PNAS PERSPECTIVE OPEN ACCESS

Interpretable algorithmic forensics

Brandon L. Garrett[a,b,1] and Cynthia Rudin[c,d,e,f,g]

Edited by Thomas Albright, Salk Institute for Biological Studies, La Jolla, CA; received March 1, 2023; accepted May 22, 2023

One of the most troubling trends in criminal investigations is the growing use of "black box" technology, in which law enforcement rely on artificial intelligence (AI) models or algorithms that are either too complex for people to understand or they simply conceal how it functions. In criminal cases, black box systems have proliferated in forensic areas such as DNA mixture interpretation, facial recognition, and recidivism risk assessments. The champions and critics of AI argue, mistakenly, that we face a catch 22: While black box AI is not understandable by people, they assume that it produces more accurate forensic evidence. In this Article, we question this assertion, which has so powerfully affected judges, policymakers, and academics. We describe a mature body of computer science research showing how "glass box" AI—designed to be interpretable—can be more accurate than black box alternatives. Indeed, black box AI performs predictably *worse* in settings like the criminal system. Debunking the black box performance myth has implications for forensic evidence, constitutional criminal procedure rights, and legislative policy. Absent some compelling—or even credible—government interest in keeping AI as a black box, and given the constitutional rights and public safety interests at stake, we argue that a substantial burden rests on the government to justify black box AI in criminal cases. We conclude by calling for judicial rulings and legislation to safeguard a right to interpretable forensic AI.

AI | algorithms | interpretability | explainability | glass box

The rapid growth in the use of artificial intelligence (AI) and algorithmic decision-making, now a "constant presence" in our daily lives (1), has far outpaced our legal system's ability to regulate the technology and ensure that our rights are protected (2). This global challenge has been deepened by the pervasive use of "black box" systems designed to be noninterpretable, meaning that its processes cannot be fully understood by laypeople or by experts (3). Sometimes, as with AI systems developed through machine learning, the system is designed not to be understandable. In other settings, human-designed algorithms are instead treated as a black box, because the government or corporations simply refuse to disclose how it works. In the criminal justice system, as we have detailed in forthcoming work, black box technology used to examine forensic evidence poses heightened risks to both public safety and to fundamental human and constitutional rights (4). Already, criminal defendants have litigated challenges, with limited success, to the use of black box systems to analyze complex DNA mixtures, risk assessments used in pretrial decision-making and sentencing, and facial recognition systems used to identify suspects (5). Yet, "[o]ne of the major obstacles to challenging potential civil rights abuses via algorithm is the opacity of such black box technology" (6).

In one telling example, a federal judge took the unusual step of ordering that the Office of the Chief Medical Examiner in New York City disclose the source code for its probabilistic genotyping software, used to analyze mixtures of DNA (7). As a result, a series of concerns regarding accuracy came to light, and the software was eventually discontinued (7). In a subsequent 2019 ruling, a state trial judge found that it was an error to rely on such forensic evidence and suggested that any convictions that resulted from use of the software should be reviewed. The judge emphasized that the software was a "black box'", which no independent expert was provided an opportunity to examine. This was particularly concerning, the judge noted, where "estimates as to the likelihood of an incorrect conclusion where there actually are four or more contributors [to the DNA sample] run to over 50%" (8).

However, many other judges have instead assumed that black box use of algorithms or AI has real value in criminal cases. Thus, in the same area of forensic DNA mixture interpretation, a Pennsylvania appellate court rejected a defense challenge, denying the request for review by independent scientists of the underlying "proprietary" software (9). The court emphasized "it would not be possible to market" the software "if it were available for free." (9). Other courts, like the New York Court of Appeals tolerate similar proprietary use of forensics in criminal cases by concluding it is reliable, based on studies done by the corporate provider (who has monetary incentives to produce favorable results), and placing the burden on the defense to show a "particularized" need for access (10). Developing a market for a product that serves the public interest could be a laudable goal. However, such rulings far too readily assume that such black box systems have been demonstrated to be accurate and that there is some substantial justification for maintaining its operation as a secret. As we describe, there are strong reasons to question that assumption.

Author affiliations: [a]School of Law, Duke University School of Law, Durham, NC 27708; [b]Wilson Center for Science and Justice, Durham, NC 27708; [c]Department of Computer Science, Trinity College of Arts in Sciences, Duke University, Durham, NC 27708; [d]Department of Electrical and Computer Engineering, Pratt School of Engineering, Duke University, Durham, NC 27708; [e]Department of Statistical Science, Trinity College of Arts in Sciences, Duke University, Durham, NC 27708; [f]Department of Mathematics, Trinity College of Arts in Sciences, Duke University, Durham, NC 27708; and [g]Department of Biostatistics and Bioinformatics, Trinity College of Arts in Sciences, Duke University, Durham, NC 27708

Author contributions: B.L.G. and C.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: bgarrett@law.duke.edu.

Published October 2, 2023.

Interpretable Versus Explainable AI

We focus in this article on the distinction between interpretability and explainability. This distinction has not been made clear by many in the computer science and legal communities. As a result, many have meant quite different things by terms like "open," or "transparent," or "interpretable," or "explainable" AI. Unfortunately, despite their "open" branding, many uses of AI still lack what we call interpretability. While we do not mean to pedantically suggest that terminology is all that matters, or that all need to agree on the same word choices, we believe that it is crucial to be precise about the definitions of certain key AI concepts. Whether the same terminology is chosen, both the theoretical computer science and the legal communities need to be more consistent in the use of these concepts.

Artificial Intelligence. First, to define relatively straightforward terms, "artificial intelligence" simply refers to machines that perform tasks that are typically performed by humans and that normally require human intelligence. "Machine learning" is a subfield of AI, and it heavily overlaps with predictive statistics. Machine learning is a kind of pattern-mining, where data are supplied to the machine, which relies on past patterns to develop methods for making recommendations for what to do next. "Deep learning" refers to neural networks, a specific type of machine learning model, which uses compositions of functions (i.e., a function of a function of a function, etc.). This makes its calculations particularly difficult for a human to understand, but also gives the models powerful predictive capacity for images, text, or time-series. An algorithm need not be created by machine learning, however. Many algorithms can and have been created by humans, and do not involve statistical models learning from data.

Interpretability. By "interpretable" AI, we refer to predictive models whose calculations are inherently capable of being understood by people. In contrast, by "explainable" we refer to efforts to provide post hoc explanations for models, which could be black box models. This distinction between interpretable and explainable is extremely important, we argue.

An interpretable AI system is a glass box system. A person can see how the AI system works and what information it relies upon in a particular instance. The predictive model is disclosed to the users. The system is a "glass box" and not a "black box." It provides information regarding the model, the factors used to provide a result, and how those factors were in fact combined to provide a result.

The underlying models, or algorithms, used by the AI may be extremely complex. However, the factors that the model ultimately relies upon may be quite simple and understandable. For cases involving complex "raw" data–like images, the algorithm can show its work in readily understandable ways. For instance, there are interpretable neural networks that show their calculations by highlighting not only what pixels they used, but how they compared the relevant parts of a current image to the relevant parts of training images in order to make their prediction. Or, some simple risk assessment instruments are depicted in a simple one- or two-page worksheet that assigns points based on certain factors, like the person's age, prior offenses, and current offense. A social worker or judge can easily see how much weight each factor has and why a person is deemed high or low risk, even if they may not understand how the data were used to generate the scheme or how accurate it is.

Transparency. Model "transparency" is different than interpretability: If the formula for the model is not shared, it is not transparent. It is possible for a model to be interpretable but not transparent, in the sense that the reasoning process behind an individual prediction is shared, but one cannot validate the model on a test set because one does not have access to the full model. It is also possible for a model to be transparent but not interpretable, which is the case for most public models, whose formulas are too complicated to understand.

Explainability. The less desirable type of approach, "explainable AI," develops black box models from data (either nontransparent or too complex to be understood by humans), and then queries the black box, in effect, to provide an account of what the algorithm may have done. These explanations do not open the black box. Instead, the researcher, without relying on an understanding of the model, relies on its inputs and outputs to generate explanations for what the key factors were, in general, or in a particular instance. In effect, this approach uses proxies to explain what the AI may have done.

Explainable AI might be better than shrouding the AI in complete secrecy, if the only alternative were a black box. Perhaps that is why many in the AI community have emphasized that explainable AI (or XAI) is a comparatively good thing. We do agree that with explainable AI, the user might better understand what might have been done than if no explanation was provided at all.

There are serious problems with explainable AI. Explanations are not always faithful to the model's calculations. In other words, explanations can often be wrong. Many explainable methods disagree with each other, illustrating why we cannot trust them. We have no way of knowing which one(s) are correct (if any actually are). Explanations also tend to be wrong on more difficult decisions (cases close to the decision boundary), which are precisely the cases where we need explanations to be correct. Explanations may not even be needed on easier decisions, because the decision may be obvious anyway. Explanations (even wrong ones) also may lend more authority to the black box, justifying its use in the first place.

We view post hoc explanations as misleading and inappropriate in high-stakes settings, like in criminal cases. That is why we view the distinction between interpretable and explainable models as an important one. As we discuss, however, there is now a powerful case that interpretable AI is superior for key forensic tasks.

The Black Box Performance Myth. We write to counter the widely held myth that the use of black box systems, despite the risk to constitutional rights, is a necessary evil, because they have an inherent performance advantage over simpler or open systems (11). In academic and policy debates, both champions and critics of black box AI argue that we face a catch-22: They assume that while black box systems are not interpretable, they achieve far greater predictive accuracy.

As one scholar put it: "making an algorithm explainable may result in a decrease in its accuracy." (12). Such claims are often repeated in the computer science, policy, and law literatures (13, 14). The proponents claim these systems represent innovation and higher performance, and therefore private markets in the creation of such black box technologies should be supported—even if they eviscerate the constitutional rights of criminal defendants. Thus, some argue that "instead of worrying about the black box, we should focus on the opportunity," that AI technology may provide (15).

Some of the most trenchant critics of black box AI similarly emphasize how AI derives its efficiency and effectiveness from its "inherently uninterpretable" associations and processes (16). One called it as difficult to understand black box AI as it is to "understand the networks inside" the human brain (17). Or, another stated that since "it may not be possible to truly understand how a trained AI program is arriving at its decisions or predictions," we are faced with a decision whether to embrace or reject the black box (18). Thus the claim that we face such a trade-off lies at the heart of efforts to both critique and retain black box and often private control over AI.

This false dilemma appears to leave society in a bind. There is a need to improve on biased and fallible human decision-making, which has contributed to record levels of incarceration in the United States (19). We cannot run database searches or regressions in our heads when making important decisions, and we can fall prey to biases.

Yet, one cannot even assess whether AI provides real benefits without knowing how the AI works, how well it works, and how it is used in practice. Not only are the benefits of black box AI unclear but the black box obscures the costs. Black box AI can magnify racial biases in existing systems, such as criminal justice (20), and early uses of AI in criminal justice have realized many critics' worst fears regarding errors, racial bias, punitiveness, nontransparency, and privacy invasions (21). Yet, AI secrecy in the criminal system is far from necessary or inevitable—it is an avoidable and poor policy choice. In the criminal system, both fairness and public safety benefit from glass box AI—and therefore, judges and lawmakers should firmly recognize a right to glass box AI in criminal cases.

In machine learning, as we will discuss, there are two regimes: one for standard tabular data (that comes in tables, e.g., demographic or criminal history data), and complex "raw" data (images, soundwaves, etc.). For tabular data, very small interpretable models can often perform as well as black box models on benchmark datasets if they are optimized carefully. For image data, where neural networks are the only technique that performs well currently, interpretable neural networks perform as well as black box neural networks. Thus, for either tabular data or images, interpretable models are generally as accurate as the best of the black box models when applied to benchmark datasets (22).

Special Cause For Accuracy Concerns in Forensics. AI systems face three basic challenges: 1) the problems of training and input data; 2) validation; and 3) interpretation and transparency. First, a "predictive model" is a formula that takes a new observation (represented by a set of features, such as statistics of the person's criminal history, age, prison misconduct history, and education) and produces a prediction (e.g., there is a 14% chance of re-arrest within 2 y of release). These predictive models—black box or interpretable—are sometimes created by machine learning algorithms, which use a database of past cases (the training data) to create the predictive model in a way that it is accurate for the past cases, and hopefully, predictive of future cases. Second, validation of the model uses a new dataset, called a "test set," which must be separate from the training data used to develop the model. That evaluation procedure should be reproducible by other researchers. Third, predictive models cannot be interpreted or explained to others, if they are "black box" models with formulas too complicated for humans to comprehend or their design is not shared with others. Conversely, predictive models are glass box models, or "interpretable" models, when the formula is understandable by humans.

Each of these three basic challenges poses particular challenges to AI systems seeking to provide forensic evidence in criminal cases. First, regarding data, criminal justice data are often noisy, highly selected and incomplete, and full of errors. As Justice Ruth Bader Ginsburg put it, although databases "form the nervous system of contemporary criminal justice operations," nevertheless, "[t]he risk of error stemming from these databases is not slim." (23). In a highly localized and fragmented system, information on other outcomes, such as arrests, jail detention, sentencing, and incarceration, may be far more lacking.

To provide just one example, "there are no nationally representative data available on the numbers of misdemeanor arrests and convictions, let alone data about pretrial detention rates, bail, or sentencing." (24). Indeed, in 2021, during the transition from the Uniform Crime Reporting (UCR) program to the new National Incident-Based Reporting System (NIBRS), about 40% of law enforcement agencies did not report data to the FBI (25). Further, where most crimes go unreported, and where criminal victimization itself, "is a relatively rare event," what law enforcement does not know is substantial (26). In general, criminal behavior is not only uncommon and hard to detect, but involves hard to predict actions and "noise."

Turning from police data to criminal court data, we observe the same types of data challenges. Outcomes in criminal cases reflect a range of subjective and discretionary decisions by various actors, including pretrial services and other social workers, prosecutors, defense lawyers, judges, and jurors. Postarrest outcomes in court often depend on negotiations between counsel, where most cases are resolved through plea bargaining, which is typically not documented, and many cases are dismissed, while cases proceeding to trial rely on judgments of jurors (27). Basic case information and sentencing data may be highly incomplete as well, and data entry failures can magnify in their consequences when consolidated in larger databases (28).

When an AI system relies on past data to form predictions about a present-moment case or situation, the data in the present case at hand may also be lacking. Something as basic as the wrong address information can lead to an erroneous arrest. In the criminal justice setting: "Errors are evidently pervasive, systematic, frequently related to behaviors and policies of interest, and unlikely to conform to convenient textbook assumptions." (26). Basic typographical errors in inputs to black box recidivism prediction models have led to catastrophic errors deeply affecting people's lives (29). In a black box system, one cannot detect such errors in application to a case. To be sure, an interpretable AI system does not alone ensure that errors in data are corrected. But it makes it possible, at least, for people to examine how the system reached a conclusion and correct the record in a case.

Second, using interpretable or glass box AI, we can far more readily validate the system and detect and correct errors in the system. A predictive model should be evaluated using test data, and that evaluation should be replicable by others, which can (practically) only happen given transparency and interpretability of the model's formula, in addition to sharing the code and formulas themselves with outsiders. In criminal justice settings, those basic types of evaluations are not often required; judges often simply assume validations have occurred without inquiring. Beyond interpretability, it is crucial that underlying code and formulas be shared; failure to share algorithms is antithetical be scientific standards regarding replicability and peer-review. In order to consider anything to be an evidence-based practice, such validations must be conducted, and all the more so if it is a practice designed to influence decisions in high-risk settings like the criminal justice system.

Third, interpretability is particularly important in legal settings, where human users of a system, such as police, lawyers, judges, and jurors, cannot fairly and accurately use what they cannot understand. To be sure, there are degrees to which users will actually understand outputs from AI systems. Even if a system is glass box and interpretable, it is also important that the outputs be more than understandable by humans, but actually understood in practice. The outputs should be explained well to the types of people that must rely on them. There has been more work on how to present quantitative information, including AI outputs, in a way that users appropriately understand and use. However, as a first step, the output must be interpretable; next, we must ensure that they are conveyed well. The utter lack of interpretability may explain why, as discussed further, judges have often not adequately scrutinized black box AI and, more generally, black box algorithms.

The Superiority of Interpretable AI in Forensics. Interpretable models are available and they can be superior to black box AI. Let us start with image data, where neural networks are the only type of AI technique that works well. In a recent study, computer scientists compared an interpretable neural network model for classifying objects with noninterpretable counterparts (30). They found that the interpretable AI system performed at a level of accuracy on par with those black box systems on benchmark computer vision datasets. The system not only explained how it reached its results, but provided visual justifications for it, by showing what features of a bird, for example, led it to conclude that it was a red-bellied woodpecker (30).

Switching to tabular data, the stakes are higher when one turns from bird identification to risk assessments used in the criminal justice system to inform decisions such as whether to detain a person pretrial or reduce their sentence. In the criminal setting, interpretable models are readily available that are small enough to fit on an index card, and research has shown that black box models do not perform any better in criminal law settings than simpler and interpretable models (31).

For example, researchers found that a simple model relying on age, gender, and prior criminal record was just as predictable as the COMPAS algorithm, which is proprietary and could rely on up to 137 inputs (32). This was the entire model and explanation:

> if the person has either >3 prior crimes, or is 18 to 20 y old and male, or is 21 to 23 y old and has two or three prior crimes, they are predicted to be rearrested within 2 y from their evaluation, and otherwise not (32).

This interpretable model was created by a complex machine learning algorithm that looked at many different factors and chose among them, combining them in a specific way to as to yield high accuracy. This exercise was not designed to produce a better risk assessment. Ideally, a risk assessment instrument would be designed not merely to predict, but to prevent harm, and would include changeable, dynamic risk factors so that legal actors can recommend interventions that can reduce risk. However, this exercise highlights how models can be quite simple, easy to understand, and without any need to resort to a black box. Researchers in the risk prediction field have long found that a small number of simple factors are modestly predictive: largely age, gender, and prior criminal activity (33). Some of those factors, however, including criminal history and even age, may be entered incorrectly in official records, and it is also very important, even with a simple model, for users to know what data it is relying on.

There have been recent efforts to understand *why* interpretable models have the accuracy of black box models. One recent theory suggests that when the prediction problem is heavily influenced by randomness (e.g., whether someone will commit a crime within 2 y could depend on any number of circumstances, and is a noisy process), there are many approximately equally predictive models, and in that case, it is likely that at least one of these models is interpretable (34).

Whatever the reasons why black box AI seems to lack a performance advantage in a range of important settings, there are strong reasons to believe it performs far more poorly than glass box alternatives, but also introduces new sources of error when used in practice. We have described how black box systems, whether AI or other types of algorithms, can lead to less accurate decision-making. Such models are harder to troubleshoot, validate, review in individual cases, and use in practice. As one of us has put it simply: "Why Are We Using Black Box Models in AI When We Don't Need To?" (35).

Rights Concerns with Forensic Use of Black Box AI. In criminal cases, judges have often deferentially approved black box systems, both those that rely on AI and those that rely on human-created algorithms, assuming that nondisclosure was justified by their great reliability. They have not conducted a careful analysis informed by law and data science. As scholars have developed in a growing body of work, due process, equal protection, confrontation, discovery, and expert evidence-related rights, each places distinct burdens of justification on the government—and unfortunately, judges have often not insisted on searching review of forensic evidence used in criminal cases (36–38).

First, the Due Process Clause ensures "against conviction except upon proof beyond a reasonable doubt of every fact necessary to constitute the crime with which he is charged." (39). The *Brady v. Maryland* obligation requires that prosecutors disclose to the defense favorable evidence, even in the absence of a request from the accused, including impeachment evidence, and evidence in the possession of other government actors, including the police (40). Recent federal rulings have focused on the obligations to disclose forensic evidence (41). Thus, if prosecutors introduce an expert presenting the results of an AI analysis, the defense should be entitled to discovery, not just regarding the ultimate result of that analysis, but also evidence that could permit the defense to ask questions about the methods and analysis performed, in order to impeach the expert. No such evidence will exist, however, unless it is a glass box AI system, where the analysis is understandable and can be disclosed. The Advisory Committee to the Federal Rules of Criminal Procedure notes Rule 16 is intended to require disclosure of scientific results and tests: "the requirement that the government disclose documents and tangible objects 'material to the preparation of his defense' underscores the importance of disclosure of evidence favorable to the defendant." (42). It is standard practice to disclose underlying documentation of forensic experts in federal cases, although state practices are quite uneven (43). There will be a pressing need to ensure use of glass box AI, because otherwise disclosures are far less readily made in discovery.

Relatedly, glass box AI is needed for defendants to benefit from effective assistance of counsel, protected by the Sixth Amendment and the Due Process Clauses. The Supreme Court has repeatedly emphasized obligations of the defense to adequately challenge forensic evidence: "Criminal cases will arise where the only reasonable and available defense strategy requires consultation with experts or introduction of expert evidence."(44).

Further, the Supreme Court's Sixth Amendment Confrontation Clause rulings have emphasized the defense right to adequately confront adverse witnesses, including forensic witnesses in court (45). The defendant's constitutional right to confront an adverse testimonial witness cannot be vindicated without the ability to interpret and understand AI evidence. If a human witness, called by the prosecution as an expert, had relied on the conclusions of a black box AI system, that person cannot be meaningfully cross-examined. The expert witness cannot explain the conclusions reached by such an AI system. Defense counsel cannot meaningfully defend a person from being confronted by AI or algorithmic result without the ability to ask questions regarding the decisions made by the system.

A glass box system also better safeguards rights under the Equal Protection Clause, which protects against purposeful discrimination of protected groups, including based on race (46). Under the Equal Protection Clause, if strict scrutiny did apply, the government might relatedly claim a "compelling government interest" supporting the use of black box AI (47). Yet, the interest cannot be compelling if there is no well-supported performance advantage to use of black box evidence. If the government is potentially obscuring potential racially disparate impacts or uses of race, with no well-justified benefit, then a judge should carefully inquire into how the system is being used. To be sure, challenges to criminal justice outcomes under the Equal Protection Clause face substantial challenges, including because of the discretion afforded to law enforcement under the Fourth Amendment (48), the discretion afforded to prosecutors as executive actors (49), and following the Supreme Court's ruling in *McCleskey v. Kemp* regarding the use of general statistical evidence to raise equal protection challenges to criminal justice outcomes (50). However, even if the courts are not receptive to equal protection claims, because of the challenges of showing discriminatory intent in particular, the risk of racial discrimination is a powerful reason not to permit black box systems in criminal cases on policy grounds. Racial discrimination is difficult to evaluate, because discrepancies in predictions between racial groups may arise due to underlying differences in the populations, underlying differences in reporting, or differences in the way the algorithm is applied in a specific location, rather than unfairness imposed by the algorithm's formula. Without access to the algorithm, it is substantially harder to pinpoint the cause of discrepancies among racial groups. Fairness and discrimination are much easier to assess when models are interpretable.

Finally, for use of AI or an algorithm by an expert witness, if the system is a black box, the parties cannot readily vet the expert to satisfy the evidentiary burden on the party seeking to introduce an expert. Thus, properly applied, *Daubert* and Rule 702 should provide substantial protections in criminal cases. Further, litigants and judges cannot adequately examine, as Rule 702(d) requires them to do, not only whether a method used by an expert is reliable, but whether it was reliably applied to the facts (51). To be sure, human experts can be a "black box," in that we cannot often fully know how they reached judgments and formed conclusions in particular cases. That is one reason why scientists have highlighted the importance of testing the accuracy of human "black box" experts, just as we would want to validate a device or an AI system.

In the past, judges have deferentially reviewed admissibility of expert evidence in criminal cases, even for highly subjective expert methods, and after the U.S. Supreme Court's *Daubert* ruling and amendments to Federal Rule of Evidence 702 tightened gatekeeping requirements for expert evidence. The National Academy of Sciences (NAS) explained in a landmark 2009 report, that where judges have long failed to adequately scrutinize forensic evidence, scientific safeguards must be put into place by the government (52). The NAS report highlighted how courts routinely found admissible a range of forensic evidence of lacking in reliability, where: "[w]ith the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source." (52). Expert testimony regarding traditional forensic evidence, which has often depended on subjective judgments made using unvalidated or unreliable techniques, has resulted in tragic wrongful convictions and lab scandals (38). To be sure, a recent amendment to Rule 702, which will take effect in December 2023, unless Congress acts, highlights the importance of judicial review of expert reliability, the burden on the party seeking to introduce the expert, and the need to review the opinions that experts reach based on their methods. Even given this rule change, there are good reasons to fear that the same lack of judicial engagement with reliability standards may occur when black box algorithms are used in criminal cases. A criminal defendant, if indigent, may be denied funds to retain an expert to examine methods or technology used by a prosecution expert (53). We have seen judges approvingly refer to studies conducted by the maker of an AI system, and not require any independent validation, or disclosures to the defense. For black box evidence, the defense faces practical barriers to challenging the evidence, and a judiciary often inclined to disregard the burden that falls on the party seeking to introduce an expert to show it is reliable scientific evidence.

Forensic AI in the Courts. As noted, a few judges have begun to raise concerns regarding black box use of algorithms in criminal cases, but unfortunately most have not granted discovery, much less relief, despite the range of constitutional and evidentiary rights implicated and the substantial interests at stake. In a noteworthy exception already noted, a New York trial judge explained, regarding a government program, called the Forensic Statistical Tool (FST) used to interpret complex DNA mixtures, that:

> [T]he fact that FST software is not open to the public, or to defense counsel, is the basis of a more general objection. This court understands the city's desire to control access to computer programming that was developed at great cost. But the FST is, as a result, truly a "black box"—a program that cannot be used by defense experts with theories of the case different from the prosecution's (54).

The FST was not an AI technology designed to be uninterpretable; it was a computer program, with code that the government simply refused to disclose; it was not transparent. Indeed, as noted, once the code was disclosed, experts reviewed it and found serious flaws.

However, many other courts have failed to require the most basic disclosures concerning use of algorithms in criminal cases. Indeed, a range of other courts in New York had admitted the same FST evidence before these errors came to light (55). A Pennsylvania court rejected a defense challenge to expert evidence concerning DNA mixture analysis, in the context of evaluating whether it was a "generally accepted" scientific methodology, but finding it was "proprietary software." (56). A range of courts have admitted DNA mixture software, by asserting it is reliable, or relying on precedent, but not clearly explaining why it should be permitted without validation or disclosure to the defense (57). At best, they have found it sufficient that the software developer claimed to have validated the software. Such systems may not be AI systems designed to be uninterpretable; they may be simpler algorithms that could readily be disclosed. Nevertheless, even in those settings, judges have failed to require disclosure to the defense or outside validation.

The most prominent legal challenge to a black box AI was brought in Wisconsin, where a defendant argued that it violated due process and equal protection rights to base the sentence on a risk assessment instrument, called COMPAS, marketed by a private company (called Northpointe), whose mathematical operations are not disclosed. In *State v. Loomis*, the Wisconsin Supreme Court dismissed these claims, emphasizing "the proprietary nature of COMPAS," and that judges have discretion when they consider the risk instrument (58). Responding to the concerns raised by the defense, the Wisconsin Supreme Court did rule that sentencing judges must be given written warnings, or a "written advisement," about the risk tool, including cautioning judges that it relies on group data (58). Those limitations seemed designed to address the serious due process concerns with the lack of transparency. However, such warnings do not open the black box to allow one to assess the operation or accuracy of the AI as used in the individual person's case.

The COMPAS system is not interpretable: One cannot know how it reached its results based on the data shared with the system, so one cannot check its correctness in an individual case. Nor can outsiders readily assess whether its approach is valid, for which they would need further access to code and formulas. The court did not address the issue of possible noise in the data, such as typographical errors, that cannot be detected if one cannot see what the AI is relying on in a particular case. In fact, quite a lot of effort has been made by scientists to understand how COMPAS depends on important variables like race and age, without a lot of success (59, 60). If the AI had been interpretable, perhaps this litigation would not have been necessary and any constitutional issues could have been avoided.

Thus, in a range of settings, courts have deferred to government claims that black box use of AI is justified, in use at sentencing, or reliance on AI by experts, and in other contexts, such as in pretrial and parole settings. In these settings, the government has often claimed black box AI or algorithms offered something advantageous.

A Glass Box Presumption. The burden on the government to justify "black box" high-stakes uses of forensic evidence should be high, given commitments to defense rights of access, nondiscrimination, and reliability of evidence. Our contribution to this literature is that for each of these analyses, without a strong performance justification, there is little justification for not making algorithms open for inspection, vetting, and explanation. Further, companies lack any clear innovation-interest in concealing the effectiveness and accuracy of products used in criminal settings. Thus, particularly in criminal cases with liberty at stake, there should be a strong legal, evidentiary, and constitutional right to glass box evidence.

To date, no legislative enactments or proposals in the United States have required open or glass box forensic evidence. There is an unfortunate reality that constitutional rights may not be enough to address these issues, where they have been unevenly enforced in criminal cases, given the challenges that largely indigent defendants face in obtaining adequate discovery, and the pressures to plead guilty and waive trial rights. In Europe, a 2016 revision to the European Union's Law Enforcement Directive (LED) restricted the use of AI in criminal cases, although enforcement of that provision in practice has been limited (61). The newly introduced Artificial Intelligence Act in Europe may provide a model for more substantial regulation of AI systems in high stakes settings, including uses by law enforcement and by courts, but once enacted, regulations will need to be set out and enforced at the national level (61). Much work will be done in the years ahead to define and implement these legal rules. We hope that they include clear requirements of interpretability for uses of AI in criminal settings.

Meanwhile, the legislative response to the use of black box AI in criminal cases has only just begun, with technology moving quickly and lawmakers and courts moving slowly. So far, a main focus of the first wave of local and state legislation in the United States has been facial recognition technology. So far 10 states have passed restrictions on certain law enforcement uses of FRT, but these restrictions are not all likely to address the problem we discuss here (62). In perhaps the farthest-reaching legislation, Washington imposed a range of quite detailed conditions and transparency requirements on all government uses of facial recognition (63). Importantly, none of those laws have required glass box use of AI for facial recognition; we are aware of no proposals to do so.

We propose that legislation require that glass box, i.e., interpretable, algorithms be mandatory for most uses by law enforcement agencies in criminal investigations. If the use of algorithms could result in generation of evidence used to investigate and potentially convict a person, it should be interpretable. Further, the system should be validated, based on adequate data. If the underlying model can be safely also open sourced, which it often should be given the simplicity of the data relied upon (particularly for scoring systems), validation and interpretability should be required by statute. For facial recognition, these systems should be validated, but not made public due to privacy and safety concerns for the general public. The National Institute for Standard and Technology (NIST) has conducted some validations of facial recognition systems. The U.S. House of Representatives considered a "Justice in Forensic Algorithms Act", which would ensure that any algorithms used in criminal cases be unrestricted by any claim of proprietary or trade secrets protection, and vetted by NIST (64). The law, which failed in committee, would have provided an important starting place.

This right to glass box AI should impose a strong presumption of interpretability for criminal courtroom uses of AI. This presumption should be used by judges when conducting due process analysis and by policymakers when deciding whether to deploy or regulate AI in a criminal system. It should take substantial evidence to overcome this presumption of interpretability. This is not to say that black box AI is never possible in criminal settings, but that the government should have to show a substantial or compelling state interest to support its use. There may be situations in which the government can offer a compelling justification to protect certain types of AI systems from disclosure, for which this presumption may be overcome. For example, a national security justification might support not making public aspects of an AI model. However, at a minimum, it should be carefully vetted by independent researchers, with appropriate security safeguards. Further, for users such as judges and defense lawyers, a glass box is necessary to safeguard defendant's rights.

It may be the case that companies would have a harder time profiting by selling interpretable AI. However, the government should not incentivize those profits or pay for a black box where an interpretable system can be just as accurate, and safeguard defense rights. Unfortunately, judges have often admitted use of black box software that is not independently validated, and judges have often been satisfied to simply rely on the studies done by the same researchers who developed the AI system. Some of that software, like in the DNA mixture setting, consisted in algorithms that the government or private companies have simply refused to disclose. There may be nothing inherently noninterpretable about those systems, and the rationale for nondisclosure is particularly thin in such settings.

Further, researchers, government agencies, and nonprofits can readily develop glass box AI, and they have increasingly done so. Researchers, for example, developed a screener with simple factors for police to forecast domestic violence (65). Other researchers developed a simple scoring system to address unnecessary use of stop and frisk by the New York City Police Department (66). Pretrial risk assessments commonly involve simple scoring systems, focusing on factors like age and prior convictions (67). Researchers created free, open-source probabilistic genotyping software for interpreting DNA mixtures (68). Simple AI systems can perform better and provide understandable information, without concealing errors inside a black box.

To be sure, technology will continue advance and there may be settings in which noninterpretable AI systems are developed that do have substantial and demonstrated performance advantages. Take the case of ChatGPT and other generative AI programs, which have both captivated and disturbed users, and which can be used to generate text, audio, and images, on any number of topics using a black box AI system. Whether it is accurate or not, may not matter if and when it is used to generate entertaining material on culture, relationships, or even law, although there are a range of other risks of the technology that may justify regulation. If it is used in court, however, to inform a decision-maker, then it would not be enough to show that it provides useful material in general. It would have to be tested, regarding its performance, on a particular task. If the AI's job is to produce information (like a search engine), regardless of whether the AI is a black box, the source and trustworthiness of that information should be directly checked. If a judge consults ChatGPT to decide whether to set bail, or to ask for an image of the crime scene, or to make a sentencing decision, we should be deeply concerned. This would be similarly corrupt (if not worse) than tossing a coin to make such decisions. Judges are permitted to leave their judgements up to coins or algorithms.

Perhaps generative AI or other systems will be shown, based on replicated validations and peer review, to perform better than existing decision-makers for some tasks. While no performance advantage has been shown in criminal settings relying on tabular data or visual identification, a performance advantage may arise in new settings. When and if that occurs, we emphasize that the government must meet a substantial burden in justifying the use of noninterpretable AI in a high-stakes setting like in criminal cases. Further, if this new AI system is not interpretable, then as a fallback, it should at minimum be explainable, with explanations faithful to the underlying black box, so that criminal justice actors can know something about its operation in at least a post hoc manner. In general, though, new legislation and judges applying existing constitutional and legal rights should aim to safeguard a right to interpretable forensic evidence in criminal cases.

Conclusion

Black box forensic evidence has become the norm in far too many important criminal justice settings. Judges, lawmakers, and executive actors have been misled by a black box performance myth. When they scrutinize algorithms, they should place a high burden of justification on those proposing to maintain nontransparent, black box forensic evidence in criminal law settings.

The US Constitution safeguards rights to a fair trial under the Due Process Clauses, Sixth Amendment confrontation rights, as well as against discrimination in violation of the Equal Protection Clause and implementing civil rights acts. Expert evidence rules should ensure that scientific evidence is carefully vetted before a criminal trial. These constitutional and statutory protections have increasingly been tested as black box evidence is deployed used in criminal settings, and the early judicial responses have not been very reassuring. Judges have rarely intervened, often because they have credited claims that proprietary algorithms are needed to generate investment in technology, or because they have assumed it is simply not practically possible to open black box technology. Judges will increasingly face pressing questions whether black box evidence is authorized, justified, and constitutional.

If we are to use AI in criminal cases—and uses of AI are proliferating—glass box (interpretable) AI can far better achieve public safety goals while protecting crucial and constitutionally guaranteed rights. The burden on the government to justify noninterpretable black box AI should be substantial, and perhaps future technologies may satisfy such a burden. For the uses of tabular data and computer vision that have been used in criminal settings, there has been no accuracy advantage to black box approaches. Indeed, some algorithms used in criminal settings have been quite simple, and not disclosed to the defense for reasons of profit or government secrecy alone. Such uses are particularly unjustified and troubling.

Interpretability can expose deeply harmful AI systems deployed in criminal settings, illuminate any benefits AI can provide, and can safeguard constitutional criminal procedure rights. In short, it is time to recognize, in criminal cases, a right to interpretable forensic evidence.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. This work was not supported by external funding. This paper is adapted from an in-depth exploration of these problems.

1. H. B. Dixon, Artificial intelligence: Benefits and unknown risks. *ABA J.* **60**, 1 (2021).
2. R. Calo, D. K. Citron, The automated administrative state: A crisis of legitimacy. *Emory Law J.* **70**, 797, 800–01 (2021).
3. Stanford University, Artificial intelligence and life in 2030: One hundred year study on artificial intelligence 46–47 (Sept. 2016).
4. B. L. Garrett, C. Rudin, The right to a glass box: Rethinking the use of AI in criminal cases. *Cornell Law Rev.* **109** (forthcoming).
5. State v. Pickett, NO. A-4207-19T4 (App. Div. 2021).
6. M. Chen, Defund the police algorithms, The Nation, 25 August 2022.
7. L. Kirchner, New york city moves to create accountability for algorithms. Propublica, 18 December 2017.
8. People v. Thompson, 65 Misc 3d 1206(A) (N.Y. Sup. 2019).
9. Commonwealth v. Foley, 38 A.3d 882 (Pa. Super. Ct. 2012).
10. People v. Wakefield, 192 N.E.3d 1152 (2022), and cert. denied sub nom. Wakefield v. New York, 143 S. Ct. 451 (2022).
11. R. Marcinkevics, J. E. Vogt, Interpretability and explainability: A machine learning zoo mini-tour. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2012.01805 (Accessed 13 June 2023).
12. A. Deeks, The judicial demand for explainable artificial intelligence. *Colum. L. Rev.* **119**, 1829, 1834 (2019).
13. N. Ram, Innovating criminal justice. *Northwest. Univ. Law Rev.* **112**, 659 (2018).
14. T. Z. Zarsky, Transparent predictions. *Univ. Ill. Law Rev.* **2013**, 1503, 1520 (2013).
15. V. Pande, Artificial intelligence's 'black box' is nothing to fear. New York Times, 25 January 2018.
16. A. Rai, Explainable AI: From black box to glass box. *J. Acad. Mark. Sci.* **48**, 137 (2020).
17. D. Castelvecchi, Can we open the black box of AI? *Nature* **538**, 20–23 (5 October 2016).
18. Y. Bathaee, The artificial intelligence black box and the failure of intent and causation. *Harv. J. Law Tech.* **31**, 890, 891 (2018).
19. J. Travis *et al.*, The growth of incarceration in the United States: Exploring causes and consequences 2 (National Research Council, 2014)
20. A. E. R. Prince, D. Schwarcz, Proxy discrimination in the age of artificial intelligence and big data. *Lowa Law Rev.* **105**, 1257 (2020).
21. A. G. Ferguson, Policing predictive policing. *Wash. Univ. Law Rev.* **94**, 1109 (2017).
22. C. Rudin *et al.*, Machine learning: Fundamental principles and 10 grand challenges. *Stat. Surv.* **16**, 1–85 (2022).
23. Herring v. United States, 555 U.S. 135, 155 (2009) (Ginsburg, J., dissenting).
24. P. Heaton, S. Mayson, M. Stevenson, The downstream consequences of misdemeanor pretrial detention. *Stan. Law Rev.* **69**, 711, 732 (2017).
25. W. Le, What can FBI data say about crime in 2021? It's too unreliable to tell (Marshall Project, 2022) (14 June 2022).
26. J. Pepper, C. Petrie, S. Sullivan, "Measurement error in criminal justice data" in *Handbook of Quantitative Criminology*, A. R. Piquero, D. Weisburd, Eds. (Springer, NY, 2010), 10.1007/978-0-387-77650-7_18.
27. B. L. Garrett *et al.*, Open prosecution. *Stan. Law Rev.* **75** (forthcoming 2023), Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3946415.
28. C. Slobogin, Government data mining and the fourth amendment. *Univ. Chicago Law Rev.* **75**, 317, 323–27 (2008).
29. C. Rudin, C. Wang, B. Coker, The age of secrecy and unfairness in recidivism prediction. *Harv. Data Sci. Rev.* **2**, 1 (2020).
30. C. Chen *et al.*, "This looks like that: Deep learning for interpretable image recognition" in *33rd Conference on Neural Information Processing Systems* (NeurIPS, Vancouver, Canada, 2019).
31. C. Wang, B. Han, B. Patel, F. Mohideen, C. Rudin, In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *J. Quant. Criminol.* **39**, 1–63 (2022).
32. E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, Learning certifiably optimal rule lists for categorical data. *J. Mach. Learn. Res.* **18**, 1–78 (2018).
33. J. Monahan, J. Skeem, Risk assessment in criminal sentencing. *Ann. Rev. Clinical Psych.* **12**, 489 (2016).
34. L. Semenova, C. Rudin, R. Parr, "On the existence of simpler machine learning models" in *ACM Conference on Fairness, Accountability, and Transparency* (2022).
35. C. Rudin, J. Radin, Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition *Harv. Data Sci. Rev.* **1**, 1–10 (2019).
36. R. Wexler, Liberty, and trade secrets: Intellectual property in the criminal justice system. *Stan. Law Rev.* **70**, 1343 (2018).
37. A. Roth, Trial by machine. *Georgetown Law J.* **104**, 1245, 1300 (2016).
38. B. L. Garrett, *Autopsy of a Crime Lab* (University of California Press, 2021), pp. 122–138.
39. In re Winship, 397 U.S. 358, 364 (1970).
40. Kyles v. Whitley, 514 U.S. 419, 437 (1995).
41. B. L. Garrett, The constitutional regulation of forensic evidence. *Wash. Lee Law Rev.* **73**, 1147 (2016).
42. Fed. R. Crim. P. 16 advisory committee note.
43. U.S.A.M. 9-5.0003.
44. Hinton v. Alabama, 134 S. Ct. 1081, 1088 (2014).
45. Melendez Dias v. Massachusetts, 557 U.S. 305 (2009).

46. U.S. Const. amend XIV, §1.
47. Palmore v. Sidoti, 466 U.S. 429, 432–33 (1984).
48. Whren v. United States, 517 U.S. 806, 813 (1996).
49. U.S. v. Armstrong, 517 U.S. 456, 469 (1996).
50. McCleskey v. Kemp, 481 U.S. 279, 297, 312 (1987).
51. Fed. R. Evid. 702(d).
52. Committee on Identifying the Needs of the Forensic Science Community & National Research Council, Council, strengthening forensic science in the United States: A path forward (2009), p. 87.
53. P. C. Giannelli, S. Antonucci, Forensic experts and ineffective assistance of counsel. *Crim. L. Bulletin* **48**, 8 (2012).
54. People v. Collins, 15 N.Y.S.3d 564, 580 (N.Y. Sup. Ct. 2015).
55. People v. Lopez, 23 N.Y.S.3d 820, 825 (N.Y. Sup. Ct. 2015).
56. Commonwealth v. Foley, 38 A.3d 882 (Pa. Super. Ct. 2012).
57. U.S. v. Russell, No. CR-14-2563 MCA, 2018 WL 7286831, at *8 (D.N.M. Jan. 10, 2018).
58. State v. Loomis, 881 N.W.2d 749, 763-65 (Wis. 2016).
59. J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).
60. E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, "Learning certifiably optimal rule lists" in *2017 Proceedings of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining* (ACM, 2017), p. 35.
61. Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing council framework decision 2008/977/JHA (Law Enforcement Directive = LED), OJ 2016 L 119/89 (2016).
62. Regulation of the European Parliament and of the Council, *Laying Down* harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (April 2021), Available at https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.
63. RCW § 43.386.070.
64. See Reps. Takano and Evans reintroduce the justice in forensic algorithms act to protect defendants' due process rights in the criminal justice system (2021), https://takano.house.gov/newsroom/press-releases/reps-takano-and-evans-reintroduce-the-justice-in-forensic-algorithms-act-to-protect-defendants-due-process-rights-in-the-criminal-justice-system.
65. R. A. Berk, Y. He, S. B. Sorenson, Developing a practical forecasting screener for domestic violence incidents. *Eval. Rev.* **29**, 358 (2005).
66. S. Goel, J. M. Rao, R. Shroff, Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Ann. Appl. Stat.* **10**, 365–394 (2016).
67. S. L. Desmarais, S. A. Zottola, S. E. Duhart Clarke, E. M. Lowder, Predictive validity of pretrial risk assessments: A systematic review of the literature. *J. Behav.* **48**, 398 (2021).
68. *See* LRMix Studio, at http://lrmixstudio.org/.