The Computing Research Association (CRA)'s Computing Community Consortium (CCC) Response to the National Institute of Justice's Request for Input on Section 7.1(b) of Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence"

This response is prepared by the Computing Research Association (CRA)'s Computing Community Consortium (CCC) by inviting CCC Council members and other members of the research community with interest and knowledge of the use of AI in justice-related scenarios to a roundtable discussion. The participants discussed the RFI and contributed to this written response document. CRA is an association of nearly 250 North American computing research organizations, both academic and industrial, and partners from six professional computing societies.

The mission of the CCC, a subcommittee of CRA, is to enable the pursuit of innovative, high-impact computing research that aligns with pressing national and global challenges. Please note any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the authors' affiliations, or of the National Science Foundation, which funds the CCC.

May 28, 2024

Written by: Nadya Bliss (Arizona State University), Kevin Butler (University of Florida), David Danks (University of California, San Diego), Stephanie Forrest (Arizona State University), Catherine Gill (Computing Community Consortium), Daniel Lopresti (Lehigh University), Mary Lou Maher (Computing Community Consortium), Helena Mentis (University of Maryland, Baltimore County), Cris Moore (Santa Fe Institute), Shashi Shekhar (University of Minnesota), Amanda Stent (Colby College), and Matthew Turk (Toyota Technological Institute at Chicago).

Summary of Key Points

1. In domains like criminal justice where constitutional rights are central, any AI system used to make decisions or recommendations to a human decision-maker should be transparent. The procedure by which humans take the AI system's outputs into account should also be transparent: for instance, how its outputs will be presented to a judge, whether police will use predictive policing or facial recognition systems as probable cause or just as a lead to be backed up by other evidence, and so on.

2. "Transparency" means a clear understanding of what input data the AI system uses, how this data is collected and curated, and how the AI system processes this data to produce its results. AI systems whose internal workings, input factors, weights, or data sources are proprietary should play no role in pretrial detention, sentencing, parole, or prison classification. This requirement, along with providing for independent audits of an AI system's performance, should be included in procedures for the procurement and use of AI in the justice system.

3. Risk assessment systems should give specific, quantitative outputs, such as "7% probability of rearrest for a violent felony during the pretrial period", as opposed to categorical labels like "8 out of 10" or "high risk". These labels are open to misinterpretation and humans typically overestimate the corresponding probabilities. Risk assessments should draw distinctions between the risk of violent and non-violent charges, between felonies and misdemeanors.

4. AI systems should be used if and when their use will actually improve the accuracy and fairness of the justice system overall. This means understanding the performance of the justice system as a sociotechnical system, including both AIs and the humans in the loop. Research is needed on how judges and other human decision-makers respond to AI systems' outputs, and how estimates of probabilities and confidence levels can be presented in meaningful and helpful ways.

5. Risk assessments should not be viewed as making predictions about individuals, but rather as providing average risk levels for defendants with similar records. The judge or other decision-maker should then consider individualized evidence in the judicial process, provided by the prosecution and the defense, that distinguishes the defendant from this reference class.

6. Data about a defendant's living situation, employment, family, upbringing, education, etc. may be useful for recommending supportive services. But they

should never be used in taking punitive actions such as detention or increasing the level of supervision.

7. AI systems should never replace judicial decision-making: they should at most make recommendations to human decision-makers. But at their best and most transparent, they can assist judges and other decision-makers. The data behind them can also help advance policy discussions — including around criminal justice reform — by pointing out that large groups of defendants are rarely dangerous to the public, that infractions that took place years ago are not predictive of inmate misconduct, and so on.

How we define Artificial Intelligence

For the purposes of our response regarding recommendations on the use of artificial intelligence in the criminal justice system, we discussed our interpretations of what constitutes AI. We present a definition here that most authors agreed with and note that it is important to define AI when developing policies and regulations regarding AI. “AI”, as it is commonly understood, spans a wide range of methods, from relatively simple algorithms such as logistic regressions whose weights are determined by applying machine learning techniques to training data, to more advanced methods like deep neural networks and large language models. The mere use of computers or spreadsheets does not constitute AI, unless that spreadsheet carries out some automated calculation that produces a recommendation or risk score intended to guide decision-making. Our recommendations below, including for transparency and independent audits, apply to this entire range of AI systems. It should be noted, however, that more advanced systems are less transparent and harder to audit.

Where constitutional rights are involved, transparency is paramount

Denying a citizen their physical liberty is one of the most fundamental and momentous actions a government can take. A person who is detained deserves a full explanation of why they are detained, and a meaningful opportunity to contest this decision. If this decision is supported in part by an AI system, then that person, and their defense counsel, need to know what data about them was used by the AI system, where this data came from, and the logic by which its recommendation was produced from this data.

This requires that any AI system used for criminal justice be transparent, as opposed to a “black box” that produces outputs using a hidden process. The idea that an opaque

system — which neither defendants, nor their attorneys, nor their judges understand — could play a role in major decisions about a person's liberty is repugnant to our individualized justice system. An opaque system is an accuser the defendant cannot face; a witness they cannot cross-examine, presenting evidence they cannot contest.[1]

This is clearest in pretrial detention, where a defendant who has not yet been found guilty of any crime is detained until their trial (or rather, in most cases today, until they plead to a lesser charge). However, the same principle applies to post-conviction decisions as well, including sentencing, prison classification, probation and parole. The people affected by an AI system's decisions, and their legal counsel, should have access to that system's reasoning.

Beyond transparency of the systems themselves, transparency also applies to the decision-making process. How and when AI systems and their outputs are involved in the judicial process must be standardized and all parties involved in a particular case must be informed when the results of an AI system contribute to a particular decision. Standardizing procedures regarding when AI can and should be used will also improve the processes of auditing and evaluating the use of AI in the criminal justice system.

Transparency vs. open source and explainability

By "transparency" we do not mean making an AI system "open source," i.e., publishing the source code of its program. This is neither sufficient nor necessary. Providing the connections and coefficients of a deep neural network or a large language model doesn't give us a human-understandable explanation of how it combines its inputs and arrives at its decisions. Nor does it explain how that network was trained, or what mistakes it might make.

Conversely, if the mathematics behind an AI system is clear, the particular software a developer uses to implement it is immaterial as long as it works as advertised and its results can be reproduced by independent actors. Thus transparency means a clear description of the internal workings of an AI system: the mathematical and logical steps it uses to produce its output.

Some tech companies complain that requiring this kind of transparency would violate their intellectual property rights, reveal trade secrets, or retard innovation. We reply that opaque, proprietary AI systems might be appropriate in many domains — recommending movies, translating speech, and so on — but they should not play a role in the justice system of a society that values individual rights and an accountable

[1] Andrea Roth, *Machine Testimony.* Yale Law Journal 126:1972 (2017).

system of government. In fact a number of tech companies have successful business models where the AI system is transparent, but the company provides helpful interfaces, integration with databases and GIS systems, technical support, and so on. If the government is considering the use of an AI system, the public has every right to require transparency, and this requirement should be implemented in procurement policies.

It is also important to distinguish transparency from the weaker notion of “explainability,” at least as the latter is currently used in computer science. “Explainable AI” or XAI is an active research area where the decisions of an otherwise opaque AI system are partially explained by identifying the variables that had the greatest impact on a decision, or the nearest counterfactual that would have led to a different decision. However, it does not require that the internal logic of the AI system be disclosed.

Explainable AI is analogous to the Fair Credit Reporting Act. The FCRA does not require that companies that produce credit scores, such as Transunion, Experian, and Equifax, disclose their algorithms. It does require that consumers be able to download, for free, the data about them that these systems use as input. Consumers can then contest this data (e.g. if it incorrectly states that they failed to repay a loan) and request that their credit scores be updated accordingly. But they cannot question the AI system itself.

Where constitutional concerns play a role, a deeper level of transparency is necessary. Defendants affected by an AI system, judges advised by it, and policymakers considering whether or not to deploy it, should have access to its reasoning: what input data an AI system uses, how this data is collected and curated, and how the AI processes it, including what weights it puts on different factors.

One can go further and ask for disclosure of the process behind an AI system’s design. This would include where its training data came from, whether this data was sufficiently representative of the population including demographic subgroups, and what training methods were used to find patterns in this data and develop a predictive system. In other words, we might want an explanation not just of what factors and weights the AI system uses, but why it uses those factors and how those weights were determined.

Beyond simply granting access to an AI system’s reasoning, attorneys for the defense and prosecution, as well as judges, need to possess a fundamental understanding of how these systems operate. This knowledge is crucial because it enables attorneys to identify potential biases or errors in the AI's decision-making process and formulate cogent arguments for or against its findings. Similarly, judges can’t hope to fairly and

accurately weigh these results if they don't understand the mechanisms which allow AI systems to generate results. We recommend establishing a repository of educational resources which attorneys and judges can access to learn what data AI systems are being trained on, how this data is being collected, and how these systems arrive at their recommendations.

Some argue that requiring transparency means sacrificing accuracy. And in domains where sophisticated AI systems are much more accurate than simpler ones, they might be worth using even if they are opaque. But where crime is concerned, opaque AI systems are not much more accurate than their transparent competitors. Predicting crime is hard, and the data is very noisy. For pretrial risk assessment, the Public Safety Assessment (PSA) is just as accurate as COMPAS, even though the former is a simple, transparent algorithm and the latter has proprietary weights.[2] Using fancier methods such as neural networks and random forests gives at best a modest improvement. In the criminal justice domain, it makes far more sense to use transparent AI systems than to struggle to explain the decisions of opaque ones.[3]

Bias in AI systems can be detected and corrected, but only if they are sufficiently transparent — and transparent AI systems can make decision-making more accountable

Many have portrayed AI systems as tools of oppression,[4] and indeed if they are used uncritically they can reproduce biases by being trained on biased data,[5] or by being used inappropriately in ways that allow human biases to re-enter decision-making processes. But AI systems can also be audited for accuracy and bias in ways that human judges cannot.[6] Human judges are obliged to explain their reasoning, but we humans are often opaque even to ourselves. We make decisions too quickly, based on stereotypes and implicit or explicit biases, and then construct justifications after the fact.

At their best, AI systems can provide a new kind of accountability, with decisions and recommendations that can be fully explained. They can be independently audited for

[2] A common definition of accuracy is the AUC or "area under the curve." This is the probability that an algorithm will correctly give a higher risk score to a random defendant who will be rearrested than to a random defendant who will not be. Both the PSA and COMPAS have AUCs ranging from 0.65 to 0.7 on various datasets. In other words, they rank such a pair of defendants in the right order about two-thirds of the time; one-third of the time they would incorrectly rank the safer defendant as higher risk.
[3] Cynthia Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature Machine Intelligence 1, 206–215 (2019).
[4] Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code.* Wiley (2019).
[5] Sandra G. Mayson. *Bias In, Bias Out.* 128 Yale Law Journal 2218–2300 (2019).
[6] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the Age of Algorithms. Journal of Legal Analysis, Volume 10, 2018, Pages 113–174 (2019).

accuracy and fairness, and their biases can be detected and corrected. But all this is only possible if they are transparent.

For instance, an AI system used in the healthcare industry was found to be biased against Black patients. A study revealed why: it had found a correlation between severity of illness and the amount of money the patient spent on healthcare, and thus concluded that people with lower expenditures were less sick. This correlation holds, of course, only for those with access to health care and the resources to spend money on it, thus creating a bias against low-income people and those in underserved communities. This AI system was subsequently modified to remove this bias. But it could be diagnosed and fixed only because it was transparent – independent researchers could see inside it and observe its internal logic.[7]

AI systems should be viewed as tools, not decision-makers

When relying on AI systems to make judicial recommendations, such as in sentencing or probation decisions, judges should always view these systems' results through a critical lens. AI systems are often subject to biases and operate based on limited data. An AI system may take into account prior convictions and prior failures to appear when setting a bond amount, but it may not take into account all of the other relevant information presented by prosecution and defense lawyers in an organic format.

These systems are always operating on partial information, and therefore should never replace humans as decision-makers. When a judge relies on an AI system to help make a judicial determination, they should be made aware of exactly what kinds of data the system was trained on so that they can weigh the model's determination against other information relevant to the case. The defense and prosecution should be made aware of this fact as well so that they can craft arguments that highlight information not considered by a given AI model.

AI systems should produce quantitative predictions that distinguish severe from less severe crimes, rather than vague labels like "high risk" or "low risk"

In addition to ameliorating explicit or implicit biases in individual decisions, AI systems — and the social science and data science behind them — can help advance policy discussions. By bringing facts from social science to the surface in judicial settings, AI can help advance the debate around criminal justice reform.

[7] Sendhil Mullainathan, *Biased Algorithms Are Easier to Fix Than Biased People.* New York Times, Dec. 9 2019.

To do this, however, risk assessment systems should not lump all types of levels of crime together.[8] Unfortunately, while they typically distinguish violent from non-violent charges, the pretrial risk assessments in use today do not make a distinction between felony and misdemeanor arrests, even though many commentators — ranging from legal scholars to the National Academies — have pointed out the need to do so.[9] Of course, a judge might feel that the risk of a misdemeanor is enough to justify pretrial detention. But if a risk assessment lumps charges of all severities together, it does not help the judge consider the "nature and seriousness of the danger to any person or the community that would be posed by the person's release" (Bail Reform Act of 1984).

Worse, many risk assessments provide the judge with an abstract label, like "5 out of 6" or "high risk" or even a color code ranging from green to red, as opposed to a quantitative estimate of the rate of rearrest. This points out the need for another kind of transparency: judges should know what an AI system's output actually means. If that output consists of the abstract phrase "high risk," our human conceptions of what that means are free to run wild, and judges or parole boards who release someone with this label are sure to be attacked in the public square: Psychologists have found that human decision-makers often overestimate the probability of bad events: mock jurors given categorical labels like "high risk" greatly overestimate the corresponding probabilities.[10]

To be useful and accountable, an AI system should make predictions that are as specific and quantitative as possible: it should say how much risk, and risk of what. This would be both more informative to decision-makers, and more amenable to studies of whether its predictions are accurate. We believe more research is needed regarding effective ways of communicating such recommendations to end users.

Finally, an AI system should provide information about its uncertainty and the confidence level of its predictions. For instance, if a defendant has an unusual record, with very few similar defendants in an AI's training data, then the AI system is perforce extrapolating to this defendant from past defendants with quite different records. In a

---

[8] Cristopher Moore, Elise Ferguson, and Paul Guerin. *Pretrial Risk Assessment on the Ground: Algorithms, Judgments, Meaning, and Policy.* MIT Case Studies in Social and Ethical Responsibilities of Computing, Summer 2023.
https://mit-serc.pubpub.org/pub/czviu6qc/release/2?readingCollection=e057132a

[9] See e.g. Timothy R. Schnacke (2017). *"Model" Bail Laws: Re-Drawing the Line Between Pretrial Release and Detention. Center for Legal and Evidence-Based Practices,*
http://www.clebp.org/images/04-18-2017_Model_Bail_Laws_CLEPB_.pdf , Christopher Slobogin *Just Algorithms*. Cambridge University Press (2021), and National Academies of Sciences, Engineering, and Medicine (2022). *The Limits of Recidivism: Measuring Success After Prison.* Washington, DC: The National Academies Press. https://doi.org/10.17226/26459.

[10] Daniel A. Krauss, Gabriel I. Cook, and Lukas Klapatch (2018). *Risk assessment communication difficulties: An empirical examination of the effects of categorical versus probabilistic risk communication in sexually violent predator decisions.* Behavioral Sciences and the Law 36(5) 532–553.

case like this, the AI system should report that its output is less certain. Developing principled techniques for calculating these confidence levels is an active area of research, as is developing effective ways to communicate uncertainty to human decision-makers.

Risk assessment systems should be regularly audited for accuracy and fairness using local data in each jurisdiction they are deployed, whenever there is sufficient data to do so

An AI system trained on data from one jurisdiction, or nationwide data, may not work as well as expected in a new jurisdiction, and the same risk score might correspond to different levels of risk to the public. The rearrest rates corresponding to a given PSA risk score differ between jurisdictions by factors of two or three: a released defendant with a score of 6 is twice as likely to be rearrested in San Francisco as in Kentucky or Kansas.[11]

More generally, an AI system that is accurate and fair in one jurisdiction might not be in another. Due to differences in demographics, policing policies, and many other factors, an AI system might display racial bias in one state or city even though it is unbiased in another. The performance of AI systems can also change over time: for instance, new diversion programs, if successful, can reduce the level of risk some defendants pose. In that case, using a risk assessment trained on data before those programs were implemented might overestimate risk.[12]

For this reason, it is vital to perform local audits of risk assessments in each jurisdiction they are used, and do so periodically, whenever there is enough data to obtain good statistics on their accuracy and various measures of bias. This has been recognized by professional associations of pretrial services agencies[13]. California state law currently requires that pretrial risk assessment algorithms be audited in each jurisdiction every three years [cite SB36], and some other state legislatures are considering similar requirements. The National Institute of Justice should encourage states and cities to reproduce this policy and to extend it to other types of assessments used in the justice system including sentencing, parole, and prison classification.

Establishing best practices for these audits is a matter for research and ongoing discussion in the computer science community. They should not be limited to overall

[11] Moore, Ferguson, and Guerin, ibid.
[12] John Logan Koepke and David G. Robinson (2018). *Danger Ahead: Risk Assessment and the Future of Bail Reform.* 93 Washington Law Review.
[13] NAPSA, National Association of Pretrial Services Agencies (2020). *Standards on Pretrial Release: Revised 2020*, https://napsa.memberclicks.net/standards

measures of accuracy. They should measure accuracy within each demographic subgroup, and various measures of disparity between demographic subgroups.

Even post-conviction decisions such as prison classification and parole should be transparent

Prison classification, i.e., the level of supervision a prisoner is under, should be transparent, as should any changes to their classification. If someone is punished and sent up to a higher-security facility, or to solitary confinement, they should know what misconduct led to this punishment, and be able to contest whether it occurred.

As in the pretrial context, transparent AI systems can also advance policy discussions. For instance, in the process of auditing a prison classification algorithm for accuracy in predicting inmate misconduct, New Mexico researchers found that infractions that took place more than two years ago had little or no relevance to an inmate's future behavior.[14] This led to a policy recommendation to the state's Department of Corrections: namely, that these older infractions should not affect an inmate's classification level, since they have little predictive value about their future behavior.

Artificial Intelligence can have beneficial impacts on the criminal justice system if we carefully outline and standardize use cases, procedures for incorporating AI in decision-making, and methods for auditing these systems

AI systems, like most emerging technologies, when being incorporated into new domains, have the potential to introduce new risks, some of which we can foresee and some which we may not yet anticipate. However, AI, when used responsibly, can have beneficial effects on the criminal justice system.

AI systems can analyze criminal records, evidence, and legal documents much faster than humans and can identify patterns and connections that may otherwise be missed, since humans are not capable of comparing hundreds to thousands of documents at the same time. Using AI to analyze large volumes of legal documents and data frees up law enforcement officers and attorneys to focus on more strategic and engaging work. AI can be used to detect unusual financial transactions that may indicate illicit activity, or analyze images of hotel rooms to help detect instances of sex trafficking[15].

---

[14] Severson, A., Guerin, P., Sanchez, R., & Moore, C. 2024. New Mexico Corrections Department (NMCD) External Classification Validation Study. Center for Applied Research and Analysis, University of New Mexico. https://isr.unm.edu/reports/2024/new-mexico-corrections-department-nmcd-external-classification-validation-study.pdf

[15] "TraffickCam | About." *Traffickcam.com*, traffickcam.com/about.

AI should never replace humans when it comes to making judicial arguments and decisions. However, establishing clear procedures surrounding the use of transparent AI systems can ultimately lead to a more efficient justice system.