistical analysis on entries in historical arrest databases, seeking to isolate the strongest correlations between a set of recorded variables and a property representing the target outcome (e.g. failure to appear, another arrest, or arrest for *violent* offense). There is some variation in the variables used: 'third generation' risk assessment measures improve on the original 'second generation' models[4] by using not only *static variables*—properties that do not change over time, like age at first arrest, having a prior conviction, gender, etc.—but also *dynamic variables* (e.g. years since last offense, employment status, present substance abuse) which are responsive to the subject's current behavior, and can reflect reduced (as well as increased) risk over time. Simplifying a bit, these tools are ultimately algorithms taking the variable values as inputs, assigning them weights, and outputting an estimate of how often someone with those features would end up with the target outcome.

There is now a new wave of tools with a somewhat different structure, more aptly described as applications of artificial intelligence. They leverage a wide array of information in vast databases to *train* a model to recognize patterns in the dataset, in order to predict the outcome value for a new entry. This method allows the trained model to *discover* previously unnoticed correlations between the target outcomes and properties in the dataset; the hazard is that they might be artifacts of the particular dataset, rather than robust connections in the underlying phenomena.[5] As these machine-learning techniques improve, programs using them have become increasingly effective at a wide variety of recognition and classification tasks, and have been pressed into service for a range of prediction tasks, too.[6] The allure of these tools is that they offer a chance not just to make an informed guess about how often an outcome will occur in some set, but to identify and intervene in a case *before the predicted outcome occurs or becomes acute*: removing pre-cancerous tumors, connecting struggling students with extra resources, or, in the case of criminal justice, *preventing* a predicted victimization.

[3] The most commonly used tools include the Arnold Public Safety Assessment (PSA), the Virginia Pretrial Risk Assessment Instrument (VPRAI), the Ohio Risk Assessment System (ORAS), the Correctional Assessment and Intervention System (CAIS), and the Level of Service/Case Management Inventory (LS/CMI).

[4] 'First generation' risk assessment refers to the unstructured clinical assessment, often based on an interview with the subject, that pre-dated the widespread use of the actuarial tools.

[5] There is another hazard in the context of criminal justice, which I will discuss later—that there will be robust connections *which are themselves unjust*.

[6] Many of which are discussed and criticized in (O'Neil 2016), (Eubanks 2018), and (Brayne 2021).

Buoyed by enthusiasm for data-driven policing and sentencing, both sorts of tools have made their way into law-enforcement at several points. To name just a few examples: PredPol and HunchLab are used by police departments across the United States to identify hotspots for property crime, assault, and auto theft. Several states use Palantir's data analysis program *Gotham* to leverage information aggregated from service and arrest databases in order to guide the allocation of police resources and aid in suspect identification and profiling. These include at least the Chicago Police Department's 'Strategic Subjects Initiative' (SSI), LAPD's 'Los Angeles Strategic Extraction and Restoration' (LASER) program, and the Northern California Regional Intelligence Center. Police departments in New Orleans and New York city had similar contracts.[7] On the post-arrest side of things, the Correctional Offender Management for Profiling Alternative Sanctions (COMPAS) program, developed in 2002, is used both to make pre-trial determinations about bail and post-trial determinations about sentencing and parole throughout Michigan, Wyoming, Wisconsin, California, and in an ever-increasing number of counties in other states.[8]

Before being swept away in enthusiasm for the objectivity, equity, or efficiency promised by these algorithmic tools, it's worth asking whether the various uses to which they are or may be put are consistent with fundamental values the criminal justice system must protect. As Attorney General Eric Holder urged,

> "Although these measures were crafted with the best of intentions, I am concerned that they may inadvertently undermine our efforts to ensure *individualized and equal justice*. By basing sentencing decisions on static factors and immutable characteristics – like the defendant's education level, socioeconomic background, or neighborhood—they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society. (Holder 2014) emphasis added.

Some of the obvious ethical concerns about using algorithms in criminal justice---bias in error rates, data looping, redundant encoding, etc.---have already been the subject of significant academic and media attention. COMPAS in particular has drawn substantial criticism for having racially biased error rates.[9] There are also a handful of epistemic concerns that have been raised

[7] 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. Palantir had confidential contracts with a number of city police departments, including NYPD, and it is unclear how many are ongoing. Other prominent clients in the United States include the Central Intelligence Agency, the Department of Homeland Security, Immigration and Customs Enforcement, Department of Health and Human Services, and the Center for Disease Control.

[8] See (Herrschaft 2014), (Kehl, Guo and Kessler 2017). COMPAS was developed and is managed by a private company (Northpointe), which re-named itself 'Equivant' in 2017.

[9] While COMPAS is equally likely to misclassify defendants of any race *in one way or another*, it is disproportionately likely to misclassify a Black defendant as high risk, and disproportionately likely to misclassify a

about the probative value of risk scores, including worries that these systems base predictions on spurious correlations, are lacking in explanatory value, and give fixed datapoints from a person's past too much weight to be epistemically reliable in predicting whether they will reoffend.[10] In this paper, I will set all of these aside in order to focus on a third type of concern visible in Holder's remarks: whether pursuing criminal justice with these tools is consistent with treating the defendant as an individual, and applying sanctions only for *what someone does* rather than for *who they are*.

The reason for this narrow focus is straightforward: while they are serious, the problems mentioned above are mostly problems with using algorithms *badly*. But no matter how we clean, debias, or supplement, all the tools in question trade in actuarial inference: the score assigned to an individual does not reflect a deep insight into his internal propensities to act in the predicted way. It simply reports the frequency of such action among people in the database who are similar to him with respect to the values of the set of variables used as predictors. Leaving aside programs focused on predicting *locations* of crimes rather than individuals, the basis for risk predictions made 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 actuarial in that 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*.

In a slightly different context—addressing the use of statistical or probabilistic evidence to establish liability in civil trials, or settle sentencing questions in criminal trials—several legal theorists, philosophers, and judges have objected that inferences of this kind violate the right to be 'treated as an individual', or would functionally make it a 'crime to belong to a reference class.'[11] In those discussions, the problematic inferences are contrasted with 'specific' or 'individualized' inference, in which an observation of the form *the subject has feature G* provides direct (albeit probabilistic) support for the conclusion *the subject has feature F*. If the right to be treated as an individual or the right to individualized suspicion excludes reliance on actuarial

white defendant as low risk. See (Angwin, et al. 2016). Angwin *et al.* focused their analysis on predictions for people arrested in Broward County, Florida, between 2013-2014.

[10] 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). See also (Bolinger, Explaining Justificatory Asymmetries between Statistical and Individualized Evidence 2021), arguing that statistical evidence in general makes at most a marginal contribution to justified credences, while introducing a risk of error concentrated on particular demographic groups, and so over the long run, the evidential value of relying on it is outweighed by the moral cost of so doing.

[11] See, e.g., (Colyvan, Regan and Person 2001), (Enoch, Spectre and Fisher 2012), (Risinger 2004), (Thomson 1986), and (Tribe 1971).

inference, as they argue it does, we might reasonably worry that it also forbids the use of any of the algorithmic tools mentioned above.[12] If so, this would be a problem not just with using the tools badly, but with using them *at all*. If the form of inference *itself* violates an important procedural right, then the only question left is whether any of the potential applications are not covered by the right. It isn't immediately obvious what the right to be treated as an individual requires, or precludes, though, because it isn't clear what it is a right *to*, exactly.

Rather than trace the constitutional grounds or legal interpretation of this right, my project in this paper is to explore its core: what moral interests might it protect, and are those interests threatened by relying on the outputs of algorithmic methods in determinations of probable cause, guilt, or sentencing? After exploring a few different ways to fill in the right (in §2-3), I ultimately propose understanding it as protecting agents' claims to a fair distribution of the burdens and benefits of the rule of law. I will not so much argue as suggest that the right to individualized treatment should be interpreted as requiring us to respect the separateness of persons, rather than as prohibiting the use of generalizations or statistical methods in inference. What it forbids is not the use of probabilistic information, but taking wrongdoing by some to justify the imposition of extra costs on others. This has significant implications for the administration of criminal justice (explored in §4-5): it permits the use of predictive tools in principle, but requires that subjects be able to anticipate which variables will be used as predictors. Furthermore, it condemns relying on various indexes of distributive injustice, or unchosen properties, to determine whom to subject to extra costs associated with criminal justice.

I will where possible treat all the applications together, but it will be helpful to have a sense of the variety of uses to which algorithmic prediction tools are put. So, very roughly: some of the tools mentioned are used to guide the application of pre-arrest scrutiny, establishing *probable cause* to subject someone to additional search or surveillance, arrest, or detention. For this purpose, what needs to be established is that the *balance of probabilities* suggest the subject is offending. At the next stage, a risk assessment might be offered to support bringing charges, or as evidence about whether the subject will either reoffend or fail to appear if released on bail before trial. In theory (if not in practice) profile evidence could be offered at trial as evidence helping to establish beyond a reasonable doubt that the defendant is guilty of the crimes with which he is charged.[13] Post-conviction, it could be offered at the time of sentencing to indicate the probability

[12] A quite different claim, also referred to 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 (Binns forthcoming).

[13] For an overview of present uses of profile evidence for probable cause and at trial, see (Harris 1998).

of re-offending if given a short sentence, or fitness for diversion into non-carceral forms of punishment (e.g. probation, supportive housing or mental health assistance programs). Finally, it can be used at any stage within the sentence duration to determine eligibility for parole.

2 Interpreting the Right

A natural place to start unpacking the right to individualized treatment is reviewing how judges have discussed the purpose and value of the requirement. Justice Stevens 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."[14] 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."[15] A flat-footed reading of these comments might yield an interpretation contrasting individualization with generalized treatment, leading us to interpret the right as something like

- *A claim that high-stakes legal 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 our determinations to be highly *tailored* to the individual. For instance, rather than having broad sentencing categories, these methods allow us to key the sentence to the strength of the correlation between the subject's specific features and re-arrest or re-conviction. 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, as (Lippert-Rasmussen 2011) points out, personalized treatment isn't always in our interest—it may lead to our being treated

[14] In *Samson v. California*, 547 U.S. 843, 860 (2006), the US Supreme Court (voting 6-3) affirmed the decision of the California Court of Appeal, that it does not violate the fourth amendment protections against unreasonable search to subject parolees to suspicionless search because it is a condition of parole that one consents to search by an officer with or without cause or search warrant. Justice Stevens authored the dissent, joined by Souter and Breyer.

[15] *United States v. Shonubi* (103 F.3d 1085 2d Cir. 1997), at 1089-1090. The defendant (Charles Shonubi) was convicted of smuggling heroin into the United States on eight separate occasions. At the original sentencing, the court multiplied the volume of heroin he was found carrying by 8, to estimate the total quantity smuggled across all his trips. Shonubi appealed on the grounds that the total amount hadn't been proved with specific evidence, and the case was sent back for resentencing. The prosecution then used statistics about drug seizures using the same method, arrested at the same airport, to estimate the total amount; Shonubi appealed again, and Judge Newman again returned the case for resentencing, explaining that "The statistical and economic analyses relate to drug trafficking generally and not to Shonubi specifically." at 1091.

worse than otherwise—and is in some tension with other weighty principles of justice, such as the generality and equal application of law, and the fair social distribution of various burdens.

More urgently, this form of individualization does not capture the connection to the individual's *responsibility* stressed in Judge Newman's comments. The problem is not that statistical determinations are inadequately *tailored*, but that they are not appropriately *responsive*: they treat the individual according to how we expect him to act based on our experience with others like him, rather than how he himself has acted. Personalization is not the point.

Moral theorists prefer to characterize the relevant obligation as a duty to be responsive to the individual's *responsible agency,* grounded in the values of autonomy or respect. For instance, (Dworkin 1977) contends that detaining a person based on actuarial prediction, however accurate, is unjust "because that denies his claim to equal respect as an individual." (Duff 1998, 155-6) also anchors the claim in respect, holding 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." (Walen 2011) articulates the content of the state's duty to respect the autonomy of its citizens in much the same way: "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." This is consonant with suggestions by (Amour 1994), (Duff 1998), and (Moss 2018) that in general taking statistical generalizations as reason to conclude that an individual is *probably* dangerous runs afoul of the individual's moral claims. A number of philosophers and political theorists have argued that something similar holds outside the legal domain: that in general, respecting others' moral autonomy prohibits basing our moral appraisals of their character on statistical evidence.[16]

But does treating individuals with appropriate respect require approaching them as a completely novel case, without an expectation that our knowledge of other cases will give us reliable guidance concerning them? One might be skeptical whether *merely predicting* that an individual is likely to offend is really a failure of respect or affront to their autonomy.[17] It's a bit delicate how to proceed here; some do argue that viewing someone as predictable in this way fails

[16] See, e.g., (Walen, A Unified Theory of Detention, with Application to Preventative Detention for Suspected Terrorists 2011), (Buchak 2014), (Moss, Probabilistic Knowledge 2018). Some maintain that we wrong others specifically when we use statistics to draw inferences that diminish the subject or would lead us to act against their interest. See e.g. (Basu, The Wrongs of Racist Belief 2019), (Schroeder 2018), (Wasserman 1992).

[17] My thanks to Patrick Tomlin, who prompted me to develop this point further.

to regard them appropriately as an agent, rather than a thing determined by external pressures.[18] But we needn't even embrace anything this strong to identify a moral problem with relying on actuarial evidence when harmfully interfering with a person. Plausibly, to justify this kind of act we must not only have reasonably high credence (above some threshold) that the action is morally appropriate, but this credence must be *resilient*. Very roughly: our present evidence must constrain our credences such that little if any new information consistent with it would cause our confidence to dip below that threshold.[19] The more harmful the interference, the more resilient the credence must be to justify it. On this interpretation, even if making a prediction based on statistics is not a failure of respect for the individual's agential *freedom*, using that prediction as grounds for harming them is a failure of respect for their agential *status*, because unless supplemented, statistical evidence cannot be adequately resilient to justify harmful interference.

At a bare minimum, civic respect and equality requires that the *default* orientation of law-enforcement to any member of the political community not be one of expressive of suspicion or disrespect. Considering a person to be *probably law-abiding* orients police to respect and protect them; considering them to be *probably lawbreaking* activates a very different script. There are reasons to doubt that in practice this difference in default orientation is primarily responsive to the *evidence* whether one is probably lawbreaking, rather than stereotypes or group-based prejudice.[20] But even if it tracked group-level rates of arrest, it seems obvious that it would fail to treat citizens as they are entitled: in the absence of specific evidence to the contrary, law-enforcement should treat a person as *probably law-abiding*. Plausibly, generalizations about 'types of people' or trends in broad demographic categories aren't sufficient to justify suspending this civic respect. Borrowing heavily from Duff's language, we might articulate these as

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

The central role this gives to respect and autonomy seems on the right track, and to capture much of the intuitive moral core of the demand that treatment be individualized. But it doesn't explain what is objectionable about the actuarial inferences in *Shonubi*. In sentencing contexts generally, guilt has *already* been established; the presumption has already been defeated by admissible, individualized evidence. If the right requires nothing more than that we treat agents as law-abiding until we have adequate particularized evidence that they aren't, then there is no

[18] See, e.g, (Marušić and White 2018), (Basu, What We Epistemically Owe to Each Other 2019), (Duff 1998).

[19] For a better discussion of resilience, see (Joyce 2005), (Buchak 2014), and (Moss, Probabilistic Knowledge 2018). For an argument that the resilience requirement explains the intuitive justificatory limits of statistical generalizations, see (Bolinger, The Rational Impermissibility of Accepting (some) Racial Generalizations 2020), and (Bolinger, Explaining Justificatory Asymmetries between Statistical and Individualized Evidence 2021).

[20] 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, et al. 2017)

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 claim 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, Regan and Person 2001), and rejected as unrealistically idealistic by (Tillers 2005). There are two ways to develop the thought that equality of standing or respect entitles individuals to be free from *extra* burden, and I find both plausible. On the one hand, we might be concerned about being subject to *disproportionate* burdens associated with law-enforcement, relative to other groups; this is the central animating idea behind (Bambauer 2015)'s explication of why statistical evidence should not be used to establish probable cause. On the other hand, we might worry about being subject to burden that they would not be subject to were we to hold all else *except their social identity* fixed; something like this the centerpiece of (Underwood 1979)'s explanation of why racial membership and other protected categories are an inappropriate base for statistical prediction. She grounds protection against the use of these and other unalterable features in the value of 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." (Underwood 1979, 1436) This emphasis on preserving the individual's *control* gives us reason to object to adding burdens to social identities that aren't themselves protected, but are either unchosen (e.g. socio-economic status), or reflect important personal choices (e.g. marital status).

There is also a more procedural gloss of the right which is well-worth considering: that it is actually a proxy for the right to a certain kind of *explanation* for the state's decisions in her particular case.[21]

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

(Vredenberg working paper) compellingly argues that the value of explanations of this kind is instrumental. Access to such explanations is a prerequisite for agents' ability to act on the political system, to hold it accountable, and form the rules which characterize the basic structure of society. The right so understood requires both more and less than that the subject be given a true account of why a legal decision concerning her has been made; the explanation offered must

[21] 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.

equip her to act intentionally to hold the decision-making body accountable. It therefore must both be intelligible to her, and 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 predictively relevant 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 interpretations highlights something of value in the intuitive content of a right to be treated as an individual, but, I think, also leaves out something important. Rather than offer a competing interpretation of 'individualization', I'd like to suggest that in fact the right doesn't protect a *unique* interest—rather, the interest underlying the right to be treated as an individual is

- *A claim to fair distribution of the benefits and burdens of public law.*

The crowning virtue of the rule of public law is its ability to shape citizens' practical reason and ground reliable expectations, enabling them to hold each other to standards which they had fair opportunity to meet. The benefits of the rule of law include many of the values articulated by the interpretations we've surveyed: expressing respect for autonomy, constraining the exercise of coercive power, ensuring that sanctions are responsive to responsible agency, and ensuring that those subject to law are in a position to challenge or reform it. The burdens, meanwhile, are the various costs associated with the scrutiny and punitive sanctions applied in the course of enforcing the laws.

So interpreted, the right to be treated as an individual requires that a person face disproportionate burden or suspicion only as a consequence of their *responsible action,* and that agents of the state default to respectful engagement otherwise. It also grounds the right ultimately in the preconditions for laws to be *fair*, both in their content and administration. What fair distribution of burdens and benefits demands depends on context: pre-conviction, every individual must have fair opportunity to avoid hostile encounters with law enforcement; at trial they must not face an unfair likelihood of false conviction; post-conviction they must not be subject to disproportionate punishment.

To afford all individuals with a fair opportunity to avoid hostile encounters with law enforcement, laws must be public, clear, and prospective.[22] These are 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. Citizens can't know *what* the law forbids if the

[22] 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).

laws are secret or inscrutable; if the laws are retroactive, they cannot act intentionally to avoid violating it.[23] Importantly, these requirements take lexical priority over considerations of administrative efficiency: the *value* of the rule of law is lost if the laws are applied in ways that do not facilitate the mutual accountability of citizens and state. Choices about the administration of public law has drastic implications for individuals' freedom, ability to pursue their life projects, and participation in the political community. Legal transgressions 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. General appeals to the value of efficient crime reduction—reducing the likelihood of suffering a private infringement of one's rights—cannot justify or compensate a person for the loss of crucial protections against suffering the State's coercive imposition of these harms.

This lens is both unifying and clarifying: it explains why each of the earlier glosses feels partly—but only partly—right. And unlike the other interpretations, which identify relatively all-purpose goods or moral interests, this reading of the right gives it a content specific to criminal law.[24] I suggest, then, that the right issues an injunction not against the use of probabilistic information or generalizations, but against a familiar form of moral aggregation. Just as invocations of the separateness of persons are in other contexts made to assert that benefits to some cannot offset harms to others, the demand that we treat people *as individuals* here asserts that wrongdoing by some does not weaken others' moral claim against the imposition of extra costs.

3 The Demands of Individual Treatment

Accepting this interpretation has relatively revisionary implications for how the right constrains the administration of criminal law. It does not *necessarily* preclude using statistical models or predictive algorithms, but it does require that the chosen procedures not subject any subgroup of the population to unfair burdens. Burdens can be unfair in an absolute sense if they are especially high, or particularly difficult to escape. They can also be unfair in a comparative sense, if they are substantially higher or more inescapable for members of the subgroup than for others.

[23] My analysis in this section strongly echoes Fuller's articulation of the value of the rule of law, particularly as developed and defended by (Murphy 2005). (Fuller 1969, 106) 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.

[24] My thanks to Tom Parr for pointing this out. Importantly, I do not mean to imply that we have a moral interest in being treated as an individual *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.

We have said that to avoid subjecting anyone to more than their fair share of burden, legal sanctions must be tied to responsible agency. It follows that it must at least be *in-principle* possible for an individual of any permissible social identity to avoid suffering a downside cost that is not born by everyone. t So if a high risk-score suffices to justify extra scrutiny, conviction, or a lengthier sentence, the predictor variables can't be tied to unchangeable identity-tracking properties like race or gender. But mere in-principle avoidability is not sufficient; the properties used to indicate criminality—and thus to determine distribution of costly legal sanctions—must also be ones 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 unstable family environment.[25] This is not to deny that we may find robust correlations between these features and criminal offending rates, especially historically—it is simply to assert that it would violate the right to individual treatment to leverage such correlations to justify the predictive application of criminal sanctions.

While I am principally focused on discussing the limitations the right imposes on algorithmic tools, it is worth noting that it has implications for non-predictive administrative decisions, too. Consider the practice of cash bail: allowing individuals to avoid pre-trial detention conditional on paying a sizable fee (typically $10,000), refunded if they appear on their scheduled court date. The immediate effect of this practice is to ensure that the most severe burdens of an encounter with the law—lengthy pre-trial detention, during which the defendant incurs a variety of costs often including job-loss—fall disproportionately on the very poor.[26] The median yearly income for people who are detained pre-trial because they cannot post a bail bond is just $15,109; 37% make less than $9,489 per year. It seems clear that this practice exposes innocents below the poverty line to disproportionately severe burdens, without offsetting benefits adequate to justify the imposition. Whether bail determinations are made in highly personalized ways, or in deference to an algorithmically produced risk score, insofar as they fail to distribute the burdens and benefits of the rule of law fairly, they violate the moral interest that animates the right to be treated as an individual.[27]

[25] This 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 (Ewing 2018), (Howard 2016), (Kim 2020), (Shelby 2007), and (Watson 2015).

[26] (Rabuy and Kopf 2016)'s analysis of data released by the Bureau of Justice Statistics found that "the median bail amount [$10,000] represents eight months of income for the typical jailed defendant."

[27] Reforms that waive bail if a defendant receives a low risk score reduce the number of poor who are subjected to pre-trial detention, but concentrate its effects more heavily on residents of poor communities of color. Rather than using risk scores to filter its effects, fairly distributing the burdens and benefits of law would require that we do away with cash bail as a general practice. My thanks to Tali Mendelberg for bringing this case to my attention, and to Sarah Stroud for pointing out the range of implications the right to individualized treatment has (if I am right) for the administration of criminal law beyond questions about the use of algorithms.

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 which variables are used, and roughly how. Note that this does not forbid 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 The Space for Prediction

That was all very high-level. To unpack the implications for algorithmic tools in particular, we should get a bit more concrete. What the right to individualized treatment requires is not that the application of legal sanctions be *personalized*, but that they express appropriate respect for each individual's autonomy and preserve the mutual accountability of citizens and state. Summarizing, individuals' claims to a fair distribution of the burdens and benefits of the rule of law entail that if a factor *f* is used as a predictor in the administration of criminal law, several conditions must be met:

- *f* must be subject to agents' deliberate control,
- The extra burdens imposed by using *f* as a predictor (increased hassle, risk of false conviction, or severity of punishment) must be outweighed by the communal benefits it yields *to the individuals who must bear these burdens*,
- the factors used (and their relative weighting) as predictors be *public or transparent*,

and relatedly,

- the basis for decisions made be sufficiently clear to facilitate civilian criticism or reform.

There will likely be some cases where predictive algorithms can be deployed while satisfying all of these conditions, but it is unlikely that many of the current applications will make the cut.

4.1 Secrecy and Strategic Gaming

One might be concerned that the strict transparency requirement I have articulated will render predictive tools useless, because making it public what variables are being used as predictors, and how, will render the algorithms vulnerable to strategic gaming. To address this, let's first get clear on the assumptions behind this objection.[28] Strategic gaming is problematic only under very specific conditions:

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

(i) the proxy criteria (the predictors) only weakly or contingently correlate with the target criteria,

(ii) the proxy properties are within subjects' deliberate control (they are alterable),

(iii) the tradeoff costs of gaming the proxy criteria are low, and

(iv) moreover this can be done without affecting the subject's true eligibility with respect to the target criteria.[29]

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 involved in strategic gaming will offset the incentive. 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 that target: 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. So while schools' transparent reliance on LSATs incentivizes students to focus on improving their test scores, this facilitates, rather than undermines, the end goal of admitting well-prepared students. The possibility of strategic gaming fails to provide even a *pro tanto* justification for keeping proxy criteria secret unless gaming would *undercut* the aims.

Similarly, if the proxy for criminal wrongdoing is robustly connected to wrongdoing—e.g. affiliation with a violent organization like the *Proud Boys*, or performance of preparatory acts like buying a high-capacity magazine for a firearm or purchasing large quantities of ammonium nitrate fertilizer—publicity can be net-beneficial. By incentivizing avoidance of the proxy, it discourages the linked criminal behavior. It also equips those for whom the proxy was misleading to avoid or politically contest decisions that rely on it, thus reducing the false-positive error rate. When the costs of a false-positive are comparatively high, these error-correcting tendencies of publicity can outweigh the costs of strategic gaming. But precisely because using predictive proxies attaches costs to behaviors that are not themselves wrongdoing, and so shapes the behavior of those subject to it, (Underwood 1979, 1438) cautions that "[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." Even where predictive, the range of properties used as proxies to guide the application of criminal sanctions will need to be tightly constrained to ensure that it does not intrude too far on autonomy.

The need to keep a proxy secret arises when all four conditions given above are met; and given the way I have characterized the right to individualized treatment, plausibly any algorithm that satisfies it will satisfy conditions (ii) and (iii): that the proxy be within subjects' deliberate

[29] I've drawn these conditions for problematic strategic gaming from (Cofone and Strandberg ms).

control, and low-cost to alter or avoid. So strategic gaming could be a genuine concern if there are compelling reasons to use a highly contingent proxy (satisfying [i]) that is strongly independent of criminal wrongdoing (satisfying [iv]). There may be many administrative decisions for which it is permissible to use secret proxies, but I contend that, with few exceptions, the administration of criminal law is not one of them. Using a property as a proxy for criminality imposes significant costs on those who have it—at *least* high risk of 'hassle factor' (the costs associated with being subjected to extra scrutiny), at worst high risk of unwarranted punishment. If the proxy is only weakly connected with criminal wrongdoing, the state cannot justify its decision to secretly use it by appeal to the harm principle, necessity, or the decision's having been ratified by a democratic decision-making process. And when reliance on the proxy concentrates the highest costs of false-positives disproportionately on an already disadvantaged subpopulation, members of that subgroup have a dual complaint against secrecy: one against the ways that attaching costs to the proxy property undermines their autonomy, and one against the way that the choice of proxy fails to treat their subgroup as political equals. When there are adequate alternative means of enforcing the law, the presumptive weight of either of these complaints defeats the marginal administrative efficiency that could be achieved by secrecy.

4.2 Opacity

One might have a thoroughly different concern about providing the level of transparency the right requires: the ways a sophisticated algorithm arrives at its predictions could simply be *too complex to understand, let alone explain.* In some cases, no matter how much we might want to be transparent about the reasons why these algorithms make the predictions they do, the best we can do is describe how the algorithm was trained. But though the thought that predictive algorithms are essentially a "black box" has captured the popular imagination, it is something of a red herring in this context. Not all predictive algorithms are uninterpretable; only those arising from unconstrained or unsupervised 'deep' learning using high-dimensional models present this particular challenge. And when they are comparably accurate, more transparent algorithms are independently preferable, since opacity can obscure errors and makes it difficult to troubleshoot. It is in fact unlikely that either high-dimensional models or deep learning methods will be necessary—or much help—for optimizing the predictive accuracy of algorithms in the context of criminal law. Though great advances have been made in recognition and automated judgment tasks, machine-intelligence has yet to stably out-perform simple rules at predicting social outcomes (like arrests), consistently plateauing around 65-70% accuracy overall.[30]

COMPAS is no exception: though it leverages a complex model, using up to 137 features of an individual's file to predict risk of being arrested for any new offense within two years of

[30] (Narayanan 2019 MS), (Yang, Wong and Coid 2010)

release, it only achieves about 68% overall accuracy.[31] What this means is that its risk predictions bore out roughly two-thirds of the time: either the person was classified as low-risk and in fact was not rearrested within two years, or they were classified as medium or high risk and were rearrested. An independent audit of COMPAS's predictions by (Angwin, et al. 2016) found that the program had a slightly lower accuracy rate specifically for those it classified as high-risk (61%), but *much* lower accuracy when predicting *violent* reoffending specifically: only 20% of those classified as highly likely to be rearrested for violent crimes actually were.

A predictive accuracy rate of 65-70% is roughly on par with the *far* simpler models used by the second-generation risk assessment tools developed in the 1970s. (Dressel and Farid 2018) found that a standard linear predictor using just 7 features (age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior crimes, crime degree, and charge) yields results comparable to COMPAS's predictor.[32] In fact, they found that untrained subjects who were given just these datapoints about each case and asked to make a prediction (without receiving any particular instruction as to *how*) also outperformed COMPAS in overall accuracy, and displayed slightly *less* racial bias.[33] Perhaps most startlingly—and underscoring just how far our prediction tools are from the imagined pre-crime oracles of *Minority Report*—all of these predictive methods were out-performed by a crude predictor with just two static factors: birthdate and number of prior convictions. While these facts should raise serious moral concerns about relying too heavily on the predictions yielded by these algorithmic tools when making high-stakes decisions, they provide one point of reassurance: uninterpretable models pose no special hurdle to transparency for our purposes, because they aren't all that useful for the applications of interest to us.

4.3 Moral Hazards of Training Predictive Models

There is an additional reason to generally avoid using deep machine learning to develop prediction tools for criminal law. These methods require substantial training data in order to learn predictive patterns, but it is treacherous to use the extant databases (requests for service,

---

[31] Northpointe invoked trade secrets to avoid disclosing the details of their model, but their in-house evaluation of their software put overall accuracy at 68%. See (Dieterich, Mendoza and Brennan 2016).

[32] At p.3. It's worth noting that since offending is measured by *arrest* (or in some cases *conviction)*, rather than directly observed, some proportion of these tools' accuracy is just their ability to predict arrest patterns, which are subject to enforcement bias.

[33] (Dressel and Farid 2018) ran two studies with untrained subjects. In the first condition, participants were given *just* the seven features listed; in the second, they were also told the defendant's race. COMPAS has a recorded overall accuracy of 64.9% for Black defendants, 65.7% for white. It has a 40.4% false-positive error rate for Black defendants, 25.4% for white; and false-negative error rates of 30.9% and 47.9%, respectively. By comparison, Dressel & Farid's subjects had an overall accuracy of 68.2% for Black defendants, 67.6% for white (in condition II this dropped to 66.2% and 67.6%, respectively); false-positive error rates of 37.1% (this rose to 40% in condition II) for Black defendants, and 27.2% (26.2% in condition II) for whites; and false-negative error rates of 29.2% (rose to 30% in condition II) for Black defendants and 40.3% (42.1% in condition II) for whites.

crime reports, arrests, or convictions) for this purpose. Information recorded in these datasets is invisibly shaped both by administrative discretion and by upstream structural injustices that artificially forced overlap between communities of color and criminogenic conditions--most especially underfunded schools and depressed economic conditions.

Some hope that we can correct for this with more datasets: given rich enough data, factors that are unrelated to the outcome of interest won't correlate closely enough with it to be reliable predictors, and so will not be learned. [34] But this optimism is misplaced when the information in available datasets is relatively sparse, or the overlap between properties is not happenstance but *artificial*, or the outcome can only be measured or represented indirectly through measures (like 'arrests') that are themselves shaped by unrelated factors (like the probability of detection, political influence, trust in the police, or familiarity with legal protections). As (Johnson 2020) demonstrates, even explicitly coding the model *not* to use properties like race or gender as predictors will not prevent it from learning to make predictions that track these features: "Where there are robust correlations between socially sensitive attributes, proxy attributes, and target features, and we've ruled out using the socially sensitive attributes, the next best thing for the program to use will be the proxy attributes." Put simply, algorithms trained on datasets in which decisions to arrest, conviction, sentence, and re-arrest have been subject to racial bias can be expected to learn correlations that, though causally spurious, are genuinely "there" in the data, projecting these traces of past injustice forward.[35]

For street crime in particular, many of the strongly correlated properties are straightforward measures of socioeconomic disadvantage: employment status, income, education level, prior contacts with police, and relative security of housing. So it is doubtful that an algorithm trained on the available datasets would be able to respect our constraint that predictors be limited to factors within subjects' deliberate control. But even bracketing these concerns about available training data, and *even if* the predictions made were highly accurate, the right to individualization as I have interpreted it may more directly preclude using deep machine-learning in developing the algorithms. A learning method which bases the risk prediction on correlations that *emerge* between very large numbers of variables and the outcomes is necessarily 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

[34] My thanks to Simon Goldstein and Stephen Finlay for this suggestion; it is developed in more detail in (J. Kleinberg, J. Ludwig, et al. 2018), at p. 136.

[35] 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 2019).

be *prospective.* So while it is true that the right as I have glossed it precludes the use of unsupervised deep machine learning, this is not—or not solely—because it is unexplainable.

5 Some upshots

We began with a simple question—is the use of algorithmic prediction tools in criminal law consistent with the right to be treated as individual?—and have arrived at a highly qualified 'yes'. On the interpretation I have offered, this right does not preclude the use of statistical methods in principle, but does significantly constrain their design. 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' before they manifest. It rather must function to announce expectations for behavior, using the coercive apparatus only to hold agents accountable to those very expectations. When legal decisions are made in ways that do not afford subjects a fair opportunity to avoid hostile encounters with law enforcement, or that impose costs disproportionately, this constitutes an unfair distribution of the burdens and benefits of the rule of law. The impulse toward secrecy must be resisted; where predictions are made, they must be based only on factors that are within agents' deliberate control, and not core to valuable exercise of autonomy.

Requiring a fair distribution strongly constrains which variables can be used as a basis for applying extra scrutiny or criminal sanction. It rules out reliance on a great many static factors (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 because they are unavoidable; the latter because using them to justify the imposition of yet more costs on the victims of upstream distributive injustice—this time in the form of increased risk of suffering unjustified state coercion—is patently unfair.

Where this leaves us depends on the application. For street crime—particularly property offenses like autotheft, burglary, or mugging—the social value of accurately predicting any particular future offense is low, especially as compared to the cost of a false positive prediction to each individual who is misclassified. And insofar as these sorts of crimes are driven by inelastic social causes, a proxy is unlikely to have strong deterrent effects, and *is* likely to track socioeconomic disadvantage. Once we also take into account the low accuracy rates, it is unlikely that the evidential value of predictive algorithms will suffice to justify their use in these contexts.

It might be possible *in principle* to design risk-assessment or crime-prediction algorithms to be independent of these variables—though I am persuaded by (Johnson 2020)'s argument that as long as robust unjust correlations (e.g. between race and rates of arrest) persist in the world, an algorithm tasked with making pattern-matching predictions will find a proxy for the forbidden variables. But it is at best unclear what evidential or predictive value a truly unbiased risk assessment tool would have. Of the extant tools, those that conditionalize on static variables *alone* presently outperform those that *also* incorporate dynamic variables.[36] We can expect that both would outperform prediction based only on the subset of dynamic variables that are not ruled out by the considerations just raised. So, while it may be possible to constrain the data used so that an algorithmic risk projection is consistent with the moral interests protected by a 'right to individualized evidence', it is unclear whether such permissible predictors will have significant evidential value.

However, white-collar crime, wage theft, and financial fraud more generally may well be appropriate arenas for the use of predictive tools. These tend to have a higher victims-per-crime ratio, and so a higher social value to predicting or identifying any single instance. They are also most commonly perpetrated by a relatively well-resourced portion of the population, for whom additional scrutiny presents little more than a hassle. The subpopulations subject to extra scrutiny, higher risk of false conviction, or longer sentences due to reliance on algorithms in financial crimes are also less likely to overlap with either a stable subgroup (like racial or SES category) or with populations already subject to intersectional disadvantage/distributive injustice, so the presumptive reasons against using a secret proxy are far less weighty.

In closing, I want to circle back to the optimistic aim of using algorithmic tools to improve the high-stakes decisions of criminal law. It is laudable to try to make determinations less biased, and to reduce the number of people subjected to unjustified or disproportionate costs in the course of law enforcement. Maybe we could make some progress toward this aim by supplementing the judgment of police officers, judges, juries, and parole boards with algorithmic assessments. But this says more about how badly distorted our unassisted decisions are than about the accuracy or fairness of the algorithmic tools. Whether it is wise to embrace these tools as an incremental improvement depends on several factors we haven't had space to work through in this paper, including what the alternative is, and how decisionmakers would be instructed to incorporate the risk predictions into their deliberations.

The most natural instruction—that a high risk score may be sufficient, but not necessary, for an adverse judgment—may actually be the worst of both worlds. If high risk is not necessary for adverse judgment, then the algorithmic tool does not constrain extant bias the decisionmakers may have toward (e.g.) giving disproportionately long sentences to defendants of color. But if it *is* sufficient, then any bias in the false-positive error rates of the algorithm is simply combined with the extant distortions---and worse, the whole decision process has a veneer of being even-handed and objective.

We might hope that risk scores could be used to alleviate bias in a different way. It is clearly unjust to base the distribution of *burdens* on unavoidable factors, but perhaps the same

[36] (Herrschaft 2014); (Dressel and Farid 2018). But see (Degiorgio and DiDonato 2013) for findings that adding dynamic factors to static demographic models in fact improves the fit of a model predicting probation revocation specifically for substance abuse.

cannot be said of distributing *benefits*. Could we simply use low risk scores to exonerate, shorten sentences, or waive bail? There is cause for concern here too. A policy of this kind would channel goods towards those who *lack* the markers of disadvantage that yield a high risk score, and so can still be expected to entrench racial and economic inequalities and compound disadvantage. There is also a dark side to partial reforms when they succeed in alleviating injustice for many but concentrate the remaining costs on people who are comparatively vulnerable or politically powerless. Once only marginalized groups face the worst costs, is much more difficult to build political will to enact the reforms necessary to correct the injustice. A partial fix may well be worse than doing nothing, then, because it allows the majority to simply look away.

6 Acknowledgements

I am especially grateful to Jeff Behrends, Stephen Finlay, Simon Goldstein, Ian Hamilton, Gabriel Karger, Seth Lazar, Chad Lee-Stronach, Tali Mendelberg, Tom Parr, Chelsea Rosenthal, Kate Vredenburgh, and Annette Zimmerman, for written comments on earlier drafts of this paper. I am also indebted to Rima Basu, Luc Bovens, Katie Creel, Marcello Di'Bello, Tom Dougherty, Anna Edmonds, Maegan Fairchild, Sarah Stroud, Robin Guong, Bernard Harcourt, Deborah Hellman, Sarah Hirschfield, Gabrielle Johnson, Geoff Sayre-McCord, Victor Tadros, Patrick Tomlin, Alex Worsnip, and Brian Zaharatos, for helpful comments and discussion of the ideas presented here. Thanks also to other participants at the 2019 workshop on the *Democratic Implications of Algorithmic Decision-Making* at Princeton University, the 2020 *Radcliffe Institute Workshop on The Ethics of Technology at Work and in Public Institutions* at Harvard University, the Dianoia Institute of Philosophy *Ethics Working Group* at Australian Catholic University, the 2020 *Philosophy, Artificial Intelligence, and Society* workshop, the 2021 works-in-progress group at University of North Carolina Chapel Hill, as well as to seminar participants at the *Center for Ethics, Law, and Public Affairs* at the University of Warwick, Rima Basu's 2020 seminar *on Belief, Evidence, and Agency* at Claremont McKenna, and the 2020 Woodrow Wilson Fellows at Princeton University, for helpful discussion of the ideas presented here.

7 Works Cited

Amour, Jody. 1994. "Race Ipsa Loquitur: Of Reasonable Racists, Intelligent Bayesians, and Involuntary Negrophobes." *Stanford Law Review* 46 (4): 781-816.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against Blacks." *ProPublica* www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bambauer, Jane. 2015. "Hassle." *Michigan Law Review* 113 (4): 461-511.

Basu, Rima. 2019. "The Wrongs of Racist Belief." *Philosophical Studies* (Routledge) 176: 2497-2515.

Basu, Rima. 2019. "What We Epistemically Owe to Each Other." *Philosophical Studies* 176: 915-931.

Binns, Reuben. forthcoming. "Human Judgement in Algorithmic Loops." *Regulation & Governance* ((on file with author)).

Bolinger, Renee Jorgensen. 2021. "Explaining Justificatory Asymmetries between Statistical and Individualized Evidence." In *The Social Epistemology of Legal Trials*, by Zachary Hoskins and Jon Robson. Routledge.

Bolinger, Renee Jorgensen. 2020. "The Rational Impermissibibility of Accepting (some) Racial Generalizations." *Synthese* 197: 2415-2431.

Brayne, Sarah. 2021. *Predict and Surveil: Data, Discretion, and the Future of Policing.* New York: Oxford University Press.

Buchak, Lara. 2014. "Belief, Credence, and Norms." *Philosophical Studies* 169: 285–311.

Burgess, Earnest. 1936-7. "Protecting the Public by Parole and by Parole Prediction." *American Institute of Criminal Law & Criminology* 27: 491.

Cofone, and Strandberg. ms. "Strategic Games and Algorithmic Transparency." *Conference Draft, on file with author.*

Colyvan, M, H Regan, and S Person. 2001. "Is it a Crime to Belong to a Reference Class?" *The Journal fo Political Philosophy.*

Degiorgio, and DiDonato. 2013. "Predicting probationer rates of recincarceration using dynamic factors from the Substance Abuse Questionnaire-Adult Probation III (SAQ-Adult Probation III)." *American Journal of Criminal Justice* 39: 94-108.

Dieterich, W, T Mendoza, and T Brennan. 2016. *COMPAS risk scales: Demonstrating Accuracy, Equity, and Predictive Parity.* Technical Report, Northpointe Inc.,.

Dressel, and Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4 (eaao5580): 3.

Duff, R.A. 1998. "Dangerousness and Citizenship." In *Fundamentals of Sentencing Theory*, by A. J. Ashworth and M. Wasik, 155-6. Oxford: Oxford University Press.

Dworkin, Ronald. 1977. *Taking Rights Seriously.*

Enoch, David, L Spectre, and T Fisher. 2012. "Statistical evidence, Sensitivity, and the Legal Value of Knowledge." *Philosophy and Public Affairs* 40 (3): 197-224.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: St. Martin's Press.

Ewing, Benjamin. 2018. "Recent Work on Punishment and Criminogenic Disadvantage." *Law and Philosophy* 37: 29-68.

Fuller, Lon. 1969. *The Morality of Law.* New Haven: Yale University Press.

Harcourt, Bernard. 2007. *Against Prediction: Profiling, Policing, and Punishment in an Actuarial Age.* Chicago, IL: University of Chicago Press.

Harris, D. 1998. "Particularized Suspicion, Categorical Judgments: Supreme Court Rhetoric Versus Lower Court Reality Under Terry v. Ohio." *St. John's Law Review* 72 (3): 975.

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.

Holder, Attorney General Eric. 2014. "Remarks to the National Association of Criminal Defense Lawyers 57th Annual Meeting, and 13th State Criminal Justice Network Conference." Philadelphia, PA, August 1.

Howard, Jeffrey. 2016. "Moral Subversion and Structural Entrapment." *Journal of Political Philosophy* 24 (1): 24-46.

Johnson, Gabrielle. 2020. "Algorithmic bias: on the implicit biases of social technology." *Synthese* s11229-020-02696-y.

Joyce, James. 2005. "How Evidence Reflects Probabilities." *Philosophical Perspectives* 19 (1): 153-178.

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 & Soceity, Harvard Law School.

Kim, Hochan. 2020. "Entrapment, Culpability, and Legitimacy." *Law and Philosophy* 39: 67-91.

Kleinberg, J., J. Ludwig, S. Mullainathany, and C. Sunstein. 2018. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10: 113-174.

Kleinberg, J., S. Ludwig, C. Mullainathan, and C. Sunstein. 2020. "Algorithms as Discrimination Detectors." *Proceedings of the National Academy of Sciences.*

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

Marušić, Berislav, and Stephen White. 2018. "How Can Beliefs Wrong? -- A Strawsonian Epistemology." *Philosophical Topics* (46): 97-114.

Moss, Sarah. 2018. "Moral Encroachment." *Proceedings of the Aristotelian Society* 118 (2): 177-205.

—. 2018. *Probabilistic Knowledge.* Oxford: Oxford University Press.

Murphy, Colleen. 2005. "Lon Fuller and the Moral Value of the Rule of Law." *Law and Philosophy* 24 (3): 239-262.

Narayanan, A. 2019 MS. "How to Recognize AI Snake Oil." *Working paper, on file with author.*

O'Neil, Cathy. 2016. *Weapons of Math Destruction.* New York: Crown Publishers.

Rabuy, Bernadette, and Daniel Kopf. 2016. *Detaining the Poor: How money bail perpetuates an endless cycle of poverty and jail time.* Prison Policy Initiative.

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.

Risinger, D. M. 2004. "Unsafe verdicts: The need for reformed standards for the trial and review of factual innocence claims." *Houston Law Review* 21: 1281.

Schroeder, Mark. 2018. "When Beliefs Wrong." *Philosophical Topics* 46: 115-127.

Shelby, Tommie. 2007. "Justice, Deviance, and the Dark Ghetto." *Philosophy & Public Affairs* 35 (2): 126-160.

Taslitz, Andrew. 2010. "Police are People Too: Cognitive Obstacles to, and Opportunities for, Police Getting the Individualized Suspicion Judgment Right." *Ohio State Journal of Criminal Law* 8 (1): 7-78.

Thomson, Judith Jarvis. 1986. "Liability and Individualized Evidence." *Law and Contemporary Problems* 49 (3): 199-219.

Tillers, Peter. 2005. "If wishes were horses: Discursive comments on attempts to prevent individuals from being unfairly burdened by their reference classes." *Law, Probability, and Risk* 4: 33-49.

Tribe, L. 1971. "Trial by Mathematics: Precision and Ritual in the Legal Process." *Harvard Law Review* 84 (6): 1329-1393.

Underwood, Barbara. 1979. "Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment." *Yale Law Journal* 88: 1408.

Voigt, Camp, Prabhakaran, Hamilton, Hetey, Griffiths, Jurgens, Jurafsky, and Eberhardt. 2017. "Racial Disparities in Police Langauge." *Proceedings of the National Academy of Sciences* 114 (25): 6521-6526.

Vredenberg, Kate. working paper. "The Right to an Explanation." *on file with author.*

Walen, Alec. 2011. "A Punitive Precondition for Preventative Detention: Lost Status as a Foundation for a Lost Immunity." *San Diego Law Review* 48: 1229.

Walen, Alec. 2011. "A Unified Theory of Detention, with Application to Preventative Detention for Suspected Terrorists." *Maryland Law Review* 70: 871.

Wasserman, D. 1992. "The Morality of Statistical Proof and the Risk of Mistaken Liability." *Cardozo Law Review* 13: 935.

Watson, Gary. 2015. "A Moral Predicament in the Criminal Law." *Inquiry* 58 (2): 168-188.

Yang, M, S.C. Wong, and J Coid. 2010. "The Efficacy of Violence prediction: a meta-analytic comparison of nine risk assessment tools." *Psychological Bulletin* 136: 740-767.