Trust and Legitimacy in an Era of Algorithmic Criminal Justice

Hadar Dancig-Rosenberg

This chapter explores the implications of algorithmic decision-making in the criminal justice system, focusing on the concepts of trust, legitimacy, and accountability. It discusses whether the transition to AI-driven criminal justice signifies a genuine regime change or merely perpetuates existing biases under the guise of neutrality. It highlights how algorithms, while promising consistency and efficiency, may undermine procedural justice principles. It proposes that integrating human discretion with algorithmic tools, alongside participatory and deliberative frameworks, could enhance the legitimacy and trustworthiness of AI-driven criminal justice systems.

A. Introduction

Algorithmic criminal justice has transitioned from a theoretical concept to an undeniable reality. The adoption of algorithmic and AI-based tools and technologies by criminal legal institutions is no longer a matter of "if." Such algorithms are now integral to various stages of the criminal legal process, from predictive policing to pretrial detention, predictive prosecution, sentencing, and post-sentencing.[1]

As time progresses and more experience is gained in using AI tools, the normative debate over the desirability of making the criminal legal system rely on algorithms with a 'mediation' of humans becomes more controversial. In the United States, the current efforts of criminal justice reform to constitute a systemic, fundamental change in the flawed existing

1 For a collection of writings demonstrating the use of algorithms along various stages of the criminal legal process, see, e.g., Andrew Guthrie Ferguson, Policing Predictive Policing, 94 Wash. U. L. Rev. 1109 (2017); Andrew Guthrie Ferguson, Predictive Prosecution, 51 Wake Forest L. Rev. 705 (2016); Megan Stevenson, Assessing Risk Assessment in Action, 103 Minn. L. Rev. 303 (2018); Richard Berk & Jordan Hyatt, Machine Learning Forecasts of Risk to Inform Sentencing Decisions, 27 Fed. Sent'g Rep. 222 (2015); Richard Berk, An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism, 13 J. Experimental Criminology 193 (2017).

criminal legal system have brought to the forefront the question of whether algorithmic justice can serve the goal of improving the broken system.[2] Scholars have highlighted arguments for or against using algorithms in criminal justice-making processes. In this chapter, I would like to reflect on the interplay of AI-driven criminal justice and the concepts of trust and legitimacy, given the ongoing debate on the desirable ways to promote criminal justice reform.

B. Al-driven Criminal Justice: A Cosmetic or a Real Regime Change?

The crisis of faith in the current criminal legal system has sparked a pressing need for a systemic change – not a cosmetic one, but a shift towards a new regime. Christoph Burchard suggested a hypothesis explaining why algorithmic predictions have become so prevalent in the US: "Many have lost faith in criminal law as a 'big experiment' that can be turned towards progress and positive reform. The underlying conflict or conflict resolution, then, needs to be experienced as something with positive potential. If this is not the case, regime change becomes an issue - from law to algorithms; from criminal law to transformative justice".[3]

Burchard points out a distinction between the concepts of trust and faith. Trust refers to a subjective perception that one has towards someone or something in concrete cases or circumstances (it could be a person, an agency, or an institution).[4] Faith is a first-order belief that relates to the normative conceptions or the ontological foundation of a regime/institution. Therefore, someone who feels disappointment following a specific experience or an encounter with the criminal legal system might feel distrust toward the system's agents whose conduct did not meet her expectations. However, at the same time, she might still have faith in the system as a

2 See, e.g., John Chisholm & Jeffery Altenburg, The Prosecutor's Role in Promoting Decarceration: Lessons Learned from Milwaukee County, in Smart Decarceration: Achieving Criminal Justice Transformation In The 21st Century 71 (Matthew W. Epperson & Carrie Pettus-Davis eds., 2017).

3 Christoph Burchard, Musings on a Vision of Predictive Criminal Justice in Light of Trust, Conflict, Uncertainty, and Coercion (unpublished draft).

4 There is a rich body of writing in social sciences on the concept of trust. In the context of criminal law, see, e.g., Joshua Kleinfeld & Hadar Dancig-Rosenberg, Social Trust in Criminal Justice: A Metric, 98 Notre Dame Law Review 101 (2022); Kevin Vallier, Social, and Political Trust: Concepts, Causes, and Consequences (Research Paper, Knight Foundation).

whole, as it carries much more than one bad experience. The system holds a set of values, goals, and rationales that are appreciated as valuable and desirable from that person's point of view. As long as the bad experience is the exception, it might break trust, but it will not undermine faith—the deep belief in the legitimacy of the regime/institution. However, if a person repeatedly experiences more and more disappointing encounters, she might start to doubt this system or institution as a whole. When anecdotal bad experiences become systematic, individuals may lose faith.

Based on this distinction between trust and faith, Burchard explained how the crisis of faith with the traditional criminal justice system led to a willingness to make a "regime change" – a transformation from clinical predictions and human-based decisions to predictive algorithms and AI-based judgments. The supposition is that algorithmic justice can constitute a new regime—a transition from one set of principles and values to another, substantially differentiated from the traditional, old regime. But does the transition to algorithmic justice indeed mark a regime change? Does the transition from conventional, liberal criminal justice run by humans to computational-oriented, predictive criminal justice reflect a change in values and policies? Is it about a profound alteration of normative principles and values, or is it only a replacement of the mere procedures and techniques of decision-making? The answer to this question seems more complicated than it looks at first glance.

One of the fundamental functions of predictive algorithms is to foresee the future according to the past. Algorithms are fed by input representing the data that have been accumulated until the moment of processing this input. This nature of algorithms makes algorithmic predictions quite conservative. Algorithmic models are developed based on past decisions made by human decision-makers. Humans fill the algorithms with content; They determine what the algorithms should measure, what weight to give each metric, what to look at, and what to overlook. Therefore, by definition, algorithms encode underlying existing human biases and tendencies. They may even enhance such tendencies by inflating their weight as predictors of risk or other outcomes of interest.

In that sense, algorithms do not seem revolutionary. Instead, they seem an easy tool to perpetuate former preferences and normative decisions, replicating and reproducing a similar set of values and principles that were determined and adopted by human decision-makers in the past. It is true that, at a certain level, they could "clean" the process from noise or

biases.[5] AI algorithms have the potential to enhance consistency among decision-makers and construct the decision-making process to be less reliant on individual biases or propensities. Also, their improvability trait ensures they 'fix' themselves in an ongoing and iterative process to reflect the most accurate output. But essentially, algorithms are still shaped in the form of human biases. In the criminal legal system, they reflect judges', prosecutors', and police officers' biases. Some scholars, therefore, have classified algorithms as a form of bureaucratic, as opposed to democratic, criminal justice.[6] It has been argued that "they may make systems more resistant to change, especially given their tendency to reflect normative facts about the world embedded in their underlying data".[7] Algorithms preserve and amplify past policies established by criminal legal professionals. They do not absorb data representing creative, untraditional, and critical standpoints of multiple stakeholders and community members.

It is not clear then that the transition to algorithms indeed marks a substantial regime change that can revive the faith lost in the criminal justice system. Indeed, people might fall into the illusion that algorithms create a regime change. However, given the ontological character of algorithms and how they are built and operated, there is a solid reason to argue that they represent the same old thing in a different package. One may even claim that algorithms create a more dangerous representation of the old system because the "different package" hides the same old thing and, thus, causes people to develop *false* faith. In other words, people may believe that algorithms constitute a substantial regime change, whereas, in fact, it is the same old story.

5 Kahneman, Sibony, and Sunstein distinguish between noise and bias. While noise is variability in human judgment that leads to inconsistent decisions, namely random errors, bias leads to systematic errors. *See* Daniel Kahneman, Olivier Sibony, and Cass Sunstein, Noise: A Flaw in Human Judgment (2021).

6 For a distinction between notions of bureaucratic and democratic criminal justice, see a symposium issue titled "Democratizing Criminal Law," published by the Northwestern University Law Review (2016). The symposium is dedicated to the dispute over whether to promote reform in the criminal justice system by adopting a bureaucratic or democratic approach.

7 Itay Ravid & Amit Haim, *Progressive Algorithms*, 12 UC Irvine L. Rev. 527, 563 (2022).

C. Trust, Legitimacy and Accountability

At this point, I would like to add two concepts to the discussion of criminal justice in the era of algorithms. These concepts are *legitimacy* and *accountability*, and they are both connected to trust.

Let me start with the concept of legitimacy and what we know about it from the body of research on procedural justice. In the 90s, Tom Tyler and colleagues demonstrated through a series of empirical studies that procedures perceived as fair enhance the sense of governmental legitimacy. This legitimacy significantly influences legal compliance far more than the perspectives that view human motivation, primarily in terms of force and incentives.[8] Moreover, fair procedures might affect compliance more than substantive outcomes.[9] When people perceive decision-making processes as fair, namely as understandable, respectful, transparent, and neutral, they ascribe a higher level of legitimacy to the decision and the decision-maker as an authority. This increases their level of compliance.[10]

Tyler and colleagues have distinguished between two components of procedural justice. The first component, the quality of interpersonal treatment, relates to the respectful attitude given by the authorities to those affected by the decision, as well as the recognition and upholding of their rights and needs throughout the process.[11] The second component, the quality of the decision-making process, refers to whether the decision was made in a neutral, transparent, equal, and unbiased manner, whether explanations about the procedure and how the decision was reached were provided by the authority, and whether the parties were given an opportunity to voice their opinions and present their positions in a way that could influence the decision.[12]

---

8 *See* Tom R. Tyler, Why People Obey The Law (2006); Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 Crime & Just. 283 (2003); Tom R. Tyler & Yuen J. Huo, Trust In The Law: Encouraging Public Cooperation With The Police And Courts (2002).

9 Tyler, Why People Obey The Law, *ibid*, at 175.

10 *See* Tyler & Huo, *supra note* 8, at 26; Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, *supra note* 8, at 284 (2003); Jason Sunshine & Tom R. Tyler, *The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing,* 37 Law & Soc'y Rev. 513, 534 (2003).

11 Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, *supra note* 8, at 329.

12 *Ibid, ibid.*

Looking at algorithms through a procedural justice lens, it seems that pessimistic-realistic critics would question their alignment with procedural justice principles. Starting with neutrality, algorithms are allegedly supposed to be more neutral and "cleaned" from noise and human biases. They are emotionless (so they cannot be in a good or bad mood), consistent, never get tired, and can provide input based on a large number of previous cases, following an analysis based on "big picture" data, reducing the weight of outliers. Whereas emotions are inherent to human nature and have been traditionally perceived by many as a potential engine for infecting discretion and increasing irrationality, inaccuracy, and discrimination, algorithms are (still) emotionless (even though some AI tools use a language of emotions when you ask them how they feel!). Yet, as mentioned below, given that algorithmic tools are fed by humans and, therefore, may substantially rely on "dirty data,"[13] critics have pointed out that this romantic supposition is naïve and false. In fact, as explained above, algorithms might perpetuate and replicate human-created discrimination and bias under the guise of neutrality. Moreover, their rigidity hinders the ability to identify unique cases that justify deviating from the pattern.

The role of emotions in criminal justice decision-making processes raises particularly interesting questions. It is worth dwelling on the interrelation between emotions and bias or noise creation. Allegedly, emotions may interfere in applying the same decision-making process and, within the process, the same considerations in similar cases, thus potentially leading to different treatment and outcomes in similar cases. Who wants to be sentenced by an exhausted, tired judge who has not had the chance to take a lunch break (even if she is known as a decent judge)?

However, this interrelation between emotions and biases or noise seems more complicated than it might look in the first place. Despite the tendency to see the vices of emotions in infecting decision-making processes, emotions might sometimes serve as tools to fix arbitrariness and to distinguish between cases that might be perceived "on the surface" as identical if you consider only certain kinds of measurable data and ignore the broader context, which is sometimes hard to measure. Think about unique cases that do not fall into typical categories of cases. Empathy, sensitivity, intuition, and compassion can sometimes lead to a more just outcome when a

13 See Rashida Richardson, Jason M. Schultz & Kate Crawford, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice, 94 N.Y.U. L. Rev. Online 15 (2019).

combination of exceptional circumstances, characteristics, life stories, and backgrounds makes an individual case one-of-a-kind. Algorithms might do no justice in such cases. They will produce only an approximate outcome based on the closest cases they encountered, not the fairest and most just outcome that fits this unprecedented case. Since they predict based on past experience, they cannot recognize outliers and exceptions, unlike humans. Emotions might increase creativity and encourage acting and thinking less patterned and more intuitively—a virtue that is needed and welcome in unique cases.

For AI algorithms, a given case could be described as an element in a mathematical set of elements. Algorithms *cannot* "see" the people behind the case. Real-life stories are reduced and translated into a collection of facts and data. In contrast, legal professionals *can* recognize defendants, crime victims, or others involved in or affected by a criminal case as individuals, even within the overloading machinery of criminal justice. Yes, unfortunately, some professionals do not do it properly. Incisive critics of the mainstream criminal legal system may even argue that many, if not most, professionals fail to treat stakeholders humanely enough. Still, even if this is correct, it does not mean that criminal justice actors *cannot* treat stakeholders humanely. Human decision-makers in the criminal legal system can develop this human capacity; they can remind themselves daily that what they do applies to real people's lives. If they are encouraged to do so (e.g., by relevant incentives set by the system), they will be *able* to do so. For this to happen, emotions must be granted pride of place because emotions are essential for "translating" a case – an element in a mathematical set, into a story of an individual—a human being with a unique life experience. Algorithms, at least for now, do not have this capacity. Instead of a human being, they "see" an abstract element.

Since algorithms cannot feel, something important might get lost in their decision-making process, which applies to humans, particularly in the criminal justice context, where moral judgments are so integral. One of the episodes of the British anthology television series "*Black Mirror*" demonstrates how emotions are essential for making moral decisions.[14] The episode shows that when soldiers put on glasses that transform the figures they see from humans to mutated humans called "roaches," and when they do not know that the figures they see are, in fact, regular people,

14 "Black Mirror", season 3, episode 5 "Man Against Fire" (written by Charlie Brooker, 2016).

they lose their compassion. Algorithms can be metaphorically compared to someone who produces output with compassion-blocking glasses. Under such conditions, respectful treatment as a procedural justice component cannot be provided. Moreover, the outcome itself might not be fair, not just the procedure. It turns out that, on the one hand, algorithms are not free from the vices of distorted emotions embedded within the data they underly. On the other hand, they cannot benefit from the virtues of emotions humans have and use in extraordinary cases to fix arbitrariness and make justice in idiosyncratic cases.

In addition, the use of algorithmic tools in the criminal justice system undermines transparency.[15] This argument is primarily based on the inherent obscurity of algorithmic systems, particularly those that use deep learning and remain opaque or hidden from human comprehension. This problem has been termed the "black box problem," which means that "observers can witness the inputs and outputs of these complex and non-linear processes but not the inner workings,"[16] including observers with computational expertise. This lack of clarity potentially conflicts with legal standards that require clear reasoning behind decisions, particularly in the context of criminal law, where the stakes are high.

To sum up, even if algorithmic tools could increase the chances of reaching the most accurate and just outcome (and as explained, this is a big question in itself), we might lose a human-friendly process that people expect to experience to acquire legitimacy for its outcome and maintain their faith in the criminal legal system. Doing justice is essential, but the appearance of justice also has its merits. In other words, while reaching the right outcome is essential, the process of reaching that outcome can be no less important, and even more important to people affected by this outcome, to see it as legitimate and, therefore, to comply with it. Furthermore, decision-making processes that are not aligned with procedural justice principles might sometimes lead to unjust substantive outcomes. As Kevin Vallier stated when discussing the best ways to cultivate trust among members of a polity, "[w]e generally want social trust to be sustained for the right reasons. Pouring the 'trust hormone' Oxytocin into the water supply might make people more trusting, but it is not a good way to promote

15 See, e.g., Alyssa M. Carlson, Note, The Need for Transparency in the Age of Predictive Sentencing Algorithms, 103 Iowa L. Rev. 303 (2017).

16 Warren J. Von Eschenbach, Transparency and the black box problem: Why we do not trust AI, 34 Philosophy & Technology 1607 (2021).

social trust. It is better to sustain social trust by giving persons morally appropriate incentives to be trustworthy, and then allowing social trust to form as a free cognitive and emotional response to observed trustworthy behavior".[17] An analogy can be drawn to the context of developing trust (on an individual case) and faith (on a systemic ground) in an algorithmic regime of criminal justice: we should not (and probably even cannot) instil trust and faith in people by memorizing a mantra of "we believe in algorithms because they make justice!". Algorithms need to be trustworthy; this can happen if they are transparent, neutral, and, at the same time, sensitive and creative; this can happen if they are open to absorbing broader information representing multiple perspectives and considerations that are sometimes hard to capture through measurable metrics. In such a reality, people will develop sustainable trust and faith as a natural cognitive and emotional response.

Another factor that influences trust is accountability. Individuals have the right to understand the decisions made by public officials in their cases. Margot Kaminski and Jennifer Urban argued that if decision-makers cannot explain their decisions, as is often the case with decisions made by algorithms, it violates the basic expectation of the individuals affected.[18] For decision-makers, the option to rely on algorithms might reduce accountability.[19] If a decision primarily depends on algorithmic judgement, it offers a strong "defence claim" to a human decision-maker. Instead of explaining and justifying a decision by using an independent judgment, decision-makers might tend to overly rely on algorithmic recommendations because, in this way, they can attribute responsibility to an external entity – the machine ("Hey, it's the algorithm, not me!"). [20] This phenomenon has been known as the "automation bias." However, when a decision depends mainly on human discretion, a potential decision-maker realizes that her discretion would be the subject of scrutiny. She will be evaluated and promoted (or not) by her decisions. Therefore, she will need to explain why she decided this and not that. The requirement to account for how and why she made

17 Vallier, supra note 4, at 4.

18 Margot Kaminski & Jennifer Urban, The Right to Contest AI,121 Columbia L. Rev. 1957 (2021).

19 See Kate Crawford & Jason Schultz, Al Systems as State Actors, 119 Columbia L. Rev. 1941 (2019).

20 Linda J. Skitka, Kathleen Mosier & Mark D. Burdick, Accountability and Automation Bias, 52 Int'l J. Hum.-Comput. Stud. 701 (2000).

that specific decision creates incentives to be more cautious, balanced, prudent, and accountable with her decisions.

D. Trust in an Era of Algorithms: A Look to the Future

What, after all, can be done to minimize the risks of utilizing algorithms in the criminal legal system without giving up the benefit of using them? How can we enhance trust and faith in an AI-based predictive criminal justice system that is not false but authentic and justified? And how can the algorithmic regime work hand-in-hand with other turns and trends in the criminal justice system seeking to reform systemic problems and make the system more democratic?

These questions portray some of the challenges the future holds. Potential solutions should combine additional toolsets of human checks and balances along the decision-making processes in the criminal justice context. One suggestion, for instance, is to combine the use of algorithms with other reforms that deviate from traditional principles and values but are based on human discretion. Itay Ravid and Amit Haim suggested designing what they call "progressive algorithms."[21] Their proposed decision-making model prioritizes accountability, transparency, and democratization principles by adopting progressive prosecutors' agendas and using them as the content according to which computational methods and algorithms would be designed. As Ravid and Haim explain, at first glance, it seems that the trends of progressive prosecutors and algorithmic justice are fundamentally at odds over a crucial issue in criminal justice reform—what role do humans play, and what potential do they have in driving systemic change: "While the promise behind the progressive prosecutors' movement puts the keys to resolving the criminal justice system's problems in the hands of humans, the computational decision-making trend sends a whole different message: the solution will arrive by limiting the presence of human discretion in the criminal process."[22] They suggest that a model combining the two trends can reconcile the alleged paradox of having both trends coexist.

Another suggestion is to consider deliberative frameworks for adopting algorithms that incorporate professionals' diverse agendas and standpoints, reflecting the cacophony of individual and public interests embedded with-

21 Ravid & Haim, supra note 7.
22 Ibid, at 531.

in the criminal justice endeavour. Since algorithms promote efficiency by saving much time on complex technical calculations, decision-makers can use the time saved to reach balanced agreements about the weight given to various kinds of complementing or contradicting considerations by the algorithms, promoting a more transparent, democratic process of algorithmic design.

A less mediated way to incorporate public views and concerns about algorithmic design and operation is to adopt a participatory framework that would consider lay stakeholders' perceptions and perspectives in addition to the professionals. An actual example of such a framework is a process undergone by the Pennsylvania Commission on Sentencing to adopt sentence risk-assessment instruments.[23] The implementation of these tools included a participatory process in which various community stakeholders, policymakers, and legal professionals were invited to contribute their input in open public hearings. Such processes can suggest ways to address concerns about the lack of democratization and transparency, which erode trust and legitimacy.

Indeed, the million-dollar question is whether humans are the inevitable solution or the root problem in the era of algorithmic criminal justice. I believe that as long as machines do not become humane (and at least for now, they don't!), human discretion must be involved in criminal decision-making processes to make them trustworthy. The notions of trust, faith, legitimacy, and accountability can help illuminate the dilemmas pertaining to the desirability of an algorithmic regime in the criminal context. Future studies should use empirical tools to uncover the public perceptions of AI's role in criminal justice and the interplay between trust, legitimacy, and the AI-based criminal justice system. Understanding the public sentiments as a significant factor driving trust can help ensure that reforms align with democratic values and accountability standards.

23 See https://www.pacodeandbulletin.gov/Display/pabull?file=/secure/pabulletin/data/vol50/50-3/60.htmlId, as mentioned in Ravid & Haim, supra note 7, at 564.