

RESEARCH ARTICLES

Ethical AI Sentencing: A Framework for Moral Judgment in Criminal Justice

Sienna Sedlarcik^a

Critical Debates in Humanities, Science and Global Justice

Vol. 4, Issue 1, 2024

Imagine standing in a courtroom as a defendant, your fate hanging out of your hands. However, something is different. It's not just the judge and the law determining your future, there is something else involved: a computer algorithm. As you await your sentence, you begin to feel a sense of unease. How could something so cold and detached, incapable of grasping the complexities of human life, have so much power over your future? Is this really justice? This was the reality for Eric Loomis (Liu, 2019). In 2013, he was sentenced to 6 years in prison, but it wasn't the judge who determined his destiny. The decision was heavily influenced by COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), an algorithm designed to assess the risk of reoffending by analyzing data. Loomis argued that his sentence was unfairly influenced by a tool he couldn't fully understand or challenge, as it labeled him as having a "high risk of recidivism" and therefore significantly increased the length and severity of his sentence. Loomis lost his case, which soon became a catalyst for AI to be used more openly and often in the criminal justice system. However, during Loomis' verdict, Justice Abrahamson highlighted the court's lack of understanding of COMAS, highlighting the need for transparency and explanation of evidence-based tools in sentencing. His case, *Wisconsin v. Loomis*, began the national debate about the role of algorithms in the justice system and raised important questions about the fairness, transparency, and the future of criminal sentencing.

AI is no longer just a myth in a sci-fi movie. It has taken hold of the world, from industries of healthcare to marketing. By the end of 2023, the market for AI amounted to around \$200 billion, and is expected to grow to over \$1.8 trillion by 2030 (Chui, 2024). But perhaps the most significant and contentious application of AI is in the criminal justice system—a system that serves as the foundation of societal order and justice. This isn't just another industry; it is a critical pillar that impacts every individual in society. Its decisions determine freedom, justice, and sometimes even life and death.

As a result, many, like Loomis, have raised questions about the moral legitimacy of incorporating AI into the criminal justice system that is traditionally based upon human judgment.

^a Sienna Sedlarcik is a high school at Staple High School in CT.

INTRODUCTION

Imagine standing in a courtroom as a defendant, your fate hanging out of your hands. However, something is different. It's not just the judge and the law determining your future, there is something else involved: a computer algorithm. As you await your sentence, you begin to feel a sense of unease. How could something so cold and detached, incapable of grasping the complexities of human life, have so much power over your future? Is this really justice?

This was the reality for Eric Loomis (Liu, 2019). In 2013, he was sentenced to 6 years in prison, but it wasn't the judge who determined his destiny. The decision was heavily influenced by COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), an algorithm designed to assess the risk of reoffending by analyzing data. Loomis argued that his sentence was unfairly influenced by a tool he couldn't fully understand or challenge, as it labeled him as having a "high risk of recidivism" and therefore significantly increased the length and severity of his sentence. Loomis lost his case, which soon became a catalyst for AI to be used more openly and often in the criminal justice system. However, during Loomis' verdict, Justice Abrahamson highlighted the court's lack of understanding of COMAS, highlighting the need for transparency and explanation of evidence-based tools in sentencing. His case, *Wisconsin v. Loomis*, began the national debate about the role of algorithms in the justice system and raised important questions about the fairness, transparency, and the future of criminal sentencing.

AI is no longer just a myth in a sci-fi movie. It has taken hold of the world, from industries of healthcare to marketing. By the end of 2023, the market for AI amounted to around \$200 billion, and is expected to grow to over \$1.8 trillion by 2030 (Chui, 2024). But perhaps the most significant and contentious application of AI is in the criminal justice system—a system that serves as the foundation of societal order and justice. This isn't just another industry; it is a critical pillar that impacts every individual in society. Its decisions determine freedom, justice, and sometimes even life and death.

As a result, many, like Loomis, have raised questions about the moral legitimacy of incorporating AI into the criminal justice system that is traditionally based upon human judgment. Specifically, this paper investigates the following question: what framework or model should we use to balance the benefits, such as expediency and efficiency, of AI sentencing against the risks, such as the fact that algorithms might perpetuate socioeconomic disparities? In this paper, I will propose a threshold model for balancing the benefits against the risks of AI in sentencing: the model sets a maximum for acceptable level of risks, beyond which the use of AI in sentencing should be deemed unjust and rejected. Furthermore, in order to ensure the model's effective implementation and foster public trust, it will be paired with public oversight mechanisms, ensuring transparency and accountability in practice.

1. AI IN SENTENCING: BENEFITS AND MORAL RISKS

The involvement of AI in the criminal justice system provides numerous compelling benefits. AI has various uses in sentencing, including facial recognition and DNA analysis to assist with evidence. It also provides risk assessments, explained in the intro, as well as automated case management and document analysis (Bush-Joseph, 2024). In what follows, I will focus on the latter aspects of AI's functions in the criminal justice system.

AI's ability to process and analyze data at remarkable speeds can greatly expedite the legal process. Complex cases that traditionally take months to resolve can be more easily digested with the use of AI, reducing delays and backlog. As of 2022, US courts faced a backlog of over 700,000 pending cases, a number that continues to grow (Bush-Joseph, 2024). This costs US taxpayers over \$14 billion annually to detain those awaiting trial. AI tools can help reduce these costs and streamline case management. AI can review vast amounts of legal documents, evidence, and case law in a fraction of the time it would take a human team, ensuring that cases move through the system more efficiently. Indeed, according to one estimate, the widespread use of AI in the legal system could potentially save the U.S. up to \$1.7 billion annually (Chui, 2024). By increasing the speed and efficiency of court cases, AI not only improves the justice system but also provides significant economic benefits.

Another significant advantage of AI in sentencing is its potential to help reduce irrationalities and inherent biases in human decision-making. Human decisions, especially in courtrooms, can be influenced by a myriad of factors, some of which are unconscious. This is a major flaw in the criminal justice system. For instance, a study found that judges were more likely to make a more favorable decision in parole hearings immediately after lunch (Taylor, 2023). AI, on the other hand, operates without these irrational fluctuations, ensuring greater consistency in sentencing and parole decisions.

However, the involvement of AI in sentencing comes with many disadvantages and ethical concerns. One critical disadvantage is the lack of transparency in AI algorithms (Schwerzmann, 2021). While it would be ideal to ensure full transparency in the AI decision making process, this is simply not currently realistic. Additionally, court cases are, on average, 50% more efficient, and, as stated before, save \$1.7 billion in US taxpayer dollars. However, some estimates suggest that the transparency and explainability of AI decisions are relatively low, with around 20–30% of the data and reasoning behind decisions understandable to an average person. This means that in about 70–80% of cases, the underlying reasoning behind AI decisions are not fully understandable to users or auditors (Schwerzmann, 2021).

Moreover, there are often socioeconomic disparities, especially racial disparities, created and perpetuated with the use of AI algorithms when sentencing (Mayson, 2018). These algorithms are typically trained on historical data which often contain racial biases. For instance, historical cases like *Dred Scott v. Sandford* (*Scott v. Stanford*, 1857) and *Strauder v. West*

Virginia (*Strauder v. State*, 1880) illustrate deep-rooted racial injustices in the legal system. For a more recent example, in *State v. Robinson* (2020), it was determined that racial bias had heavily influenced Robinson's death penalty sentence, leading to its reversal. This case highlights a situation where racial bias was identified and corrected, but many instances of such bias are less apparent and not caught, and therefore can easily be embedded into AI systems, perpetuating systemic injustice.

If the data we use to train this technology comes directly from past cases the decisions of AI will most likely reflect past racial biases, discrimination in policing, sentencing, and more. As a result, AI tends to overestimate black people's relative riskiness and underestimate that of white people (Mayson, 2018). Algorithms predicting higher recidivism rates for certain racial groups can lead to disproportionately harsher sentencing, reinforcing a feedback loop of racial disparity. Additionally, black male offenders received sentences that were 13.4% longer than those of white male offenders, and black defendants were 38% more likely to be sentenced to death in the US than white defendants in cases with similar circumstances (U.S. Sentencing Commission, 2023). The challenge lies in integrating AI in a way that supports, rather than undermines, the fairness and humanity of the justice system.

In addition to the above-mentioned concerns, there is an even more fundamental concern about the moral status of such algorithms: lacking sentience, it seems that AI algorithms are not geared to be involved in morally relevant decision-making. This concern will likely remain even if AI algorithms are greatly improved in the future, and even if they can inherit all the merits of human judgment and avoid all the human weaknesses. To see why, imagine a case with two hypothetical judges: Judge Judy, who relies on her own judgment for sentencing, and Judge Joe, who strictly follows an AI program's recommendations. Is the morality lost when judges like Joe fully rely on AI? Even if AI could match human judgment, does the complexity and nuances of sentencing might still require human decision-making? These questions will be discussed in further sections. This case displays the importance of setting a framework for the supervision of AI so that it doesn't take full, unchecked power of sentencing decisions, as exhibited through Judge Joe. The challenge lies in developing an ethical model that effectively balances the benefits of AI with its potential risks. Furthermore, the ethical implications extend to both the convicted individuals and the victims. The court system is used as a condemnatory function, which is defined as the expression of society's disapproval and moral judgment of criminal behavior, often through punishment or sentencing. Therefore, victims aren't gaining the justice and societal support they deserve, meaning AI lacks the moral status to make these significant decisions.

2. BALANCING THE BENEFITS AGAINST THE RISKS: A FRAMEWORK

As AI becomes increasingly prevalent in the criminal justice system, the focus has shifted from *if* AI should be used to *how* it should be integrated responsibly (Caldwell et al., 2020); very few people still think that AI should not be used in criminal sentencing at all (though see Schwerzmann, 2021 for arguments that we should abolish the use of AI in criminal sentencing). However, given the potential risks of AI, how can we justify the integration of AI into the criminal justice system? One might be tempted to view this through a utilitarian lens: there are compelling reasons to believe that the potential benefits of AI — such as increased efficiency, consistency, and transparency — can outweigh the potential risks, if handled correctly, in order to support the greater good of the public. However, there are problems with this utilitarian approach. On the one hand, it undermines fairness and justice. Although the majority of the public may benefit from the implementation of AI into the criminal justice system, the defendants are at risk of being at a large disadvantage. It is crucial to address the moral imperative of ensuring fairness, particularly in light of the racial biases that AI systems may inadvertently perpetuate. Fairness, a principle deeply rooted into our Constitution and Bill of Rights, sits at the cornerstone of our society, often taking precedence over efficiency. For instance, the Sixth Amendment to the Constitution guarantees the right to a jury trial. This process may be slow, less efficient, and costly, but ensures that a defendant's fate is determined by a group of people, rather than a single judge. Therefore, it is important to respect and continue this idea of sacrificing expediency for justice, while also evolving these amendments created over two centuries ago to mold into today's society. Furthermore, as we shall see in the next section, it is challenging to calculate the expected utility and net benefits of using AI in sentencing as it is difficult to measure or compare its advantages and disadvantages.

Therefore, I propose that we adopt another framework in order to maintain an ethical, yet efficient integration of AI into the criminal justice system that avoids the above-mentioned problems with the utilitarian framework. This framework contains two components which I shall explain in detail in the next two sections. First, it sets *a maximum level of risks* (e.g., a maximum level of racial bias) in AI sentencing that we are willing to accept to ensure that fairness and justice are upheld. It also asserts a “trade off” structure to effectively balance the benefits of efficiency, savings, and consistency, against risks, such as transparency. I will explain this threshold model in Section 4. Second, the framework also emphasizes the importance of earning public support and trust, as this is one of the biggest challenges. To achieve this, I will outline a model in Section 5 for implementing effective public oversight and ensuring that the public has a meaningful voice in decision-making.

3. ALGORITHM TRANSPARENCY, BIAS, AND AUDITING

Let us start with the problem that AI algorithms lack transparency. Something that has been implemented into AI more commonly is Explainability Scores, known as xAI (Deeks, 2021), which is a metric used to assess how easily and effectively an AI system's decision-making process can be understood by humans, including non-experts. The concept is that Explainability Scores "provide insight into the internal state of an algorithm, or to human-understandable approximations of the algorithm." (Deeks, 2021, p. 1834). This technology can be easily accessed by sentencing AI systems, and work as a consistent check and baseline for how they operate. It can be used for COMPAS, Public Safety Assessments, PredPol (predictive policing), case management systems, and more. If a case decision has an Explainability Score of 80 out of 100, this would suggest that the system is quite transparent, as about 80% of the reasoning for a particular decision is easily understandable to the average person. However, a system with a score of 30 out of 100 would be considered a "black box"—its decision-making process would be opaque, making it difficult for users to understand, trust, or challenge the AI's conclusions. Opaque algorithms like these can undermine one's sense of fairness and trust, particularly when utilized by the government. Currently, the decisions of AI are a "black box" due to its low Explainability Scores, with the majority being 20-30 (Deeks, 2021). Therefore, my proposal is to mandate that each AI system be equipped with an Explainability Score to ensure a standardized measure of transparency and accountability. Although the use of COMPAS was allowed in Loomis' case, Justice Abrahamson acknowledged that "this court's lack of understanding of COMPAS was a significant problem in the instant case." There are numerous ways AI can reach a high explainability score, one being a "subject-centric" approach. This would provide individuals with details about similar decisions made by the algorithm, allowing them to see how different factors (e.g., age, number of arrests, etc.) would have influenced the algorithm's recommendation if altered. Another is an exogenous "model-centric" approach. This would mean creating simpler models (like decision trees) to offer a more understandable view of how the algorithm operates and how it created specific decisions (Deeks, 2021).

To address the trade-offs between efficiency and transparency in AI's role in the criminal justice system, a balanced framework should be established. Setting a minimum Explainability Score of 75 out of 100 ensures a reasonable level of transparency while still leveraging the benefits of AI. Additionally, I propose that a minimum Explainability score of 80 out of 100 should be enforced for criminal cases. This score is selected based on industry standards, aiming to provide sufficient clarity about AI's decision-making processes without compromising efficiency. To quantify trade-offs, the framework would propose that sacrificing one point in the Explainability Score corresponds to either \$X million in taxpayer savings or a X% increase in case processing efficiency, as long as the explainability score remains above

the 75-80% threshold. For the purpose of this paper, I am just proposing this framework, not specific numbers, as this will most likely require trial and error when implemented to create exact trade offs. For example, an opportunity of 5% more efficiency in data analysis in a criminal case would mean sacrificing 2 points of an Explainability Score from a 90-88. This would be acceptable, as it is still over the 80 out of 100 minimum. This approach provides a clear method for evaluating the trade-offs between transparency, efficiency, and cost savings. Additionally, the framework should include provisions for continuous monitoring and adjustment to address potential biases and ensure that AI systems uphold ethical standards. By implementing these guidelines, we can strive for a fair integration of AI that balances practical benefits with ethical considerations.

Unlike transparency and other factors of AI, bias and racism in decision-making are difficult to precisely measure quantitatively to compare, making it necessary to establish a new framework to address these issues. Human nature inherently involves biases for or against certain groups, which remains a significant flaw in the criminal justice system. For instance, on August 3, 2023, the Sixth Circuit Court of Appeals overturned Leon Liggins' drug conviction after the federal district judge presiding over his trial openly remarked that Liggins, an African American man, "looks like a criminal to me." (Che, 2023). Research reveals that "implicit bias accounted for roughly six percent of the variation in actual behavior" (Rachlinski, 2009, p. 1201). With judges handling approximately twenty-one million cases at the state level and seventy thousand cases at the federal level annually, this translates to about 1.26 million defendants facing outcomes influenced by racial bias. Additionally, Black males receive sentences that are on average 13.4% longer than White males, while Hispanic males receive sentences 11.2% longer.

Nevertheless, these numbers are challenging to measure with precision. It is difficult to discern whether these disparities stem from racial biases or from socioeconomic factors related to upbringing. Given this context, it is imperative that AI systems in the criminal justice system are held to a high ethical standard. Though ideal, the current goal is not to introduce a completely transformative solution but to integrate AI in a way that mitigates existing biases. A clear standard must be established to ensure that AI does not exceed the level of the impact of racial bias on judge's decision-making observed in the human case. Furthermore, when comparing outcomes across demographic groups, AI should demonstrate equal or improved consistency and fairness in sentencing compared to human judges.

Additionally, to ensure the consistent performance and fairness of AI systems, it is crucial to implement regular audits and evaluations of algorithms, ideally on a quarterly basis (Kim, 2017). However, conducting a bias check for every individual case is impractical. Instead, AI systems should be designed to self-monitor and detect biases in their decisions. Utilizing diverse datasets for training can help minimize the risk of discriminatory outcomes. Additionally, AI algorithms should be programmed to cross-check

each other against established frameworks to ensure adherence to fairness and transparency standards. This approach not only streamlines the monitoring process but also reinforces the commitment to equitable AI practices.

4. LEGAL GUIDELINES AND PUBLIC OVERSIGHT

The greatest challenge facing AI in the criminal justice system, despite the strategies proposed in the previous section, is earning public trust and moral approval for its decisions. To achieve this, it is essential to demonstrate that AI sentencing practices align with current laws and constitutional principles, and do not rely on outdated precedents that do not reflect contemporary values of justice and individual rights. For example, in the United States, adherence to the Equal Protection Clause of the Fourteenth Amendment (Wilkinson, 1975), ratified in 1868, is crucial to ensuring that AI does not perpetuate biases present in historical data from before its enactment. It is also important to understand various legal standards set in place, specifically the Administrative Procedure Act (APA). This law is intended to ensure that agencies provide a satisfactory explanation for their decisions (Rubin, 2003). However, it has been out of date since the day it was written, and has not been upheld. Therefore, it is critical to implement this Act further in order to maintain justice and order with the use of AI algorithms in criminal justice sentencing.

AI should never be the sole determinant of legal decisions; judges and juries, who are knowledgeable about the specifics of each case, must retain ultimate control and accountability. This ensures that verdicts reflect current societal values and legal standards. Furthermore, involving elected representatives in evaluating the credibility and accuracy of AI assessments is vital for gaining public trust. This is critical because only humans possess a “front row” perspective on current public movements, evolving values, and societal changes.

Another valuable approach is the concept of *meaningful public control*, as proposed by Isaac Taylor (2023). Many algorithms, especially those used in critical areas like criminal justice, are developed by private companies with commercial interests. To ensure these algorithms align with broader societal values and ethical standards, it is essential to shift their oversight to public entities. Public oversight would ensure that the development and implementation of AI technologies are conducted transparently and in accordance with legal standards.

Involving the public in ongoing discussions about the ethical and social implications of AI is crucial. This engagement allows for the collection of diverse perspectives and feedback, which can help address potential biases and ensure that AI systems reflect contemporary societal norms and values. By creating forums for public input and integrating this feedback into the development process, we can foster greater accountability and trust in AI technologies. Additionally, public control mechanisms, such as regulatory bodies and advisory committees, can provide oversight and ensure that AI

applications do not deviate from established ethical guidelines and societal expectations. This approach helps to ensure that AI technologies serve the public good while respecting and upholding Constitutional values.

CONCLUSION

In conclusion, integrating AI into the criminal justice system requires a careful balance between its benefits and ethical considerations. While AI has the potential to enhance efficiency, consistency, and cost savings, it also raises important concerns about fairness and transparency. To address these issues, a comprehensive framework is essential.

Firstly, ensuring algorithm transparency is crucial. Implementing ExplainabilityScores, with a minimum of 75 for general use and 80 for criminal cases, along with regular audits, can mitigate concerns about the “black box” nature of AI systems. This will provide clarity on decision-making processes and enhance trust in the technology.

Secondly, addressing biases is vital for maintaining fairness. High ethical standards must be set, with a goal of keeping racial bias below that of humans, and ensuring an even distribution across different demographics. Regular evaluations of AI systems should be conducted to prevent the perpetuation of existing disparities.

Lastly, AI should serve as a tool to assist, not replace, human judgment. Judges and juries must retain ultimate authority to ensure decisions reflect contemporary values and legal standards. Public trust and moral approval are foundational for successful AI implementation. This requires meaningful public oversight and engagement, shifting AI development and monitoring to public entities to align with democratic principles. By fostering transparency, accountability, and inclusivity, we can harness the benefits of AI while upholding justice and fairness.

Published: November 05, 2024 EDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

References

- Bush-Joseph, K. (2024, February). *Outmatched: The US asylum system faces record demands*. Migration Information Source. <https://www.migrationpolicy.org/research/outmatched-us-asylum-system>
- Caldwell, M., Andrews, J. T., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science*, 9(1), 1–13. <https://doi.org/10.1186/s40163-020-00123-8>
- Che, C. (2023, August 4). Conviction reversed over judge's remark that Black man 'looks like a criminal.' *The New York Times*. <https://www.nytimes.com/2023/08/04/us/judge-remark-black-man-criminal.html>
- Kim, P. T. (2017). Auditing algorithms for discrimination. *U. Pa. L. Rev. Online*, 166, 189.
- Liu, H. W. (2019). Beyond State v Loomis: Artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology*, 27(2), 122–141. <https://doi.org/10.1093/ijlit/eaz001>
- Mayson, S. G. (2018). Bias in, bias out. *Yale L.J.*, 128, 2218.
- Rachlinski, J. J. (2009). Does unconscious racial bias affect trial judges. *Notre Dame L. Rev.*, 84, 1195.
- Rubin, E. (2003). It's Time to Make the Administrative Procedure Act Administrative. *Cornell L. Rev.*, 89, 95. <https://doi.org/10.2139/ssrn.412584>
- Schwerzmann, K. (2021). Abolish! Against the use of risk assessment algorithms at sentencing in the US criminal justice system. *Philosophy & Technology*, 34(4), 1883–1904. <https://doi.org/10.1007/s13347-021-00491-2>
- Taylor, I. (2023). Justice by algorithm: The limits of AI in criminal sentencing. *Criminal Justice Ethics*, 42(3), 193–213. <https://doi.org/10.1080/0731129X.2023.2275967>
- U.S. Sentencing Commission. (2023). *Demographic differences in federal sentencing*. United States Sentencing Commission.
- Wilkinson, J. H., III. (1975). The Supreme Court, the Equal Protection Clause, and the three faces of constitutional equality. *Virginia Law Review*, 945–1018. <https://doi.org/10.2307/1072429>