Beyond "Fairness": Rethinking the Use of Algorithmic Predictions in Criminal Justice

Tashmia Sabera
University of Wisconsin-Madison
sabera@wisc.edu

Abstract

This paper critiques the widespread use of predictive algorithmic tools in criminal justice, such as COMPAS, arguing that concerns about fairness and accuracy, while important, fail to address a deeper ethical issue: the infringement of the right to be treated as an individual. Drawing on Renee Jorgensen's work, I argue that fairness-based reforms are insufficient because predictive punishment is incompatible with the demands of negative retributivism, the theory of punishment most compatible with the right to be treated as an individual. Given the high stakes of criminal law and the inherent trade-offs involved in algorithmic fairness metrics, I contend that algorithmic predictions should not be used to justify punishment or policing decisions. However, I propose that algorithmic tools can be ethically employed in developing policies aimed at crime reduction, provided they are used to identify causal factors rather than to predict individual behavior. To this end, I advocate a pluralist framework: negative retributivism should govern punishment and policing, while rights-based consequentialism should inform long-term policy goals. This approach aims to clarify when the use of algorithms in criminal justice is unjustified and when it may be justified with critical revision.

Introduction

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is an algorithmic tool used in many U.S. courts to predict recidivism risk, guiding decisions on bail and additional sentencing for convicted offenders. In 2016, a report revealed that COMPAS makes inaccurate predictions and produces racially biased results (Larson 2016). This finding drew academic attention to the justification of imposing predictive punishment by algorithmic tools. The problem of inaccurate predictions and racial bias in predictive algorithmic tools like COMPAS has largely been framed in terms of a narrow conception of fairness — typically focused on increasing accuracy or eliminating bias (Corbett-Davies 2023, Berk 2021). However, one could argue that a deeper ethical issue is at stake — one that cannot be resolved merely by ensuring fairness: predictive punishment itself violates the right to be treated as an individual, and therefore, lacks moral justification.

Renee Jorgensen rightly identifies this potential breach of the right but seeks to address it solely by promoting greater fairness in the use of algorithmic tools in criminal law (Jorgensen 2022). In this paper, I argue that this narrow conception of fairness fails to uphold the basic requirements of the right to be treated as an individual which requires meeting the standard of negative retributivism in the context of criminal justice. In my view, algorithmic predictions are incompatible with negative retributivism, and therefore, should not be used for punishment or policing. Additionally, the high stakes involved in criminal law give us a strong reason to avoid relying on algorithmic predictions for decision-making, as the fair determination of risk scores involves inherent trade-offs, making it impossible to satisfy all fairness metrics simultaneously (Kleinberg 2016).

Nevertheless, with critical revision, algorithmic tools can be used to enact social policies aimed at reducing crime. This proposal is compatible with views suggesting that machine learning is ill-equipped for making *predictions*, yet it can be used to identify variables to inform a causal model aimed at understanding the social, structural and psychological factors that drive crime (Barabas 2018). In this way, we can leverage algorithmic tools to reduce crime without violating the right to individualized treatment. I propose adopting a pluralist and comprehensive theory of criminal justice, in which negative retributivism informs our punishment and policing laws, while rights-based consequentialism guides long-term policies for crime reduction. Overall, my aim is to provide the philosophical underpinning of a framework that clarifies where the use of algorithms is unjustified and where it can be justified with critical revision.

Right to be Treated as an Individual in Criminal Law

In my view, treating people as an individual in the context of criminal law means not punishing them for anything beyond their own actions. To develop this argument, I adopt Eidelson's account of what it means to treat people as individuals, as it aligns with my intuition about the nature of this right. Eidelson argues that treating people as individuals is not merely a matter of fairness or conscientiousness; rather, it is involves treating them as *ends in themselves* (Eidelson 2013). The right protects a unique interest of individuals — namely, respect for autonomy. Drawing on a widely shared consensus among philosophers including Joseph Raz and Joel Feinberg, Eidelson emphasizes that to be autonomous is to be the author of one's own choices and actions.

According to this view, a person is treated as an individual if two conditions are met: the character and the agency condition. The *character* condition requires us to interpret an individual's past decisions originating from themselves, while the *agency* condition demands that we acknowledge an individual's future capacity to reflect on their decisions. This includes recognizing their ability to engage in second-order reflections on their first-order desires. For instance, a person may have a first-order desire for violence, however, we should also recognize their second-order capacity to reflect on and manage this disposition.

One might object that the requirement seems overly demanding, especially when it comes to applying statistical generalizations on individuals, rendering the account implausible. However, Eidelson argues that the conditions do not prohibit relying on statistical evidence, provided that the statistics are based on the relevant information concerning a person's authorship and allow room for their future autonomous decisions. Moreover, the right does not require going out of one's way to gather individualized evidence; rather, the collection and evaluation of evidence should meet a standard of *reasonableness*.

Among the theories of punishment, negative retributivism most effectively satisfies the conditions of the right to be treated as an individual. This is because negative retributivism holds that punishment ought to track deservingness in a fitting way (Quinton 1954). Under negative retributivism, guilt is a necessary but not sufficient condition for punishment, meaning that punishing the innocent or imposing excessive punishment on the guilty is unjustified. When punishment is imposed, two essential sentencing constraints must be upheld: proportionality and fairness (Zimmerman 2019).

Proportionality can be of two types: ordinal and cardinal. Ordinal proportionality focuses on ranking offenses, ensuring that more serious crimes receive harsher punishments than less serious ones (Walen). It does not specify the exact severity of punishment, but rather that punishments are ordered justly. For example, if theft is less serious than murder, then theft should have a lighter punishment than murder. Cardinal proportionality, by contrast, concerns the actual severity of punishment and whether it appropriately reflects the gravity of the crime. It is about calibrating punishment, not just ranking it. For instance, if a fine of $10,000 for theft is too excessive, and one year in prison for murder is too lenient, then both fail the test of cardinal proportionality. Overall, proportionality ensures that punishment is based on an individual's actions, rather than on statistical generalization tied to their demographic identity.

Fairness should be regarded as a separate and distinct constraint within negative retributivism. This standard requires that the nature of punishment be fair in itself and that it does not exacerbate the adverse effect of structurally unfair background conditions, such as racial inequality (Zimmerman 2019). Treating people as individuals requires adherence to both proportionality and fairness. Later in this paper, I show that most attempts to address the use of algorithmic tools in criminal law focus solely on this conception of fairness, overlooking the constraint of proportionality. This is a significant shortcoming, given the high stakes involved in criminal justice.

Due to space constraints, I limit my discussion to retributive and deterrent theories of punishment and therefore do not engage with other theories such as expressivism or restorative justice. Deterrence theory holds that punishment is justified as long as it deters potential offenders through its severity and certainty. This theory allows for punishing an innocent person and imposing harsher punishment on the convicted to deter others (Boonin 2008). It is incompatible with the right to be treated as an individual because it disregards both the character and the agency conditions. It does not meet the character condition because it fails to consider subject's past decisions or actions as the sole basis for punishment. Moreover, deterrence theory does not recognize that individuals, even those prone to crime, have the second-order capacity to reflect on and regulate their first-order desires. Therefore, deterrence theory is incompatible with the right to be treated as an individual.

Jorgensen on the Right to be Treated as an Individual

Jorgensen addresses whether the use of algorithmic predictions in criminal justice infringes upon the right to be treated as an individual. In her view, the right does not protect any unique interest; rather, it requires that the procedures of criminal law be justifiable to individuals subject to them. She argues that the right entails a claim to fair distribution of the benefits and burdens of public law and therefore, the

use of algorithmic predictions in criminal justice does not, in itself, infringe it.

It is unclear why the right should be understood as not protecting a unique interest, given that its very purpose is to affirm the value of individuality. I believe the most compelling way to uphold this right is to interpret it through the lens of individual autonomy — something that cannot be secured merely by ensuring procedural fairness or conscientiousness. By failing to recognize this crucial dimension, Jorgensen overlooks what lies at the heart of the right — self-authorship of individuals — and reduces it to a narrow conception of fairness.

Even if one accepts that the right does not entail a commitment to individual autonomy, autonomy remains a distinct moral value, and Jorgensen's conception of fairness fails to satisfy its demands. She outlines three conditions under which the use of algorithmic predictions in criminal law can be justified to individuals; control, transparency and burden. The control condition is satisfied if the predictive factor is subject to the agent's deliberate control; transparency requires the predictive factors to be sufficiently clear to facilitate civilian criticism and reform; the burden condition holds that any unavoidable extra burdens must be outweighed by the corresponding benefits to the individual. By burden, she refers to increased hassle, the risk of false conviction or the severity of punishment.

The control condition holds that when criminal law employs predictive inferences, the relevant factors must lie beyond the subject's deliberate control. It is problematic in two ways. First, it presupposes that imposing additional punishment on the basis of prediction is justified, thereby overlooking the *agency* condition of autonomy. Second, factors that appear to be within the subject's deliberate control may, in fact, be closely tied to factors beyond their control. For instance, the control condition might require that COMPAS exclude questions related to zip code, since neighborhood is not within the subject's deliberate control and often tied to subject's racial background. Yet COMPAS also asks about a subject's frequency of feeling discouraged, or their associations and values. At first glance, these may not seem beyond the subject's control. However, our experiences are deeply shaped by our environment. Someone who grew up in a crime-prone area and attended a poorly funded school is more likely to have friends with prior arrests. Similarly, a person from a marginalized community is more likely to feel frequently discouraged due to systemic bias. It remains unclear how far we must go to sever the ties between factors beyond our control and those seemingly within it- but the likelihood of entanglement is very high (Benjamin 2019, Eubanks 2018). Jorgensen acknowledges this issue, referring to it as the "entrapment" problem, and proposes the burden condition as a solution: balancing the burden imposed by predictive factors against the benefits they provide.

However, the burden condition faces serious objections too. First, it fails to satisfy the requirements of ordinal proportionality, as Jorgensen proposes using actuarial predictions for white-collar crime, wage theft and financial fraud, but not for street crimes like auto-theft or burglary, or even for violent crimes like homicide. This is problematic because, as argued earlier, the right to be treated as an individual is best secured by negative retributivism, which treats ordinal proportionality as a key constraint. It violates ordinal proportionality by failing to impose harsher punishment on more serious crimes. Even setting that view aside, it seems counterintuitive for white collar offenders to face greater scrutiny and harsher punishment than those who commit street or violent crimes, which are generally considered more serious. Jorgensen defends her view by claiming that actuarial predictions for homicides or street crimes are less accurate and more likely to track socio-economic disadvantages. These empirical claims seem somewhat ad hoc. Moreover, socio-economic disadvantage is not the only factor beyond an individual's control; white-collar offenders may also be subject to such uncontrollable influences used in the actuarial predictions.

Furthermore, Jorgensen regards criminal law as analogous to other areas of public law, such as land or constitutional law. However, criminal law is a distinct domain involving high-stakes decisions about individual rights (Hart 2008). Standards acceptable in other public law contexts may not suffice here, given the uniquely severe consequences. For example, the widely accepted evidentiary standard in the civil law of common law countries is the balance of probabilities, under which a party wins the case if their version of events is slightly more convincing than their opponent's. In contrast, the evidentiary standard in criminal law of common law legal systems is proof beyond a reasonable doubt. Under this standard, if there is reasonable doubt regarding any part of the prosecution's case, the accused must be acquitted.

Finally, there are good reasons to be skeptical about the effectiveness of deterrent and predictive punishment. Bernard Harcourt argues that actuarial punishment not only undermines principles of just punishment but also fails to prevent crime (Harcourt 2019). Similarly, Ian Hacking highlights how classification can shape self-perception and behavior, potentially reinforcing the very patterns the system aims to predict (Hacking 1995). In this way, the recidivism risk scores generated by algorithmic tools may become self-fulfilling prophecies regarding the likelihood of re-offending. Thus, using actuarial predictions to inflict additional punishment lacks moral justification, both in terms of its violation of autonomy and its questionable effectiveness.

Rethinking Algorithmic Prediction in Criminal Justice

Criminal justice has many possible goals: the reduction and prevention of crime, the reformation and rehabilitation of offenders, and the pursuit of justice. The right to be treated as an individual serves as a safeguard for individuals in the pursuit of achieving these goals. A reasonable question arises: if we have access to statistical predictions of potential offenses, what should the state's response be in balancing the right to individualized treatment with the goals of criminal justice, such as reduction and prevention of crime?

Before answering that, let me outline the general philosophical framework I propose for the criminal justice system. I support adopting a comprehensive theory of criminal justice, as crime should be addressed as a single, but multifaceted, problem. My idea is partly motivated by John Braithwaite and Philip Pettit's comprehensive theory of criminal justice (Braithwaite and Pettit 1992). They emphasize the need for a comprehensive theory of criminal justice, arguing that the standards applied to punishment should govern the exercise of police discretion. Appealing to the principle of simplicity, they advocate for a consequential approach, contending that retributivism alone cannot address all aspects of criminal justice. At the same time, they remain committed to protecting individual rights. Overall, they recommend a rights-based consequential theory.

Building on this pluralist approach, I propose a comprehensive theory that includes not only punishment and policing but also reformative and restorative perspectives. Long-term policies, such as poverty reduction and education, should guide crime prevention, grounded in consequentialist reasoning, so long as they respect individual rights. Overall, I propose a pluralist, comprehensive theory of criminal justice.

My proposal is as follows:

Rule 1: Negative retributivism should guide punishment and policing.

Rule 2: Rights-based consequentialism should guide policy decisions aimed at crime prevention, with the overarching goal of protecting individual rights.

Rule 1 entails that actuarial inferences should not justify the punishment of offenders. Moreover, for the same reason, the application of such inferences should be avoided in policing practices so as not to reinforce racially biased surveillance (Eubanks 2018, Browne 2015). Rule 1 takes precedence over Rule 2. According to Rule 2, consequentialist policies may only be pursued to the extent that they do not violate the principles of negative retributivism. This pluralist theory is compatible with negative retributivism, as negative retributivism allows for plausible forward-looking instrumental concerns (Zimmerman 2019).

While predictive punishment is unjustified and incompatible with the right to be treated as an individual, this does not mean we must reject algorithmic tools altogether. Rather, they should be used in ways that respect and preserve individual rights and autonomy. States can employ algorithms to design long-term crime reduction policies, but this must be done critically and cautiously, as studies highlight the risk of algorithmic injustice and human rights violations (Marjanovic 2021). In particular, algorithmic predictions may inform long-term policies under Rule 2, provided marginalized and vulnerable groups are not adversely affected. To that end, we must ensure that such algorithmic tools are not biased against marginalized people (Noble 2018). Sandra Hardin's idea of standpoint epistemology could be helpful for developing participatory design processes that recognize and address systemic biases, ensuring that AI systems do not perpetuate existing inequalities (Hardin 1991).

That said, in this paper, I do not intend to offer a detailed account of how algorithms should be implemented in policy decisions. My general suggestion is that any such use must be compatible with rights-based consequentialism.

Objections

A. If Preventive Detention Is Justified, Why Not Predictive Punishment?

A potential objection to my view arises from the widespread acceptance of preventive detention in many legal systems. While punishment typically carries greater legal consequences and social stigma, one could argue that the suffering caused by preventive detention is nearly equivalent to predictive punishment. By "greater legal consequences," I mean that punishment follows a conviction, which, in most cases, becomes part of the permanent record of the convicted person and affects their future access to many areas of life — such as employment, education, and more. However, by highlighting the suffering involved in both, one could reasonably ask: why, then, should we justify the former but not the latter?

First, negative retributivists do not endorse all forms of preventive detention, as many are overly broad, arbitrary, and fail to track desert. However, detention may be justified in rare cases of serious offences that pose grave harm. Since punishment is justified based on desert, it is not contradictory, as posing grave harm can be seen as a form of desert. Importantly, the suspicion must be based on the subject's actions, not their group membership. Moreover, many cases in which preventive detention seems justified may already fall under a separate offence, such as attempted crime. For example, in many legal systems, preparing to commit armed robbery is recognized as a separate offence from robbery itself, as the preparatory actions themselves constitute culpability. If that is so, only a few situations would require preventive detention.

Overall, negative retributivism justifies preventive detention based on highly suspicious actions, not predictions or group membership. It is justified because negative retributivism operates on the basis of deservingness. Thus, negative retributivism can maintain opposition to predictive punishment without self-contradiction.

B. Preventing Grave Consequences

An objection may arise from the desire to prevent serious harm to many people. For example, if a convicted person is assessed as having a high risk of recidivism in committing a grave offence, such as inciting genocide or committing war crimes, why wouldn't preventive incapacitation be justified?

The issue can be explored through the thought experiment of killing young Hitler: if we could travel back in time before Hitler initiated the Nazi regime, would it be justified to kill him to prevent the future harm? My response is grounded in autonomy, the intuitive moral core of the right to be treated as an individual. Its character condition implies that no one should be punished for acts they have not committed, while its agency condition assumes that individuals can reflect on their harmful dispositions and regulate them. Together, these principles make it unjustifiable to incapacitate or punish someone solely based on the possibility of future harm, even when the gravity of that potential harm is significant.

Conclusion

One problem with Jorgensen's approach — and the traditional fairness-focused responses to algorithmic predictions in criminal law — is their narrow conception of fairness. The right to be treated as an individual is not secured by fairness alone. It must also protect individual autonomy, which requires grounding in a punishment theory that fully upholds all aspects of this right.

This paper has shown that negative retributivism aligns with the standard of treating individuals as autonomous agents. It prohibits the use of actuarial predictions — such as those from tools like COMPAS — for imposing additional punishment. Therefore, such inferences should not be used in criminal law. However, their use may be permissible in long-term crime reduction policymaking. There is a growing body of literature that explores how algorithms might be used in policymaking (Perry and Uuk 2019, Green 2021). While I have not explored the technical details of such algorithmic implementation, I suggest that such decisions should be guided by rights-based consequentialism.

Overall, I have proposed a framework that clarifies when the use of algorithms is justified and when it is not. A key takeaway of this framework is that, to achieve holistic justice in criminal law requires moving beyond a narrow conception of fairness.

Acknowledgements

The author is grateful to Annette Zimmerman for her detailed and insightful feedback, and also thanks the anonymous reviewers for their helpful comments.

References

Barabas, C.; Virza, M.; Dinakar, K.; Ito, J. and Zittrain, J. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In Conference on *Fairness, Accountability and Transparency*, pp. 62-76. PMLR, 2018.

Benjamin, R. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code.* Cambridge: Polity Press.

Berk, R., Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50(1): 3-44.

Boonin, D. 2008. *The Problem of Punishment*. Cambridge: Cambridge University Press.

Braithwaite, J.; and Pettit, P. 1992. *Not just deserts: A Republican Theory of Criminal Justice*. Oxford University Press.

Browne, S. 2015. *Dark Matters: On the Surveillance of Blackness*. Duke University Press.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R. and Goel, S. 2023. The Measure and Mismeasure of Fairness. *Journal of Machine Learning Research* 24 (312): 1-117.

Eidelson, B. 2013. Treating People as Individuals in *Philosophical Foundations of Discrimination Law*, ed. Deborah Hellman and Sophia Moreau, 203–227. Oxford: Oxford University Press.

Eubanks, V. 2018. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

Green, B. 2021. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing* 2(3): 249-265.

Hacking, I. 1995. The Looping Effects of Human Kinds. In *Causal Cognition: A Multidisciplinary Debate,* edited by Sperber, D. Premack, D. & Premack, A.J., 351–394. Clarendon Press/Oxford University Press.

Harcourt, B. E. 2019. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* University of Chicago Press.

Hart, H.L.A. 2008. *Punishment and responsibility: Essays in the Philosophy of Law*. Oxford University Press.

Jorgensen, R. Algorithms and the Individual in Criminal Law. 2022. *Canadian Journal of Philosophy* 52(1): 61-77.

Jeff, L.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. We Analyzed the COMPAS Recidivism Algorithm <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

Noble, S. U. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. In *Algorithms of oppression*. New York University Press.

Perry, B. and Uuk, R. 2019. AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk. *Big Data and Cognitive Computing* 3(2): 26.

Quinton, A. M. 1954. On punishment. *Analysis* 14(6): 133-142.

Harding, S. 1991. *Whose Science? Whose Knowledge?* 1991. Ithaca: Cornell University Press.

Walen, A. "Challenges to the Notion of Retributive Proportionality" <https://plato.stanford.edu/entries/justice-retributive/challenges.html> Accessed: 2025-04-19.

Zimmermann, A. 2019. Criminal Disenfranchisement and the Concept of Political Wrongdoing. *Philosophy & Public Affairs* 47(4): 378-411.