Machine Learning Algorithms: Should Accuracy Override Transparency When Predicting Recidivism in the Criminal Justice System?

Examination Number: B180230

7967 words

MSc Epistemology, Ethics and Mind
## CONTENT TABLE

1. Introduction ......................................................................................................................... 3
2. Machine Learning Algorithm to predict recidivism: accuracy and transparency ........ 5
3. The argument of Equivalent opacity .................................................................................. 7
   3.1 Publicity transparency ...................................................................................................... 7
   3.2 Intelligibility & opacity of human decisions ................................................................. 8
   3.3 Against double standards: the requirement of equal treatment ............................... 9
4. Response to Equivalent opacity: ......................................................................................... 10
   4.1 Human and MLA opacity are different ......................................................................... 11
       4.1.1 Humans can give explanations .............................................................................. 11
       4.1.2 Human explanations have a regulatory and normative role .................................. 14
       4.1.3 Human decision-making is part of an organisational system .............................. 15
   4.2 Double standards of transparency (P3)......................................................................... 16
   4.3 Individual accuracy and transparency .......................................................................... 17
       4.3.1 Identifying mistakes in individual cases ............................................................... 18
       4.3.2 Improving how we understand recidivism ........................................................... 18
5. An overall objection: should we value transparency over accuracy? ............................. 20
6. Conclusion ........................................................................................................................... 21
7. Bibliography ....................................................................................................................... 23
Machine learning algorithms: should accuracy override transparency when predicting recidivism in the criminal justice system?

1 INTRODUCTION

Machine learning algorithms (MLAs) to predict recidivism - the likelihood that an offender will reoffend – are increasingly used for decisions on pre-trial detention, sentencing, probation, and parole. Accuracy in adjudication is a central value in the justice system and transparency is a vital attribute for explainability, oversight, and contestability of decisions. However, the use of machine learning algorithms (MLA) has seemingly made accuracy and transparency into conflicting priorities. The primary reasons in favour of MLAs, as opposed to using individual assessments by humans or simpler actuarial tools, is that MLAs improve the accuracy of predictions, overcome the individual biases of human agents, and can help improve cost efficiencies. Predictive accuracy is often presented as the primary selling point; improved accuracy will help safeguard the community and can result in a better rationalisation of resources in the prison system. But MLAs are also a black box with no insight or explainability on how predictions are reached. There are two different ways in which algorithms can be a black box: either due to unintelligibility or due to lack of access. Unintelligibility is caused by the intrinsic complexity of machine learning techniques: the more complex the algorithm, the greater the opacity. Lack of access is due to corporate proprietary measures used to protect the algorithm as a trade secret. The lack of transparency of MLAs raises epistemic and ethical concerns, such as on fairness, on an individual’s right to explanation, on contestability of decisions, and on oversight and accountability in the judiciary (Grant et al., 2023; Grimmelikhuijsen, 2023, 2023; Mittelstadt et al., 2016; Rudin, 2019).

A common argument used to neutralise concerns on opacity of algorithms claims that lack of transparency is as much a problem with humans as with MLAs. As such, it is misplaced to require higher standards of transparency from MLAs compared to humans, and to use transparency concerns as a valid reason against MLAs. All other things being equal, we should favour the use of MLA to improve the accuracy of recidivism prediction. I will refer to
this as the argument of equivalent opacity\(^1\) (Bagaric & Hunter, 2022; Chiao, 2022, 2023; Peters, 2023; Thomsen, 2022; Zerilli et al., 2019).

The concern this paper will address is whether the opacity of black-box algorithms is equivalent to the opacity of human decisions in predicting recidivism, and if greater accuracy should be privileged over transparency concerns. I will argue that the use of black-box algorithms to predict recidivism does increase opacity of decision-making in the justice system, and that it is justified to call for methods that are more transparent. The argument of equivalent opacity provides a reductionist view of human decision-making which creates a false equivalence with algorithms. Furthermore, transparency is necessary to help identify, and correct errors made in individual risk assessments and to allow for improvement in how the justice system manages recidivism generally.

I will begin in section 2 by describing the use of MLA to predict recidivism, and the importance of both accuracy and transparency. I will then explain the argument of equivalent opacity in section 3, based primarily on Chiao (2022) and Zerilli et al (2019). In part four, I will raise three objections to the argument of equivalent opacity, providing a response on why human and algorithmic opacity are not equivalent and why it is justified to request more transparency from algorithms. I will also address possible objections to each point. In part five I will address an overall objection to my thesis.

First though, I will clarify the scope and terminology used in this paper. While opacity is caused by two very different issues – propriety measures or the complexity of algorithms– I will refer to them both interchangeably as machine learning algorithms (MLA) or black box algorithms. I will treat them separately only in as far as the argument of equivalent opacity treats them as separate notions (notably as set forth by Chiao, 2022) but for the most part will treat them the same.

I will use the term transparency, to refer to the capacity to provide a reasonable explanation (explainability) of a decision that is understandable to human stakeholders. Within the context of recidivism prediction, I define the adequate level of transparency to be the one which enables the right to an individual explanation on decisions, that allows for contestability of decisions and of oversight in the justice system.

\(^1\) Also referred to as ‘argument of equal opacity’ by Peters 2023 and ‘argument from the limitations of human reasoning’ by Maclure 2021.
In the literature, there is discussion on the different meanings of transparency, interpretability, and explainability, and in the world of AI these words can have different technical implications. It is beyond the scope of this paper to go into technical differences between explainable AI (XAI) and interpretable models and which ones are more suited to meet transparency requirements (See for example Creel, 2020; De Bruijn et al., 2022; Peters, 2023; Rudin, 2019). If decisions made by an algorithm are explainable and intelligible in an adequate way, as defined above, then presumably transparency concerns will no longer be applicable, but neither will the argument of equivalent opacity be necessary.

Finally, it is a worth mentioning that several studies find that complex black-box algorithms are no more accurate than interpretable models for predicting recidivism. (Grant et al., 2023; Ingram et al., 2022; Rudin, 2019). Rudin (2019) holds that the belief that there is a necessary trade-off between accuracy and interpretability is a ‘blind myth’ that has lead researchers to prioritise work on black box algorithms and neglect interpretable algorithms (Rudin, 2019, p. 207). These findings could suggest that the debate on transparency and accuracy is resolved. However, considering that black-box algorithms continue to be developed and used to predict recidivism, that proprietary measures also are a cause of opacity, and that arguments of equivalent opacity continue to be mobilised, I sustain that it remains pertinent and important to address the transparency concerns raised by the use of black box algorithms. For sake of argument, this paper will address the argument of equivalent opacity taking as a given that this is in the case of a black box algorithm having higher accuracy rates than other methods to predict recidivism, as higher accuracy is the main selling point used in their favour.

2 MACHINE LEARNING ALGORITHM TO PREDICT RECIDIVISM: ACCURACY AND TRANSPARENCY

MLAs to predict recidivism produce a risk score showing whether an individual has a low or high risk of reoffending. The MLA is trained using large quantities of data to identify patterns that can be used to predict an individual’s future behaviour (McKay, 2020; Mittelstadt et al., 2016). The data used covers a person’s demographics, health and attitudes, their social and physical environment, living circumstances, and criminal record. Data about the offender is collected through self-report assessments and via interviews carried out by evaluators such as probation officers or social workers (Desmarais et al., 2016). Risk assessment aims to identify high risk offenders in order to focus more resources on their control, incarceration or
rehabilitation, and at the same time, triage out the low risk offenders who require less attention and resources (Hannah-Moffat & Struthers Montford, 2019; Van Ginneken, 2019).

Having a high accuracy rate of predictions is a key aim of risk assessments. Their credibility relies on being able to perform accurately. Reaching accurate verdicts in general is a central value of the justice system. If the justice system makes too many mistakes, it cannot fulfil its duty and loses reliability. (Bagaric & Hunter, 2022; Chiao, 2023; Enoch et al., 2012; Ryberg & Petersen, 2022). Chiao calls the importance of accuracy in the justice system the ‘anodyne thought’, where he argues that seeking accuracy of verdict is central in most theories of adjudication that value truth, and thus seeking accuracy should generally be favoured, even if in some circumstances competing values may take precedence (Chiao, 2023, p. 3).

Transparency on how decisions are made is an important quality of the justice system. Transparency is necessary for the right of individuals to have an explanation, for contestability of decisions to be possible, and for oversight and regulation of the justice system. Vredenburgh argues for explanation to be an individual right, grounded in the interest of “informed self-advocacy” to represent one’s needs and interests within an institution (Vredenburgh, 2022, p. 212). Jongepier et al underline the necessity we have as deliberative agents to be able to understand decisions that have a significant impact on us. (Jongepier & Keymolen, 2022). In addition to the individual right to an explanation, Bibal et al (2021) add that public authorities – such as the judiciary – rely on explanations to take ownership of the legality of decisions and to help determine if legality and due process is respected in case of contestation. (Bibal et al., 2021, p. 156).

The importance of transparency and accuracy in the justice system is uncontroversial. However, the apparent trade-off made by MLAs results in a loss of transparency which is cause for concern. The argument of Equivalent Opacity attempts to by-pass this concern. Once it is established that opacity in decision-making exists as much with humans as it does with MLA, the transparency concerns raised by the use of MLA lose their argumentative force. I shall therefore first address the argument of Equivalent Opacity, before arguing in favour of algorithmic transparency.

---

2 Vredenburgh defines informed self-advocacy as a set abilities which allows an individual to represent their own “interests and values to decision-makers and to further those interests and values within an institution.” (Vredenburgh, 2022, p. 212).
3 THE ARGUMENT OF EQUIVALENT OPACITY

The argument of equivalent opacity posits that we should weigh the transparency of MLA against the real situation of transparency in the justice system, rather than against an ideal of transparency (Chiao, 2022, p. 35). Chiao (2022) proposes looking at two different limitations to transparency raised by MLAs: transparency as publicity and transparency as intelligibility. Publicity is about providing access to existing information, rendering it open to public scrutiny, and reflects the type of algorithmic opacity that is caused by trade secrets and proprietary measures. (Chiao, 2022, p. 36). Transparency as intelligibility refers to the explanations given by decision makers on how they arrived at their decision, and reflects the type of opacity caused by algorithmic complexity (Chiao, 2022, p. 41). I will look at each of these in turn.

3.1 PUBLICITY TRANSPARENCY

One common demand on MLA is that trade secrecy be dropped to allow the algorithm to be open to public scrutiny (Chiao, 2022). This includes providing transparency on any factor which contributes to how the MLA reaches a decision, such as information on the content and provenance of the training data used, on the source code, and access to the algorithm before and after training, among others (Chouldechova, 2020, p. 2).

When we compare with the justice system, however, Chiao points out there is no systematic transparency requirement to make public every step of human legal decision-making. For example, the stages of every decision taken in court are not laid out for inspection, judges do not need to explicitly lay out their inference process, and jury secrecy actually protects the jury from opening up to public scrutiny, with jury deliberations taking place in a ‘black box’ (Bagaric & Hunter, 2022, p. 124; Chiao, 2022, p. 37).

When reason-giving is required, Chiao sustains that the reasons given are rarely a detailed description of the decision-making process and what counts as a reason can vary greatly. In many cases, a few sentences with a brief description can suffice. Chiao uses the example of the European Court of Human Rights that has a blanket requirement to provide reasons, but where reason-giving in most cases, is no more than a bureaucratic exercise using boiler plate sentences to get the paperwork through, rather than providing fully fledged reasons (Chiao, 2022, p. 38). Even when reasons are provided, this is much less than the comprehensive transparency to reveal every step of a process that is being demanded of algorithms.
To summarise, principles of public transparency which require having access to each step of the decision making process, shouldn’t be used as a blanket rejection of algorithms because this principle isn’t required of human decision making in the justice system (Chiao, 2022, p. 40).

### 3.2 Intelligibility & Opacity of Human Decisions

Intelligibility refers to a decision makers’ ability to explain their decisions in a way that can be understood by stakeholders. The complexity of MLA can render it extremely difficult or impossible to comprehend how data is correlated to reach predictions. As a result, decisions which impact the lives of individuals are based on a process that no one really understands. The machine ‘rationale’ that will lead an MLA to reach a decision are completely different to the type of factors or rationale that humans would use as an explanation. Creel (2020) describes this as a fundamental ‘difference in kind’ that renders the algorithms inherently opaque and incomprehensible for humans:

> “Even when information that is sufficient for machine classifications or predictions is available, its difference in kind and scale means that it explains little to us.” (Creel, 2020, p. 586);

Although experts have used different techniques to develop post hoc explanations on algorithmic decision making, the process necessarily simplifies what is happening, focusing only on some aspects of the algorithm to render the explanation comprehensible to humans. As a result the post hoc explanations aren’t reflective of the full algorithmic process nor faithful to what the algorithm is actually computing (Creel, 2020; Rudin, 2019). Rudin states that calling these explanations ‘explanations’ is misleading and risks giving false impressions on how decisions are reached (Rudin, 2019, p. 207).

The argument of equivalent opacity requires that we compare this to the transparency of human decisions in the justice system. Numerous studies in social psychology and cognitive science, including on judges, show that human decisions are influenced by an array of extraneous factors, such as cognitive bias and framing, that are outside the decision-makers’ awareness (Chiao, 2022; Cohen, 2015; Tversky & Kahneman, 1981). For example, a 2011 study on the impact of food breaks on judges’ parole decisions found that the rate of positive rulings would be around 65% following a food break, and drop to almost zero throughout the day until after the following food break, when positive rulings would return to the initial 65%
(Danziger et al., 2011). Another study looking at the anchoring effect on judges showed how anchors – the tendency to rely on the first piece of information provided - influenced judges’ sentencing in the hypothetical cases presented to them. This included using irrelevant anchors such as dice rolls, where higher dice rolls tended to result in judges opting for longer sentences (Englich, Mussweiler and Strack 2006 as cited in Chiao, 2022, p. 46).

The work of social psychologist Jonathan Haidt suggests that reasoning around moral judgements occurs after the judgement takes place, and not before. Reasoning is used post hoc to rationalise and develop justifications that are viable (Haidt, 2001). Haidt sustains that when we seek explanations for our moral decisions we cannot access the cognitive processes that caused them, but rather we consult “a pool of culturally supplied norms for evaluating and criticizing the behaviour of others” (Haidt, 2001, p. 822) which we can then use to justify our judgements.

These and other studies show that many of the cognitive processes behind human decision-making are as opaque to us as the processes in MLA. And the explanations we give are not a full description of all the factors that influence how we make decisions. Some authors even suggest that we have more information on the factors that contribute to MLA decision making then to the full range of factors that contribute to decision making in humans (Bagaric & Hunter, 2022, p. 131; Thomsen, 2022, p. 262). Since we cannot know the innerworkings of decisions and explanations at best only partially describe how decisions are really carried out, then both MLA and human decision-making face comparable problems of opacity.

### 3.3 Against Double Standards: The Requirement of Equal Treatment

Zerilli et al (2019) assert that the transparency requirements made on the use of MLA in criminal justice are of much higher standard than what we normally require of the justice system. Such as requests to only use algorithms that can be inspected internally. Given that opacity in MLA and humans can be equated, he argues that there is no reason to make stronger requirements on algorithms than we do on humans. (Zerilli et al., 2019, p. 668). To meet the same standard of explanations we use for humans, Zerilli proposes that for algorithmic decisions we use what Daniel Dennet (1981) calls the ‘intentional stance’ level of explanation.

Intentional stance explanations are those which attribute mental properties – such a desires and beliefs – as reason-giving explanations. Intentional stance explanations are the way
humans normally explain their own decisions and those of others. This is as opposed to using either the physical stance which concerns physical laws or the design stance which looks at an engineering level of explanation (such as for an organism or an artefact) (Zerilli et al., 2019, p. 669). Using the intentional stance means attributing ‘mental states’ to an algorithm. For example, to explain why an image recognition algorithm identifies a wolf in a photo, one would attribute a mental state as if the algorithm thought “I believe that snow in an image indicates a wolf; image X had snow in it, so I concluded it must be a wolf” (Zerilli, 2022, p. 15). Using a design explanation for an algorithmic decision would be unnecessary and a double standard because we do not require that level for a satisfactory explanation of human decisions.

The argument of intelligibility shows that human explanations are incomplete but are satisfactory for humans, so post-hoc explanations of MLA decisions framed as an intentional stance should also be satisfactory even if they are not faithful to the whole algorithmic process. If we hold both to the same standard, this means that it is unnecessary to understand the innerworkings of algorithms and unnecessary to have full explanations on how the decision-making process of algorithms takes place. As such, and all other things being equal, there is no reason why transparency concerns should trump the benefits of increased accuracy of MLA to predict recidivism, as doing so would constitute a double standard of transparency.

In summary, the argument of equivalent opacity states that lack of transparency of MLAs should be measured against the real transparency of human decision making in the justice system, and not ideal transparency requirements. Given that publicity transparency and intelligibility transparency in humans are no better than in MLA, using principles of transparency to object to MLA is not justified, and a similar level of explanation for decisions should be acceptable to both even if both are incomplete or not fully representative of the decision making process.

4 RESPONSE TO EQUIVALENT OPACITY:

A key premise throughout the argument of equivalent opacity is that we should benchmark the transparency of MLAs against the transparency of human decision-making in the justice system, and not against ideal transparency standards. A limitation of this approach is that it tends to focus on how opacity in humans resembles MLAs but doesn’t consider their differences. A more instructive comparison would be to see whether MLA and human
decision-making are equally able to help meet adequate transparency levels. I will take as an adequate level of transparency to be one which enables the individual’s right to an explanation, and allows for contestability of decisions and oversight in the justice system.

Taking this as a starting point, I will address three main problems with the argument of equivalent opacity. First, MLA and human decision-making cannot be equated because humans can provide explanations and MLAs cannot. Human explanations have normative and self-regulatory attributes that MLAs lack. Secondly, the intentional stance is not satisfactory for MLAs used to predict recidivism because there is a difference in how MLA and humans can ‘operationalise’ transparency to provide an adequate explanation. Finally, I will argue that a certain level of transparency is needed for improved accuracy of individual recidivism predictions and to help improve how the justice system addresses issues around recidivism.

4.1 Human and MLA opacity are different
There are three ways in which human decision-making differs from MLA that haven’t been considered in the argument of equivalent opacity: humans can give explanations, their explanations have a regulatory and a normative role, and human decision-making is part of an organisational system. These three attributes place human decision-making in a better position to meet the adequate level of transparency required to enable the right to an explanation, to allow for contestability of decisions and oversight in the justice system.

4.1.1 Humans can give explanations.
Humans can give explanations when MLAs cannot. These explanations provide a basis for feedback, critique, and contestability of decisions. For recidivism prediction, evaluations by psychologists, the use of actuarial tools that are not ‘black boxes’, and decisions by probation officers, parole boards or by judges, all have potential for explanation. It may be that these explanations are not always of good quality, or not consistent, but in principle the explanatory potential is present and can be requested and contested. However, once this decision is made by an MLA, access to an explanation is no longer possible.

In the argument on transparency as publicity, Chiao (2022) aims to show that explanations in the justice system are not as transparent as is often assumed by those who raise opacity concerns on the use of MLA. He sustains that lifting trade secrets to access the innerworkings of MLAs is akin to requiring courts and judges to systematically lay out every step of their reasoning processes behind each decision. As principles of transparency are not mobilised to
require that courts and judges do this, then there is no basis to use principles of transparency to push back on the use of MLAs protected by trade secrecy. (Chiao, 2022, p. 40).

This argument overlooks that the necessary conditions for providing an explanation are not the same for humans and for MLAs. The aim of human explanations isn’t to describe the causal factors behind decision making, nor to reconstruct the full reasoning process behind each decision, but rather to see how a decision is answerable to the explanation, where the legitimacy of the decision is tied to the quality of the explanation. To achieve this, a human explanation doesn’t necessarily have to lay out every step of the reasoning process to be seen as a reasonable explanation for a decision. Even in the circumstance that an explanation describing step-by-step reasoning is deemed necessary, it could still be possible to carry out without it becoming a systematic requirement for all explanations in court. On the other hand, lifting proprietary measures to access an MLA’s innerworkings, such as the data used and the design elements, is arguably a necessary condition to get an explanation on an algorithmic decision because there may be no other way to do so. This is why we can require this of MLA without making it a systematic requirement for courts. Therefore, it is not because this is not a systematic requirement of courts and judges, that it is not a justified requirement to make of MLAs.

Chiao also points to the example of jury secrecy to show that principles of transparency are not always a requirement in all the justice system (Chiao, 2022, p. 36). Having insights into jury discussions would help provide explanations on verdicts reached, but it is accepted that jury discussions happen in a ‘black box’. So why not accept the same from MLAs in predicting recidivism? The difference here is that jury secrecy serves a purpose for justice, such as to avoid jury members being harassed or put under pressure, and to encourage free discussion with the interest of reaching the best verdict possible. But for MLAs there is no similar justification that links the use of trade secrets as a necessary condition for the quality of the prediction or as necessary for the functioning of justice. In addition, secret jury deliberations do not impede contestation, explanations, or oversight. While it isn’t possible to access jury discussions, the whole trial– the process, the evidence, testimonies, and arguments presented – are available for scrutiny and contestation by the affected parties. All the elements that feed jury discussions are available. This is not comparable to the lack of transparency caused by an MLA to predict recidivism where these elements are not available.
Chiao also uses examples of weak reason-giving in courts and by judges to illustrate that the principles of transparency applied in practice are so low that they cannot justify requiring access to the inside workings of an algorithm. However, weak reason-giving by humans in the justice system doesn’t mean that reasons can’t be provided per se. Reason-giving remains possible, but faces constraints, and constraints can be challenged or overcome. Concerning the European Court of Human Rights (ECHR), Cohen (2015) explains that the high number of cases managed by the court leads to limitations in time and resources and in the capacity to provide detailed reasons for all cases. Instead, the court prioritizes qualitative reason-giving for a minority of cases it considers of higher importance and uses boilerplate reasons for the rest. In fact, courts and judges across jurisdictions have to regularly balance competing rationales for and against reason-giving, as Cohen states: “In practice, judges may balance democratic rationales for reason-giving against institutional, cognitive, and pragmatic reasons not to give reasons.” (Cohen, 2015, p. 490).

These descriptions show that activities linked to transparency, such as balancing competing rationales for reason-giving, are being regularly renegotiated within the justice system. Consequently, reason-giving may at times be unsatisfactory or inconsistent, but in all cases, there is a potential for explanation and a possibility to improve. However, this is not the case with MLAs where the potential for explanation is absent.

One objection is that the ‘potential for an explanation’ isn’t enough of a reason to always push back on MLAs. For example, on pre-trial detention and bail, a judge’s opacity is arguably greater than that of an MLA: explanations are often of poor quality or non-existent and decisions are very likely prone to bias. A judge’s discretionary power means that there is no need to justify how they weigh up the risk that the accused will re-offend or the risk she will not turn up to court. The judge’s risk analysis for bail and pre-trial detention can be based on impressions they have of the defendant, without any significant explanation, or evidence, and many jurisdictions lack official guidance on how to reach these decisions (Bagaric & Hunter, 2022, p. 124). If an MLA is used, at least there is consistency in how risk is decided, based on objective criteria rather than on arbitrary impressions. Using an MLA for predicting recidivism in these circumstances by nature of its impartiality appears to improve transparency of the system compared to a judge.

Since MLAs are not subject to cognitive bias and prejudices, this gives the impression that their predictions are non-arbitrary. However, as Leo Breiman illustrated with what he coined as the ‘Rashomon effect’, for any given problem, such as recidivism prediction, it’s possible
to develop a multitude of predictive models with the same level of accuracy (Breiman, 2001, p. 206; Chouldechova, 2020, p. 4). Each model will be as accurate as one another but will have its own characteristics, patterns, and structures. This means that multiple models will be equally accurate in predicting recidivism but could make different predictions for the same individual. So, although the algorithm may not exhibit individual bias in the way a judge does, there is still something arbitrary about how MLAs work and how their decisions can impact an individual.

In addition, using an MLA isn’t going to solve the underlying problem of lack of explainability and contestability in the example of pre-trial detention. The opacity of the judge doesn’t justify rendering a decision permanently opaque by using an MLA, when other options could be available that would make the system fairer and less opaque. For example, it’s possible to establish parameters to help decision making on bail and pre-trial detention, to use algorithms that aren’t black boxes to help the judge in her decision-making process, to provide an account and feedback to judges on their own past decisions, among other solutions.

4.1.2 Human explanations have a regulatory and normative role.

The second difference with human explanations not taken into account in the argument of equivalent opacity, is that explanations have a self-regulatory and normative role. People tend to align their behaviour to the explanations they give. Called ‘mindshaping’, this view holds that people commit to their self-ascribed explanations in order to be more predictable to others (Peters, 2023). Peters uses the mindshaping view to argue that the role of explanations in self-regulation contributes to improving the transparency of human decisions, as the explanations will influence future decision making (Peters, 2023, p. 9). Similarly, in a comparative study on reason-giving in courts, Cohen (2015) describes how explanations play a regulatory role that helps to improve the accuracy of judges. Reason-giving in judges is a form of self-discipline that helps to limit the discretionary power of judges and improve the quality of decisions made (Cohen, 2015, p. 511). In this sense reason-giving plays a normative role, as the legitimacy of the decision is tied to the quality of the explanations given.

So even if explanations on decisions don’t take into account the hidden cognitive processes and biases, and even if there may be a ‘post-hoc rationalisation’ to justify initial moral judgements as Haidt sustains, explanations do play a role in future decisions, and the
legitimacy of decisions becomes tied to the quality of explanations provided. This confers human explanations with a certain reliability that post hoc explanations derived from MLA don’t have. MLA post hoc explanations have no link with the algorithm and play no self-regulatory nor normative role.

4.1.3 Human decision-making is part of an organisational system.

Arguments positing that opacity between humans and MLA are comparable often refer to humans as a single entity. However, within the justice system, this ‘human’ is part of a broader system, where work is carried out with peers and colleagues and where there is regulation and oversight of the judiciary and other professionals. There are judicial review committees, units overseeing professional standards, ethical commissions, codes of conduct, and complaint procedures, in addition to the appeals process. Professionals must internalise processes acquired in their training, professional values, include self-reflection and critical analysis on justice issues and decision making. Some of these internalised features are also part of the ‘bundle’ of unseen cognitive processes behind decision making (Peters, 2023, p. 8). Maclure describes judicial reasoning as a social institution that was designed to help overcome bias, in her words “with the purpose of neutralizing to the greatest extent possible the cognitive limitations of individual judges.” (Maclure, 2021, p. 10).

So, while officials may have discretionary power over certain decisions, there is also a regulation and oversight system which isn’t considered in the argument of equivalent opacity. This system helps to check discretionary powers and guarantee the rights of individuals. When you take a decision-making process that was a part of this system and place it in a black-box, such as an MLA, you are essentially cutting off the decision-making process from the rest of the regulatory and normative system, and reducing its institutional transparency.

Admittedly, it doesn’t necessarily matter for the MLA if it is cut out from the regulatory system, because an algorithm doesn’t need the same regulations as humans do: it won’t be subject to individual bias or mood swings or conflicts of interest, that can impact its decisions. And an MLA can be regularly monitored for fairness, checked for predictive validity, and updated with new training data or adjusted to new policies.

However, even if for the functioning of the MLA it doesn’t matter if it is cut from the rest of the system, it does matter for the stakeholders in the system. It matters for the individuals who are being scored by the algorithm and by the officials who are using the risk score to make
decisions. Officials cannot explain how the score was reached and individuals who are scored are not able to access an explanation nor contest it. The justice system in general no longer has access to the reasons behind decisions it must stand by, nor can it question how these decisions are made.

4.2 **Double Standards of Transparency (P3)**

Zerilli (2019) argues that intentional stance explanations are satisfactory for humans, and so they should be satisfactory for MLA. In a broad sense, I also argue that standards are the same for both: adequate transparency is one that enables an individual’s right to explanation, allows for contestability, and for oversight in the justice system. However, in a finer sense how we produce explanations that reach these standards will differ between MLA and humans. Given that the ‘rationale’ of an MLA is so different to that of a human, it is hard to see how to confidently attribute mentalistic properties (such as ‘belief’) as suggested by Zerilli, to complex data patterns used to predict recidivism, knowing that it can result in a misleading picture on how the algorithm reached its decision. Neither is it clear that this is always the most useful explanation for an MLA decision (Günther & Kasirzadeh, 2022, p. 380).

Günther and Kasirzadeh (2022) illustrate the usefulness point with the example of an airplane crash that was caused by a fault in the algorithm that managed airplane stability. From an intentional stance perspective one could explain the fault by saying that the “The system ‘decided’ to push down the nose to control for the false ‘belief’ that the airplane goes up” (Günther & Kasirzadeh, 2022, p. 377). However, it was the design stance level of explanation which permitted technicians to identify why the algorithm had such a ‘belief’ so that they could fix it; notably, that the algorithm was designed to depend too heavily on one sensor which happened to be the sensor that malfunctioned.

Admittedly, in some situations, using the intentional stance may help provide clarity on the decision of an MLA, such as the example Zerilli provides with the picture recognition of a wolf or dog. It may also depend on what the algorithm is being used for. In picture recognition it is much easier to verify an error – either the MLA recognised the dog or it didn’t. It is arguably, easier to figure out why it is making the mistake; that all the pictures of wolves it trained with had snow. But with recidivism prediction, it isn’t possible to verify a mistake in the same way, because it’s about the prediction of future human behaviour. We cannot identify whether a mistake in prediction has been made, and need to use the
explanation to help decide whether the prediction seems valid – as an overall justification for the decision. It’s as if we had to decide whether the recognition of the wolf is valid without having access to any pictures, but simply by looking at the patterns in the data used to predict future pictures. Must the global modelling made by the MLA be considered? Or only the specific factors that contributed to the individual decision? Is it even possible for a human to reconstruct a ‘rationale’ or is the machine logic – the patterns found in the data – beyond the rationale of human explanation as suggest by Creel (2020) when he says it represents a difference ‘in kind’?

To illustrate mistakes made with such explanations, Rudin (2019) points out that many MLAs used to predict recidivism use data on age and criminal history, and don’t explicitly use race. However, as age and criminal history are often correlated with race in data sets, it’s possible for a post hoc intentional stance explanation to say “the MLA believes you will reoffend because of your race” even if race isn’t an explicit data entry. Rudin argues that this explanation misrepresents what the MLA is actually doing. This sort of shortcut can be misleading and not help to identify how the correlation of indicators results in bias and unfairness in of decisions (Rudin, 2019, p. 208).

In conclusion, we cannot apply a blanket strategy on both MLA and humans on what types of explanations should be carried out to meet the defined standards of transparency. Looking into the innerworkings of an algorithm may be necessary to understand how a decision is reached. But in humans – aside from this being hardly ever feasible – it isn’t necessarily relevant to provide an adequate explanation for a decision.

### 4.3 Individual Accuracy and Transparency

A last overall objection to the argument of equivalent opacity is that it presupposes a disconnect or negative correlation between accuracy and transparency, presumably because from the perspective of MLA, increased accuracy has been associated with growing complexity and less intelligibility. However, increased accuracy can also be associated with increased transparency. There are two ways in which accuracy is connected to transparency in recidivism predictions: it can help identify mistakes in individual cases and it helps improve how we understand recidivism.
4.3.1 Identifying mistakes in individual cases

In recidivism prediction it isn’t possible to know whether the risk score is mistaken until months or years later because the prediction concerns future human behaviour. In addition, for ‘false positives’ (an individual who is mistakenly given a high-risk score of reoffending), it’s possible that we can never really know whether a mistake was made. If the individual doesn’t reoffend it could be interpreted as being someone who was high risk, but who thanks to the prison programme was rehabilitated. This difficulty in identifying mistakes underlines the importance in understanding how the risk score is calculated. In addition to allowing for contestability, transparency can help improve the accuracy of individual cases by reducing errors in the process.

Errors on individual cases can occur due to human error in collecting and uploading data, a risk that increases with MLAs which require a large amount of data (Grant et al., 2023; Rudin, 2019). Mistakes also occur if a relevant element in an individual’s life is excluded from the risk analysis. It should be necessary for decision makers such as parole boards and judges to be aware of any relevant individualised evidence which is not used in the algorithmic prediction. Being able to provide an explanation on how a risk score is obtained opens the possibility for contestation, which can help identify errors in the process that need to be corrected or adjusted.

In MLAs there can also be mistakes which result from overfitting. This is when an MLA identifies patterns resulting from “noise” in the training data which shouldn’t be generalized to real world situations. An MLA can achieve a high accuracy rate in the training data but be prone to error in real life situations because they create patterns in the training data which are not applicable to real life situations. Such as the MLA using snow to decide that the picture was of a wolf and not a dog. The likelihood of this happening increases with the complexity and amount of data used, and can also be caused because the training data is not representative of the real population (Grant et al., 2023, p. 11). Access to internal MLA workings can help identify and reduce overfitting errors that can impact individual cases.

4.3.2 Improving how we understand recidivism.

Transparency around how recidivism is predicted, contributes more broadly to deliberations on how we define recidivism, and what the implications are for criminal justice and society at large. Measuring recidivism is not a value neutral exercise and there is no single approach to calculating recidivism that can claim objective supremacy. Different methods are used to
predict recidivism and to measure accuracy, which can vary across algorithms, predictive models used, and on the data collected.

For example, as re-offense is not directly observable, a variety of proxy-indicators can be chosen to represent re-offense, such as re-arrest, being charged with a crime, being convicted, or missing a parole meeting, among others (Biddle, 2022). There are also different time periods that can be selected for recidivism, ranging from within 6 months up to several years after release from prison. There are choices on what concepts of fairness are used to evaluate fairness in a system, on how to measure accuracy such as whether reducing false negatives or false positives are prioritized, and on how to define the cut off points in risk scores to define what is a ‘high risk’ or a ‘low risk’ category. Each of these choices incurs different trade-offs that are not value free (Biddle, 2022).

Biddle points out that using re-arrest as an indicator produces more data more quickly, which is in the interest of developers who depend on large data sets (Biddle, 2022, p. 332). However, using re-arrest as an indicator means that more people are counted as reoffenders, and this will disproportionately impact people living in areas that are often targeted by law enforcement. An arrest will confirm the accuracy prediction of the algorithm but won’t reveal for example that the person acted in self-defence or was in the wrong place at the wrong time and arrested. The algorithm was accurate if accuracy is measured by an arrest, but maybe the person didn’t commit a crime. Biddle points out that criminals who are not often arrested – such as for white collar crimes – will be less represented in the data and less likely to be classified as high risk reoffenders (Biddle, 2022, p. 332).

Accuracy in the justice system therefore also depends on being aware of the value laden choices being made on recidivism prediction to see whether and how the criteria need adjusting or contesting. ‘Black boxing’ all these value laden choices into an MLA will ‘cover up’ and render opaque any human valence that has been embedded into the system through human decisions, trade-offs and assumptions on recidivism, accuracy, and relevant data.

In conclusion, when we compare how MLA and human decision-making fare, in terms of reaching the stated transparency goals, human decision making is more promising. Explanations and contestation are possible, explanations are not disconnected from the decision-making process but can feed it, and all of this happens within an oversight system where the concepts and theories of justice continue to evolve through the practice of justice. Shortcomings of transparency do exist, linked to organisational problems and competing
priorities within the justice system, due to human errors and misconduct, due to political and economic pressures, and due to corruption. But contestation, oversight, and a right to an explanation remain feasible throughout.

5 AN OVERALL OBJECTION: SHOULD WE VALUE TRANSPARENCY OVER ACCURACY?

An overall objection to my argument is that having explanations or more knowledge on how we make decisions isn’t necessarily more desirable than overall accuracy, even if this may decrease errors made in individual cases. According to Chiao’s ‘anodyne thought’ accuracy in adjudication must generally be favoured, even if other values at times may take precedence. David Enoch et al (2012) uses a thought experiment that illustrates the value of accuracy over knowledge in the justice system:

“Suppose you have to choose the (criminal) legal system under which your children will live, and you can choose only between systems A and B.

System A is epistemologically better: perhaps its courts only convict when they know (or think that they know) the accused is guilty, or perhaps they only convict based on sensitive evidence, or perhaps they convict only based on evidence that normically supports the conclusion that the accused is guilty.

System B is not as good epistemically as System A. But System B is more accurate, so that the chances of System B convicting an innocent are lower than the chances of System A doing so.” (Enoch et al., 2012, p. 212).

Most people would arguably go for the less knowledgeable and more reliable system B, where less innocent people go to prison. This can be used to illustrate the interest of using MLA to increase accuracy rates in recidivism prediction, even if there is less transparency on decision making.

However, one needs to consider a slightly modified version of the thought experiment to have the full picture. In this case, System A, in addition to having a lower accuracy rate compared to B, also has the possibility to contest convictions. System B, on the other hand has no possibility to contest convictions. In system A there’s a higher likelihood for an innocent person to be convicted compared to B, but in system B, when an innocent person is convicted, they cannot appeal. Now choosing which system you prefer your children to live in is maybe not so obvious.
Ryberg & Peterson (2022) suggest that before assuming whether there must be a trade-off between accuracy and transparency, it is necessary to identify if the cost associated with lack of transparency can be overcome through other means. For example, if the cost is identified as losing trust in the justice system, then in order to maintain trust, a judge can explain to the decision subject why the algorithm is used and how the judge integrates it into her decision making (Ryberg & Petersen, 2022, p. 69). However, if we take the ethical cost to be for the right of individuals to have an explanation and for contestability of decisions and oversight in the justice system, it is hard to see an alternative solution other than using methods – algorithmic or other - that are not black-boxes.

6 Conclusion

Machine learning algorithms (MLAs) to predict recidivism are used in high stake decisions which impact the lives of individuals. One of the main claims in favour of MLAs is that they increase the overall accuracy rate of predictions. However, this comes at a cost in transparency as it isn’t possible to obtain explanations on how the predictions are carried out. Proponents of MLA assert that transparency concerns are unfounded because algorithms are no more opaque than human decision-making in the justice system. I have argued that this claim of equivalent opacity provides a reductive description of human decision-making which results in a false equivalence between MLA and human decision-making. It doesn’t take into account that human explanations are embedded in a web of self-regulatory, normative and institutional systems, that MLAs cannot replace. This explanatory capacity is necessary to respect an individual’s right to explanation, and allows for contestability and oversight within the justice system.

I have suggested that even if MLAs may improve general accuracy rates, transparency in how decisions are made can improve the accuracy of decisions on individual cases by helping to identify errors and by considering evidence relevant to the case. This is all the more important as mistakes in recidivism predictions are not open to immediate verification, unlike algorithms used in picture recognition or diagnosis. Finally, I have argued that transparency is needed to allow for broader deliberation on how recidivism is defined and measured. There are many possible approaches to predicting recidivism and each approach will be value laden and include human assumptions about crime, accuracy and risk that shouldn’t be hidden, but remain open to deliberation. Admittedly, the transparency of human decision-making in the justice system faces numerous shortcomings and challenges. However, the solution to these
problems isn’t to place a high stakes decision in a black-box algorithm that no one can explain or contest.


