Position: AI's growing due process problem

Sunayana Rane [1 2]

Abstract

AI systems are now ubiquitous in real-world decision-making. However, their use is often invisible and almost always difficult to understand for the ordinary people who now come into contact with AI regularly. As these AI-driven decision-making systems increasingly replace human counterparts, our ability to understand the reasons behind a decision, and to contest that decision fairly, is quickly being eroded. In the United States legal system due process includes the right to understand the reasons for certain major decisions and the right to openly contest those decisions. Everyone is entitled to due process under the law, and human decision-makers have been required to adhere to due process when making many important decisions that are now slowly being relegated to AI systems. Using two recent court decisions as a foundation, this paper takes the position that AI in its current form cannot guarantee due process, and therefore cannot and (should not) be used to make decisions that should be subject to due process. The supporting legal analysis investigates how the current lack of technical answers about the interpretability and causality of AI decisions, coupled with extreme trade secret protections severely limiting any exercise of the small amount of technical knowledge we do have, serve as a fatal anti-due-process combination. Throughout the analysis, this paper explains why technical researchers' involvement is vital to informing the legal process and restoring due process protections.

1. Introduction to Due Process

Due process is recognized as an important constitutional right in the United States, formally described in the 5th and 14th amendments to the Constitution. The 5th amendment says no one shall be "deprived of life, liberty or property without due process of law." The Due Process Clause of the 14th amendment, adopted during post-Civil-War fears that confederate states may find ways to deny newly-recognized rights for African Americans, uses the the same words to assert that states may not deprive any individual of these rights. Due process does not apply to all decisions, but it does apply to many important government actions.

Due process as interpreted in the American legal system, even as far back 1868 when the 14th amendment was ratified, had come to mean "a certain core procedural fairness" that people could expect of government actions in their lives. In particular, due process involved the right to "notice," "the opportunity to be heard," and "a determination by a neutral decisionmaker according to some fair and settled course of judicial proceeding" (Eberle, 1987). Due process is deeply intertwined with the ideal of fundamental fairness of the legal system. Equal protection and due process have together won substantial victories for civil rights (Brown v. Board of Education, 1954), abortion rights (Roe v. Wade, 1973), LGBTQ rights (Lawrence v. Texas, 2003; Obergefell v. Hodges, 2015), and disability rights (Youngberg v. Romeo, 1982).

This type of fundamental tension between fairness and powerful interests is not new. However, the use of AI systems to skirt due process requirements *is* new and unusually dangerous due to their invisibility and staggering spread (Citron & Pasquale, 2014). Ordinary people are up against technically opaque systems they cannot fight, nor understand, in areas ranging from state disability aid to healthcare, from getting a mortgage to staying out of prison. (The Guardian, 2021; ProPublica, 2016). The IRS contracts out the creation of AI systems that produce taxpayer profiles from social media data, which they then use to choose targets for tax audits (Houser & Sanders, 2016). The same contractors that market these tools to federal agencies also often sell tools to private entities, such as credit scoring companies (Hurley & Adebayo, 2016). While some legal scholars sounded the due process alarm over a decade ago on automated scoring tools replacing human decision-making (Citron, 2007), **this paper takes the position that today's more sophisticated and largely uninterpretable AI systems present a new kind of danger: AI systems whose behavior cannot be**

[1] Department of Computer Science, Princeton University [2] University of Chicago Law School. Correspondence to: Sunayana Rane <srane@princeton.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

explained even by their creators are not only a threat to due process, they are fundamentally incompatible with it. Due process protections provide one important way in which we can start to unravel deceptive business practices and secretive uses of AI systems, because government entities are subject to oversight and reporting requirements that private parties are usually exempt from.

2. An Adversarial System

The legal system is at its core an inherently adversarial system. Although the constitutional due process provisions guarantee us certain rights, the legal system is practical in its understanding that there are those who would violate those rights, if given the opportunity. The adversarial system is designed to allow parties to contest these violations.

Due process provides the right to challenge a capricious, arbitrary decision – in other words, a right to at least get a fair shot to argue against an unfair decision. And yet the problem for AI-driven decision-making becomes immediately clear, as explainable AI (xAI) techniques for sophisticated models are still a work-in-progress: when we don't fully understand our AI systems, how can we explain their decisions? How can those explanations be challenged and overturned when they are capricious, untrustworthy, or unfair? Courts have shown a troubling tendency to think of AI systems as inherently mathematical and therefore "unbiased" and superior to inevitably biased human decision-makers, and have therefore upheld their use in many cases where the consequences are heartbreaking and the process clearly unfair (The Guardian, 2021; ProPublica, 2016).

The adversarial system is undermined when the underlying decision-makers cannot be interrogated and cross-examined as intended. If the AI research community's understanding of AI systems is based on (already imperfect) interpretability methods that are only understandable to those with substantial training in computer science, then lawyers are woefully unprepared to challenge an AI system's decision-making in court. This means that ordinary people are being subject to arbitrary, strange AI-driven decisions that we AI practitioners are all too familiar with, but without our ability to intelligently challenge the AI system's behavior. They are left with little recourse. Challenges to AI-driven decision-making often fail (State v. Loomis, 2016). Alarmingly, these AI systems are often developed and sold by companies who claim they are "proprietary" trade secrets, even when the AI systems are shown to be racially biased (ProPublica, 2016). How can a judge who doesn't understand such an AI system intelligently ajudicate it without more information?

2.1. Notice and Hearing

Courts have interpreted the due process clauses to mean that the government cannot take away certain rights and privileges without proper 'notice and hearing' to those affected. The purpose of the 'notice' is to inform a person of reasons for the impending decision, and the purpose of the 'hearing' is to give them the opportunity to contest that decision (Eberle, 1987). Notice and hearing are key to the idea of fairness in the process.

However, 'notice' given by an AI system is often impossible to understand, with the reasons for the decision obscure to the staff in charge and often even to the creators of the AI in question.

As Goodman (Goodman, 2021) points out, AI systems "do not provide any opportunity for meaningful cross-examination, knowledge of opposing evidence, or the true reasoning behind a decision." Thus the 'hearing' part of the 'notice and hearing' requirement is watered down substantially by simply relegating the heavy-lifting to an AI system, and subsequently using the AI system itself as a 'reason.'

Now that AI systems are increasingly replacing human decision-makers in matters that affect each of these domains and many others, due process is under threat in a very different way – AI systems cannot, in their current forms, guarantee due process in the way a human decision-maker is expected to. Both machines and humans are fallible, but AI models display a range of wildly unpredictable behavior far outside the expected distribution for human decision-makers (The Guardian, 2021; NYTimes, 2021; NPR, 2019). Although important interpretability and explainability research continues (Doshi-Velez & Kim, 2017), we are still far from being able to explain the causality behind most model behavior in a thorough and satisfying way. Even in those simpler models which are inherently more easily interpretable, such as regression-based models and shallow decision trees, causality and its requisite mathematical assumptions are nuanced and easily misunderstood without technical expertise. This problem becomes commensurately more dangerous as models get bigger and more sophisticated: for today's largest models, even those training the models do not understand the causal links underlying exactly *why* a model made a particular decision.

3. What has changed? Modern AI's unique perils

Over a decade ago, a small group of legal scholars raised the alarm about any kind of quantitative scoring tool having due process concerns. Are modern AI systems really any different from the old scoring tools discussed in treatises like Citron (2007)?

The difference between hard-coded scoring tools discussed in earlier discussions of due process and technology and today's more sophisticated AI systems lies with the fundamental fairness question that is so key to the legal fulfillment of due process. Due process guarantees that people at the receiving end of an AI-driven decision have the right to know the "why" behind that decision. In the case of previous technologies, including more rudimentary scoring systems used in applications like credit scoring or social obedience scoring, the issue was that the hard-coded scoring algorithm's interpretable values were not revealed to the public or to key oversight agencies.

With modern AI systems, however, this problem is compounded by an even more fundamental due process incompatibility: in most cases, even the AI's creators, even if they were to invest in state-of-the-art interpretability and explainability techniques and be fully transparent with the public about the results, still cannot fully understand (let alone explain to a layperson) *why* the AI system produced the output/decision that it did. This is not merely a transparency issue that can be remedied through more stringent regulation; it is a true incompatibility between due process and modern AI, which now invisibly permeates our lives at an unprecedented scale and severity.

Due process was intended to protect our right to ask a decision-maker *why* a decision has been made, so that we can contest the validity of the *why* in court. Therein lies the fundamental incompatibility between today's models and due process protections–one that cannot be remedied simply through additional transparency requirements: The *why* remains elusive to the best of us.

Case study 1: Disability rights

AI-driven due process violations are particularly harmful to vulnerable groups who lack the resources to effectively fight back. In one of several documented disability-rights cases (The Guardian, 2021), an AI system that had replaced a social worker decided to cut a cerebral palsy patient's state aid, which he used to pay for the helper he needed for basic functions like using the bathroom.

Few of the people subject to such arbitrary and opaque decisions know, and indeed should be expected to know, how to fight back. In one notable class-action case (KW v. Armstrong, 2016), the Idaho ACLU represented adults with developmental disabilities who relied on state aid to live in their communities and who had their welfare benefits cut by the Idaho Department of Health and Welfare (IDHW). The IDHW used an AI system to make these cuts using data collected by a hired contractor. The details of this tool, they claimed, were proprietary 'trade secrets' and they refused their disclosure to the public (an argument that the court rejected). The Idaho ACLU successfully argued that the IDHW provided insufficient notice and violated due process in cutting aid budgets, because the notice provided "made it very difficult for a participant to determine why his budget had been reduced and left him unable to effectively challenge the reduction" – key components of notice and hearing as guaranteed by due process.

The details of the statistical budget tool are, to a technical audience, concerning at best. Only 733 data points are used to create the model, with over half of the past participant data discarded and key groups underrepresented. The "software program runs a spreadsheet" that calculates dollar amounts based on individual needs. It then auto-generates the 'notice' provided to the individual whose budget has been cut. Whether the spreadsheet component is just a front-end or the entire AI system is run through spreadsheet macros is unclear. The appeals process is lengthy and cannot be navigated by people with developmental disabilities on their own, despite the fact that 39% of these individuals do not have a legal guardian who they can turn to for help in appealing the decision, and many do not live with relatives. By the IDHW's own estimate (for which they did not provide any sound empirical evidence, indicating that the true proportion could be much higher) the tool would give a whopping 15% of individuals an inadequate budget. The IDHW had not conducted the annual recalculations which they admitted were needed to update the tool, and "ha[d] never checked to ensure that the current tool [was] not reducing participant budgets arbitrarily."

Citing an older case, the court in *KW v. Armstrong* reaffirmed that due process protections were intended to "insure fairness and ... avoid the risk of arbitrary decision making" (Carey v. Quern, 1978). Yet it is this very lack of reliable fairness, this prevalence of arbitrary behavior that plagues even our most advanced AI systems. Technical AI researchers know all too well that there is currently no easy solution for it.

The most dangerous part of this trend is the fact that the court proceedings often don't even mention what kind of 'AI system' is being used, and often acquiesce to demands that any information about the AI system remain a trade secret. In a court of machine learning researchers, our first question might be 'Is it logistic regression, a random forest, or a 100B parameter language model?' We would then be able to proceed to some reasonable mitigation strategies based on the type of model in question. A logistic regression would perhaps have to be explained by a feature importance ranking along with a plain English explanation of its meanings. A random forest could have an enforced depth limit, or perhaps a visualization of a single tree-based classifier would be required to show which variables are splitting the data and why. Both could have an enforced minimum accuracy, precision/recall, and other key metrics with required disclosure to the public. In each of these cases, with a plain English definition, perhaps the person in question could understand and contest the AI system's purported reasons for a decision. If an LLM told someone they are no longer entitled to disability benefits however, most AI researchers would soundly reject that claim and entirely disallow the use of a type of AI system that, despite its powerful abilities, often fabricates, hallucinates, and is large and opaque enough to be incredibly poorly suited to any kind of post-hoc interpretability analysis that would even come close to meeting the standards of the due process notice requirement.

The nature of the AI systems used often raises technical questions that are important for the AI research community to consider, because we are uniquely positioned to help elucidate and disentangle these thorny issues. For example, where do we draw the line (or the Venn diagram) between statistical tools, machine learning, deep learning, and AI? Most courts are currently unaware that there is a distinction, and therefore cannot engage intelligently on the question of precisely when due process protections are lost. Another recent case on disability rights contested the use of a statistical score, called the SIS (Supports Intensity Scale) score, which also resulted in reduced aid budgets (LS v. Delia, 2012). There is, once again, disappointingly little information about how individual states use the SIS questionnaire and other information to arrive at a 'score,' but based on information from the American Association on Intellectual and Developmental Disabilities (AAIDS) who have created the SIS assessment and scale, the scale itself seems to be a statistical tool created to help understand how an individual places on a distribution of those with intellectual disabilities (American Association on Intellectual and Developmental Disabilities, 2024). While such statistical tools themselves can be useful, their translation to aid budget decisions is often opaque and harmful. Banning the tools doesn't make sense, but requiring extreme transparency in their use for budget determinations does.

Despite the many technically questionably decisions made in the creation and use of IDHW's AI-driven budget tool, *KW v. Armstrong* is one of the few cases where real progress was made: the AI system and the (lack of) data used to train it were discussed in depth in the opinions released to the public, the efforts to shroud the details in secrecy under 'trade secret' protections were largely overruled in the public interest, and the vulnerable population whose rights were cruelly trampled eventually won the case. Perhaps most importantly, established Supreme Court precedent Goldberg v. Kelly (1970) had decades earlier deemed disability benefits protected property interests, the reduction of which required the kind of informative notice the lawyers in *KW v. Armstrong* could later contest on due process grounds.

However, this outcome is far from the norm. Most of those at the receiving end of AI-driven caprice are unaware of the fact that the capricious decision upending their life was even made by an AI system, and are therefore entirely powerless to contest it. There often aren't precedents that set up protections so nicely (such as by marking welfare benefits as protected property interests, as in (Goldberg v. Kelly, 1970)) as to make it feasible to contest AI-driven encroachments on established, nuanced rights. Which way a case will turn often rests on state laws, with little universal protection for disadvantaged populations. Appeals courts are often split on decisions regarding the use of AI systems. Without technical guarantees, best practices, and common-sense transparency requirements designed by those extremely comfortable with technical details, we will increasingly have AI systems that are cheap and "efficient" but also "secret, biased, underparticipatory, unaccountable, and intrusive on the privacy of low income and vulnerable populations" (Spaulding, 2020).

4. Interpretability for due process

Some legal scholars have argued that judges should demand explanations from AI-driven decision-making systems using explainable AI (xAI) techniques (Deeks, 2019) to "open the black box," but courts have rarely explored any technical details of these AI systems in practice. Furthermore, as we know, xAI and interpretability are still in their formative stages, and currently cannot conclusively and causally explain a large neural network's decisions. Other types of models, like complex tree-based ensembles, can be equally difficult to interpret. Those using AI systems often don't even have to disclose which type they are using (or even that they are using an AI system to make the decision at all).

If interpretability is to help protect due process rights, the tools we use to understand AI systems' decisions must be as universally accessible as the AI systems are ubiquitous. Unfortunately, they are currently far from easy to use for everyone. Current interpretability research is rather inaccessible to the outside world. Increasingly, those making important decisions about how these AI systems can and cannot be used in the world are *not* computer scientists. If the only people able to to understand and evaluate an AI system are those with years of computer science training, then due process violations will be impossible to catch and litigate.

Some machine learning researchers have argued that black-box models (including all deep learning models) should be disallowed from high-impact applications, and that inherently interpretable (usually simpler) models be used instead (Rudin, 2019). However, the performance gains of deep learning models, and in particular large (increasingly uninterpretable) models, have ensured that this advice has not been heeded. When should it be necessary to use only inherently interpretable models, or no models at all? This section engages with how different types of AI systems exhibit varying degrees of interpretability and compatability with due process – in situations requiring due process protections, some methods are more savageable than others.

4.1. Regression-based AI systems

Regression-based predictive AI systems, including logistic regression for classification, are perhaps the easiest to interpret and probe. The easiest way to determine the importance of each input feature, such as race or gender, to a certain decision is to examine the coefficient corresponding to the input feature.

Regression-based models can also be probed and adjusted for highly correlated variables, such as race and socioeconomic status, so that variables that unwittingly become proxies for protected class variables can be spotted and their effects mitigated. However, to do this type of analysis, model details including their inputs and coefficients have to be made available for study and tweaking. When AI systems are deemed proprietary, even the simplest regression coefficients become uninterpretable for lack of access to them by the larger community.

4.2. Neural networks

Simple neural networks are often still explainable to some degree. A regression-based model is, after all, mathematically equivalent to a fully-connected single-layer neural network (Zhang et al., 2023). Here the single layer of weights and biases, mapping directly to each of the inputs, makes it straightforward to understand the relative importance of each feature based on its corresponding coefficient. However, once the networks get deeper, even by just a few layers, it becomes increasingly difficult to make conclusive judgements about why the model made a particular decision.

Even the most technically sound, intuitively sensical approaches to demystifying neural networks can have unintended consequences that are difficult to grasp for those without a computer science background. For example, while saliency maps are important intuitive tools, further research has raised questions about their efficacy in truly capturing the effects of training in CNNs (Adebayo et al., 2018). This is a level of nuance that most judges, juries, and ordinary people will not (and should not be expected to) understand. At least until our interpretability methods are translated to simple, intuitive tools accessible to laypersons, there are areas in which the existence of interpretability tools should not be used as an excuse for a model's use – there are areas in which we should not be using these models at all.

4.3. Large language models (LLMs)

Moving beyond the scale of simpler neural networks and CNNs are the LLMs and foundation models of the past few years. Here the interpretability work becomes more difficult, even for those with considerable technical expertise. Just as it is difficult to map human neurons to particular behaviors and decisions, it has proven quite difficult to map LLM behaviors concretely and causally to vector-level representations (Sucholutsky et al., 2023; Broniatowski et al., 2021). This is an important ongoing research challenge, and we will have better answers in time.

In the meantime, one troubling thing about LLMs used for decision-making is the common misperception among laypersons that the AI system can just 'explain its decisions' using natural language, which makes it different from previous AI-driven decision-making systems – while those were just tools, this is a human-like decision maker. Without technical expertise in just how next-token production works in LLMs, it is tempting to believe that a self-explaining system can fulfill due process requirements of notice and hearing just as a human would.

This is only going to get worse as models become more powerful – the natural tendency is to credit an LLM for behavior it *seems* to be exhibiting. It is difficult to intuitively understand that meanings of the words it is using are different from what we think they are, and the explanations are even more dangerous because they are meaningless and often false while seeming very believable on the surface. As AI safety researchers, we can mitigate the effects of this by, at the very least, ensuring that the shared language we use with LLMs will also have shared *meaning* in the way it does between two human interlocutors (Rane et al., 2024) – that they are aligned not just in proclaimed values, but in fundamental concepts, language, and cognitive ability as well. Humans lie too, but they can also be held accountable for those lies. Humans can usually depend on the shared meaning of the words they use with one another; it is imperative that we get to the stage where we can do the same with AI.

4.4. CART-based AI systems

Often the workhorse of the machine learning toolkit, classification and regression tree-based (CART) methods enjoy widespread use and reliably good performance. Tree-based AI systems continue to be used extensively in practice to replace human decision-makers in narrow tasks that don't require human-like conversational or reasoning ability. For this reason, it is well worth looking into how these AI systems can be translated into useful interpretability information for judges and juries to consider.

Tree-based ensemble methods such as random forests

(Breiman, 2001) are most commonly used in practice, but they also tend to be some of the most difficult to interpret into plain English. Single decision trees, however, can be quite interpretable when they are not too deep – each partition of the data can be understood and, with context, perhaps even explained and challenged. A shallow decision-tree can begin to look a lot like a flowchart that a human decision maker might follow, based on certain variable thresholds, when making a determination. However, the deeper a tree becomes, the more the variable splits start to become difficult to explain post-hoc, and the more uninterpretable it becomes (Molnar, 2022). Unfortunately, as with neural networks, the predictive power of tree-based methods usually increases substantially with greater depth. This tradeoff between easily understood and challenged tree-based AI systems and more powerful (in terms of predictive accuracy) tree-based AI systems is something that should be explored further in real-world settings.

4.5. Statistical tools

Some 'AI' tools used in practice are often straightforward statistical models. However, even a simple thresholding tool based purely on statistics is not inherently interpretable and should be subject to review and inspection by independent third parties. Important questions include: what kind of data was used, what kind of distributional assumptions were made and why, what kind of performance reviews were conducted and on which test data? Was *any* test data held out? Was validation forward-looking or backward-looking in time? All of these questions and answers are difficult for laypersons to understand even in the case of statistical models, and the challenges grow as models become more complex. Subjecting even the most basic statistical models to thorough interpretability requirements is a first step towards getting individuals the information they need to contest poorly-made decisions.

4.6. Interpretability, explainability, and plain English

It is difficult to overstate the importance of interpretability tools that are accessible to everyday people; however, this does not excuse such tools from having technical rigor and from providing technical details when required. It is not an either-or between technical details or simple-English details – we need both. AI alignment research has identified goals of aligning AI with humans at the representational, conceptual, behaviorial, and values levels (Rane et al., 2023; Sucholutsky et al., 2023; Rane et al., 2024). To restore due process protections, we need a better understanding of models at *every single one* of these layers. As AI researchers it is our task to translate this understanding to "plain English" as we acquire it. The world needs these explanations, even if our best explanation is, for now, an acknowledgement that we don't fully understand why these models behave in unexpected ways, and that their behavior will remain unpredictable in the near future.

Case study 2: In the criminal justice system

The area of the law where the stakes are perhaps the highest is the criminal justice system. Unfortunately, this is one place where opaque AI systems continue to be used extensively, despite persistent bias and fairness issues. In one of the most public-facing examples of this, a 2016 ProPublica exposé revealed the extensive use of criminal risk assessment AI systems, used to determine the length of prison sentences, which were biased against black defendants (ProPublica, 2016). While the fairness research community has engaged extensively with algorithmic bias since then, most technical researchers are surprised to learn that the legal world has not disallowed the use of such AI tools; on the contrary, many advocate for their expansion within the criminal justice system.

In a 2018 court case over the COMPAS algorithm, the Wisconsin Supreme Court upheld the use of COMPAS and similar risk-scores in making sentencing decisions as long as the judge also had other reasons for justifying the decision (State v. Loomis, 2016). Unfortunately, unlike the disability rights case discussed earlier, this time the court upheld complete 'trade secret' protections for the 'proprietary' nature of this AI system. The case has been criticized in legal scholarship for failing to protect defendants' due process rights (Freeman, 2016) despite the clear and public display of the AI system's racial bias (one of many potential biases, not all of which are as easily measurable as race or gender bias).

Troublingly, the court ruled that because the risk score was "not determinative" in deciding the final outcome of the case, and because the trial court "would have imposed the exact same sentence without it," that its use did not constitute a due process violation. This assertion is based in existing legal understandings of whether something can be "probative" without being "determinative" in the outcome of the case. Yet upon closer inspection with a lens of simple common sense, it is a strange statement that seems to disregard both human cognitive biases and the extremely harmful racially-biased nature of the AI outputs – the court implies that it is okay to use the harmful AI system as a *factor* in making a decision because it didn't *significantly* impact the final decision. If the AI system doesn't have a significant impact, and it is clearly racially biased, then why allow such a demonstrably harmful tool to be used at all? If it is valuable to use, then clearly it *does* have a significant impact on the final decision, and therefore that impact should be carefully and openly scrutinized. It must be one or the other.

The proprietary nature of the AI system made it nearly impossible for the judges to understand any details about its technical nature. Even if they had the technical expertise to know which questions to ask (which is a lot of ask of someone without computer science training of any kind), the trade secret protection allowed Northpointe, the company selling COMPAS, to skirt the relevant questions. In her concurring opinion, Justice Abrahamson wrote "this court's lack of understanding of COMPAS was a significant problem in the instant case. At oral argument, the court repeatedly questioned both the State's and defendant's counsel about how COMPAS works. Few answers were available." She then recounted that Northpointe tried to submit an amicus brief regarding the accuracy and efficacy of the COMPAS AI system, which the court denied. Presumably the court was concerned about the conflict of interest Northpointe had in introducing this information, along with the impossibility of independently verifying quantitative metrics due to the aforementioned trade secret protections. However, Justice Abrahamson's view was that the court *should* have allowed Northpointe to file the amicus brief because "[t]he court needed all the help it could get."

This sentiment sums up two problems: First, courts often *do* need all the help they can get in understanding the technical details of an AI system. Without AI researchers involved in the process, they will be left without it. Second, when trade secret protections for AI systems are upheld, courts have no information about the AI system whatsoever and may choose to trust, as their only way of understanding the AI system, the information voluntarily provided by the company creating the AI system – information that will inevitably be in the company's own interests. This is absolutely unacceptable. Without independent analysis of AI systems like COMPAS, including basic audit and transparency requirements, vital due process protections are increasingly lost.

While the court allowed future use of AI systems like COMPAS, it required a list of 'advisements' and 'cautions' to be provided to judges along with the AI systems' risk assessments scores. These are:

1. "The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined."

2. "Because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders — not a particular high-risk individual."

3. "Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism."

4. "A COMPAS risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed. Risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations."

5. "COMPAS was not developed for use at sentencing, but was intended for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole."

It is unclear how, if at all, these so-called 'cautions' and 'advisements' actually protect defendants' rights. A court has full discretion on whether and how to take this list into account, and there is absolutely no guarantee that the list will have any impact whosoever on how much the risk score affects a judge's decision. Perhaps most importantly, there is still absolutely no mention of what type of AI system was used or how it was evaluated. Was it a logistic regression? Was it a decision tree? Was it a language model? No one knows, and no one has the right to know, because even the most basic safety information is protected a trade secret. Even if the courts and the lawyers involved had the technological expertise to understand how to evaluate the model (and they don't), their door would be slammed shut and barred by the sweeping trade secret protections courts have upheld for these AI tools.

4.7. Trade secret protections

Northpointe claimed, and the Wisconsin Supreme Court upheld, sweeping trade secret protection due to the "proprietary" nature of the COMPAS AI system. While there are certainly reasons to allow *some* protection to reward resources invested in innovation, trade secret protections must allow for public safety provisions. There are ways to protect individuals without unduly harming commercial interests.

One of the most well-known trade secrets is the Coca-Cola formula. It remains fiercely guarded to this day, after over a century of sales. However, trade secret protection does not exempt the Coca-Cola company from providing the required list of ingredients on all their products, as well as a standard-format nutrition label. Presumably, this is because we have decided that consumer protection is an important factor to balance with trade secret protections. We feel it is unfair to leave ordinary people in the dark about something so crucial to their well-being. One ingredient prominently listed on the label of Coca-Cola beverages in the U.S. is high-fructose corn syrup, used to sweeten the soda. Without this vital ingredient information, global public health and safety studies would not have been able to find a link between products containing high-fructose corn syrup and obesity and diabetes on a global scale (Goran et al., 2013). Governments would not have been able to use this information to institute special taxes on sugary beverages, and launch public health campaigns to inform us all about healthy choices. A total deference to trade secret protections would have left consumers, academics, and independent agencies in the dark.

As dangerous as unfettered soda-drinking may seem from these studies, there is something that makes unfettered AI-driven decision-making even more dangerous: individuals can (and have always been able to) simply refuse to drink soda for any reason. They have both an ingredient list and a choice. They cannot, however, refuse to be subject to AI-driven decision-making – indeed in many cases they do not even have the right to *know* that they are being subject to AI-driven decision-making – and they certainly aren't provided with a nutrition label for an AI system. Surely a disregard for due process, individual rights, and simple human dignity cannot continue to be broadly justified using trade secret protection as a blanket excuse – a balance of proprietary protections and public transparency is called for.

5. Alternative Views

There are several viewpoints that take alternative positions on the issue of due process and AI. First, there is the view that broad and sweeping trade secret protections should be ironclad to protect innovation (Klein, 2023). Certain advocates for the COMPAS system have also opined that however biased these systems may be, they are still faster and less problematic than the alternative human decision makers (judges). There is also the view that courts themselves often take, discussed in the previous section, which is that racially-biased AI systems are permissible as long as they come with a warning label that tells judges about their pitfalls. To the author's knowledge, there has not been a comprehensive study of how resistant judges are to the cognitive biases that usually affect human decision-making in cases where they are permitted to view racially-biased output but mandated not to let it bias their final decision.

There is also a prominent *gap*, not necessarily (or consciously) in position but rather in awareness, between the technical machine learning community and the legal practitioners who have historically upheld due process rights. While the machine learning community is largely aware of models' technical issues (motivating research efforts in trustworthiness, xAI, interpretability, and safety), this knowledge has not translated effectively to legal practitioners, who are unprepared to litigate the technical nuances that lead to legal due process violations. The alternative view, then, is perhaps that the due process implications of today's AI systems are not sufficiently impactful to study or address.

6. Due process as an ideal

Companies have often stated that it is difficult to operate in a world with rapidly-changing AI regulation standards. Due process provides not only an actual limit on government action, but also a guideline for private action; it can help provide the public and private sectors with an understanding of what kind of AI-driven behavior is acceptable and what is not. If the outcome of using an AI system would seem arbitrary or unfair (Rane, 2024), if notice cannot be provided with sound, technically verifiable reasons for the AI's decision, if a fair hearing cannot be guaranteed with the opportunity to contest the decision, then it is likely that the AI system in question violates the spirit, if not the letter, of due process. Companies and government entities wishing to anticipate these regulatory risks and create better AI systems can use due process as a minimum checklist for acceptable AI-driven behavior. Just as privacy scholars have argued that stable, settled privacy best practices increase consumer trust and are good for business (Waldman, 2018), AI systems (and boundaries within which AI should not be used) that honor the spirit of due process protections will increase public trust and mitigate regulatory risks.

7. Informing the legal process

As AI researchers, we take the lead in helping the world understand what acceptable AI-driven behavior should be, what it (from a technical standpoint) *cannot* be, and what it *should not* be. This requires far greater engagement with the real world in which our AI systems are now deployed. Our task is clear:

1. To illustrate with empirical evidence when and how due process and other fundamental rights may be under threat from AI-driven decision-making.

2. To investigate how much information we need to thoroughly interrogate and evaluate models, and develop new techniques for doing so that can protect *some* degree of proprietary protections in some contexts.

3. To research simple and intuitive ways to explain to ordinary people how an AI system is reaching a particular decision, and to clearly convey when it is impossible to know this.

4. To provide courts with easily-understandable technical arguments for why, in many scenarios, AI systems should not be subject to sweeping trade secret protections and other mechanisms for secrecy – especially when they infringe on the due process rights of vulnerable, ordinary people.

5. To highlight areas where AI decision-making should

not replace human decision-making at all, and to explain the reasons why.

Acknowledgments

Thank you to the faculty and students of UChicago Law School, Princeton University, and the ICML position paper track audience for their feedback.

Impact Statement

There are many potential societal consequences of this work, all of which are thoroughly discussed in the main text.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

American Association on Intellectual and Developmental Disabilities. SIS FAQs. https://www.aaidd.org/sis/faqs, 2024. [Accessed 19-01-2024].

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Broniatowski, D. A. et al. Psychological foundations of explainability and interpretability in artificial intelligence. *NIST, Tech. Rep*, 2021.

Brown v. Board of Education. 347 U.S. 483, 74 S. Ct. 686, 98 L. Ed. 873, 1954.

Carey v. Quern. 588 F.2d 230, 232, 7th Cir., 1978.

Citron, D. K. Technological due process. *Wash. UL Rev.*, 85:1249, 2007.

Citron, D. K. and Pasquale, F. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

Deeks, A. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7):1829–1850, 2019.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Eberle, E. J. Procedural due process: the original understanding. *Const. Comment.*, 4:339, 1987.

Freeman, K. Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. *North Carolina Journal of Law & Technology*, 18(5):75, 2016.

Goldberg v. Kelly. 397 U.S. 254, 90 S. Ct. 1011, 25 L. Ed. 2d 287, 1970.

Goodman, C. C. AI, can you hear me? promoting procedural due process in government use of artificial intelligence technologies. *Rich. JL & Tech.*, 28:700, 2021.

Goran, M. I., Ulijaszek, S. J., and Ventura, E. E. High fructose corn syrup and diabetes prevalence: a global perspective. *Global public health*, 8(1):55–64, 2013.

Houser, K. A. and Sanders, D. The use of big data analytics by the IRS: Efficient solutions or the end of privacy as we know it. *Vand. J. Ent. & Tech. L.*, 19:817, 2016.

Hurley, M. and Adebayo, J. Credit scoring in the era of big data. *Yale JL & Tech.*, 18:148, 2016.

Klein, M. A. Trade secret protection, multinational firms and international trade. *International Economics*, 173: 325–342, 2023.

KW v. Armstrong. 180 F. Supp. 3d 703, D. Idaho, 2016.

Lawrence v. Texas. 539 U.S. 558, 123 S. Ct. 2472, 156 L. Ed. 2d 508 , 2003.

LS v. Delia. NO. 5:11-CV-354-FL E.D.N.C., 2012.

Molnar, C. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

NPR. Feds Say Self-Driving Uber SUV Did Not Recognize Jaywalking Pedestrian In Fatal Crash. https://www.npr.org/2019/11/07/777438412/feds-say-self-driving-uber-suv-did-not-recognize- 2019. National Public Radio (NPR).

NYTimes. Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html, 2021. The New York Times.

Obergefell v. Hodges. 575 U.S. 994, 135 S. Ct. 2071, 191 L. Ed. 2d 953, 2015.

ProPublica. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentenc 2016. [Accessed 19-01-2024].

Rane, S. The Reasonable Person Standard for AI. In *Forty-first International Conference on Machine Learning*, 2024.

Rane, S., Ho, M., Sucholutsky, I., and Griffiths, T. L. Concept alignment as a prerequisite for value alignment. *arXiv preprint arXiv:2310.20059*, 2023.

Rane, S., Bruna, P. J., Sucholutsky, I., Kello, C., and Griffiths, T. L. Concept alignment. *arXiv preprint arXiv:2401.08672*, 2024.

Roe v. Wade. 410 U.S. 113, 93 S. Ct. 705, 35 L. Ed. 2d 147, 1973.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Spaulding, N. W. The ideal and the actual in procedural due process. *Hastings Const. LQ*, 48:261, 2020.

State v. Loomis. 881 N.W.2d 749, 2016 W.I. 68, 371 Wis. 2d 235, 2016.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Achterberg, J., Tenenbaum, J. B., et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

The Guardian. What happened when a 'wildly irrational' algorithm made crucial healthcare decisions. https://www.theguardian.com/us-news/2021/jul/02/algorithm-crucial-healthcare-decisions, 2021. [Accessed 19-01-2024].

Waldman, A. E. *Privacy as trust: Information privacy for an information age*. Cambridge University Press, 2018.

Youngberg v. Romeo. 457 U.S. 307, 102 S. Ct. 2452, 73 L. Ed. 2d 28, 1982.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. *Dive into Deep Learning*. Cambridge University Press, 2023.