The Use of Algorithms in Decision-Making: An Ethical Conundrum

Amber Francisca Ling Yi Rong

Abstract

This essay considers the ethical complexities of using algorithms in decision-making processes in the justice system in the UK, with a focus on the principle of explainability. While algorithmic decision-making tools offer improved efficiency and consistency, their implementation within justice systems raises serious accountability, transparency, and fairness issues. The paper argues that it is fundamentally challenging to establish and measure standards of explainability because of stakeholder's varying requirements, the technical obscurity of the "black box" system, and legal issues related to proprietary rights. It also considers the ethical challenge of interpretability and completeness through case studies of COMPAS and Loomis. The essay concludes that although complete explainability remains elusive, greater inclusivity in AI design, regulatory responsibility, and user autonomy can promote ethical governance and enhance public trust in algorithmic decision-making within the justice system.

Introduction

Algorithms have become an integral part of modern life, and there appears to be a growing interest in their use within the UK justice system. Lord Chancellor Shabana Mahmood has expressed support for utilising these systems to address the mounting backlog of court cases, which has reached a record-high level of 73,000 in December 2024 from 38,000 in December 2019.[1] While the promise of enhanced efficiency is appealing, studies have also highlighted the critical importance of receiving explainability from these algorithms. This matter is reflected in the European Ethical Charter, yet the charter provides little clarity on how the standard of explainability should be met.[2] This paper thus contends that explainability is an important consideration in algorithmic decision-making in the justice system and explores the

[1] GOV.UK, 'Courts reform to see quicker justice for victims and keeps streets safe' (GOV.UK, 12 December 2024) <https://www.gov.uk/government/news/courts-reform-to-see-quicker-justice-for-victims-and-keepsstreets-safe> 16 December 2024.

[2] The European Commission for the Efficiency of Justice, 'European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment' (December 2018) <https://rm.coe.int/ethical-charteren-for-publication-4-december-2018/16808f699c> accessed 13 December 2024.

technical, practical and ethical challenges that make defining and achieving explainability particularly difficult.

The Importance of Explainability in Algorithmic Decision-Making

Explainability, defined as a human-interpretable description of how a decision-maker arrived at a particular decision after considering a set of facts,[3] has become a central focus in ethical and policy discussions worldwide. This is distinct from transparency where transparency answers "what happened" by making users aware they are dealing with AI, while explainability clarifies how the system reached its decision. [4] Jacob Turner notes that transparency is a recurring theme across various ethical codes in different areas of expertise,

reflecting widespread concerns about the risks of opaque AI systems.[5] However, this need for explainability is even more pronounced in the context of justice systems, as the ICO Report reveals that jurors overwhelmingly prioritise explanations in criminal justice scenarios over every other scenarios, such as healthcare.[6] This preference stems from the value explanations provide in challenging decisions, building trust, ensuring fairness and to prove the absence of

[3] Doshi-Velez and others, 'Accountability of AI Under the Law: The Role of Explanation' (2017) Berkmana Center Research Publication, 4.
[4] Arsen Kourinian and Mayer Brown, 'Addressing Transparency & Explainability When Using AI Under Global Standards' (Bloomberg Law 2024) <https://www.mayerbrown.com/-/media/files/perspectivesevents/publications/2024/01/addressing-transparency-and-explainability-when-using-ai-under-globalstandards.pdf%3Frev=8f001eca513240968f1aea81b4516757#:~:text=Global%20AI%20standards%20often%20 group,decision%20was%20made%20using%20AI> accessed 18 August 2025.
[5] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan Cham 2018) [6] ICO, 'Project Explain interim report' (June 2019) <https://ico.org.uk/media/2615039/project-explain20190603.pdf> accessed 12 December 2024.

bias. [6] Burrell reinforces this by stating that explanations support individual autonomy, allowing individuals to contest decisions and maintain agency of their treatment.[7]

The issue of "machine bias" and inaccuracy further underscores the need for explainability in algorithmic decision-making in the justice system. This is problematic as these tools make morally consequential and socially significant decisions.[8] One notable example is the COMPAS algorithm, which calculates the likelihood of recidivism of prisoners for their release. To illustrate this issue, ProPublica's research revealed that COMPAS falsely flags Black defendants as future reoffenders at nearly twice the rate as White defendants (45% compared to 23%).[10] Dartmouth's finding also adds that COMPAS is no more accurate than predictions made by non-experts in the justice system.[9] These cast significant doubt on the reliability of such systems. Richard Susskind[10] emphasises that functional AI must be transparent and capable of heuristic reasoning, which means systems should be open to challenge and improvement. This reinforces why providing explanations is essential to understanding, improving and ensuring the reliability of decision-making tools, while also enabling checks and balances to prevent bias and discrimination.[11]

---

[6] ibid.

[7] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) < https://dx.doi.org/10.2139/ssrn.2660674> accessed 12 December 2024.

[8] Joel Walmsley, 'Artificial intelligence and the value of transparency' (2021) 36 AI & Soc 585, 588. [10] Julia Angwin and others, 'Machine Bias' (*ProPublica*, 23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 3 December 2024.

[9] Julia Dressel and Hany Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism' (2018) 4(1) Science Advances eaao5580.

[10] Richard Susskind, 'Expert Systems in Law: A Jurisprudential Approach to Artificial Intelligence and Legal Reasoning' (1986) 49 MLR 168, 173.

[11] The Royal Society, 'Explainable AI: the basics' (November 2019) <https://royalsociety.org/newsresources/projects/explainable-ai/?utm_source=report&utm_medium=print&utm_campaign=aiinterpretability/> accessed 10 December 2024, 9-10.

Lastly, taking a broader view, Lisa Webley underscores that explainability is essential for upholding the core principles of natural justice.[12] These principles demand that decisions are to be made impartially, with fair notice and an opportunity to respond.[13] Explainability would therefore allow individuals to understand the rationale behind decisions and evaluate the system's legality and robustness. [14] Explainability is also especially important for legal professionals like lawyers and judges, whose actions influence not only individual clients but also public trust in the entire legal system.[15] This is because lawyers uphold the rule of law and must act transparently and responsibly to maintain confidence in legal processes.[16] Any breach of professional standards risks eroding public trust, triggering a chain reaction that could potentially destabilise both the legal profession and the justice system, and ultimately undermining the rule of law and societal confidence in the system.[17]

The Uncertainties in Defining 'Explainability' Standards

Despite these needs, there remains significant uncertainty about how we should define the standards for explainability due to the varied needs of different stakeholders. This is because different users require different forms of explanation tailored to their specific contexts.[18] As Doshi posits, an explanation must provide the correct type of information in

[12] Legal Services Board Podcast, 'Ethics, Technology and Regulation' (21 May 2020) <https://legalservicesboard.podbean.com/e/ethics-technology-and-regulationLinks to an external site.> accessed 14 December 2024.
[13] Committee for Privileges and Conduct, Third Report of Session 2017–19, Further report on the conduct of Lord Lester of Herne Hill, HL 252.
[14] Legal Services Board Podcast, 'Ethics, Technology and Regulation' (21 May 2020) <https://legalservicesboard.podbean.com/e/ethics-technology-and-regulationLinks to an external site.> accessed 14 December 2024.
[15] ibid.
[16] ibid.
[17] ibid.
[18] The Royal Society, 'Explainable AI: the basics' (November 2019) <https://royalsociety.org/newsresources/projects/explainable-ai/?utm_source=report&utm_medium=print&utm_campaign=aiinterpretability/> accessed 10 December 2024.

order for it to be useful.[19] A key principal for this is that they must enable the human observer to assess how much a particular input influenced the output.[20] In other words, explanations should present digestible information tailored to their specific audience. This shows that the explanations must be highly context-dependent, as it would only be meaningful if the AI-based decision can be explained at an individual level,[21] highlighting the complication in defining 'explainability' standards. Thus, for AI developers, the critical question is not merely whether a system is explainable or whether one model is more explainable than the other.[22] Instead, it is whether the system can deliver the specific type of explainability needed for a particular task or user group.[23]

For example, independent auditors or system developers would require the publishing of algorithms to know the technical and transparent details about how the system functions.[26] However, such explanations are not suitable for laypersons, who lack the expertise to interpret technical details.[24] Instead, counterfactual reasonings, which detail the specific factors leading to an individual output and explore the what if scenarios, what would need to change for AI to make a different outcome, are more helpful to help them understand and contest decisions.[28] Regulators, by contrast, might require explanations on data processing to ensure that the system operate within the bounds of established regulations.[25] This highlights that explainability cannot follow a "one-size-fits-all" approach,[26] which contributes to the

[19] Doshi-Velez and others, 'Accountability of AI Under the Law: The Role of Explanation' (2017) Berkmana Center Research Publication, 4.
[20] ibid.
[21] Hans de Brujin, Martijn Warnier and Mrijn Janssen, 'The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making (2022) 39 Government Information Quarterly 101666, 101669.
[22] The Royal Society (n 19) 14.
[23] ibid.
[26] ibid.
[24] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) < https://dx.doi.org/10.2139/ssrn.2660674> accessed 12 December 2024. [28] The Royal Society (n 19) 13-14.
[25] ibid 19.
[26] ibid.

challenge in defining its standards. As such, a user-centric mindset must be adopted.[27] This would consider the specific context and needs of each stakeholder to ensure that explanations are both relevant and meaningful.

Uncertainty in Achieving 'Explainability' Standards

Technical challenges

Achieving explainability in algorithmic decision-making fundamentally requires transparency, an ability to see the inner workings of the algorithm. However, this is hindered by the technical "black box" nature of AI systems.[28] This opacity arises from the complexity of machine learning models, particularly those involving techniques like deep learning.[29] This is because these systems do not follow pre-programmed rules. [30] Instead, they learn relationships and patterns through a layered structure, developing their own decisional rules.[31] These rules are typically unintelligible to humans, as they lack alignment with human concepts or symbolic reasoning. Therefore, they can even be too complicated for expert users[32] and programmers[33] to understand.

A proposed solution to mitigate is the use of post-hoc explanations[34] or proxy models[35] which analyses and interpret decision-making process after decisions are made. However, even

[27] Margaret Hagan, *Law by Design* (2021).
[28] Han-Wei Liu, Ching-Fu Lin and Yu-Jie Chen, 'Beyond *State v Loomis*: artificial intelligence, government algorithmization and accountability' (2019) 27 IJLIT 122, 135.
[29] ibid.
[30] ibid.
[31] ibid.
[32] The Royal Society (n 19) 8.
[33] Han-Wei Liu, Ching-Fu Lin and Yu-Jie Chen, 'Beyond *State v Loomis*: artificial intelligence, government algorithmization and accountability' (2019) 27 IJLIT 122, 135.
[34] Hans de Bruijn, Martijn Warnier and Marijn Janssen, 'The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making' (2022) 39(2) Government Information Quarterly <https://doi.org/10.1016/j.giq.2021.101666> accessed 19 December 2024.
[35] The Royal Society (n 19) 13.

such methods are not without challenges. One example is SHAP, a widely used method for estimating feature importance in AI systems.[36] However, research shows that SHAP's outputs can diverge from exact computations, especially when models involve more than a handful of features.[37] This highlights that SHAP is better at approximating relative importance, rather than providing precise explanations. Such limitations reinforce the Royal Society's concern that post-hoc explanations may not always be accurate or faithful and could even mislead users. [38] Similarly, Lipton[39] warned against the embracing interpretability that is optimised to satisfy subjective demands, since it can lead to explanations that are plausible but misleading. Illustrating this, Lipton drew parallels with human behaviour, where subjective rationales often conceal biases in processes like hiring and college admissions. To conclude, the technical black box remains a barrier to meaningful explainability, as alternative measures like post-hoc explanations may only provide partial transparency and risk misleading users.

Proprietary barriers

Balancing explainability with the protection of private interests presents another challenge in achieving explainability. This is also known as the legal "black box", [40] where proprietary protections such as trade secrets and intellectual property laws restrict access to the inner workings of AI algorithms.[41] This opacity often exists to maintain competitive advantage and stay ahead of adversaries. [42] The Council of Europe in European Ethical Charter

[36] XuanXiang Huang and Joao Marques-Silva, 'On the failings of Shapley values for explainability' (2024) 171 <https://doi.org/10.1016/j.ijar.2023.109112> 18 August 2025.
[37] Ibid.
[38] The Royal Society (n 19) 13.
[39] Zachary Lipton, 'The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.' (2018) 16(3) Queue 31, 42.
[40] Han-Wei Liu, Ching-Fu Lin and Yu-Jie Chen, 'Beyond *State v Loomis*: artificial intelligence, government algorithmization and accountability' (2019) 27 IJLIT 122, 135.
[41] ibid.
[42] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) < https://dx.doi.org/10.2139/ssrn.2660674> accessed 12 December 2024.

acknowledges this as a substantial barrier to achieving full transparency.[43] They posit that trade secret laws, in particular, limit access to the source codes of proprietary software, complicating efforts to ensure accountability and fairness.[44]

This obstacle is starkly illustrated in the *Loomis*[45] case, where the defendant contests the lack of transparency in the COMPAS risk assessment system. Despite these concerns, the Court upheld the proprietary nature of COMPAS, preventing disclosure of how risk factors were weighted or how risk scores were calculated. Similarly, Northpointe, the company behind the COMPAS, also refused to disclose its calculation methods to ProPublica, which criticised them for bias and inaccuracies in predicting recidivism among Black defendants.[46] A similar concern has been raised in the UK. The legal reform organisation JUSTICE, in its review of algorithmic tools in the justice system,[47] warned that biased and opaque systems such as COMPAS and Geolitica risk becoming an unlawful and arbitrary exercise of power. These case studies underscore the practical difficulties in meeting explainability standards, as corporate secrecy obstruct access to essential information necessary for true transparency and by extension, meaningful explainability.

[43] The European Commission for the Efficiency of Justice, 'European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment' (December 2018) <https://rm.coe.int/ethical-charteren-for-publication-4-december-2018/16808f699c> accessed 13 December 2024.
[44] ibid.
[45] *State v Loomis* 881 NW2d 749 (Wis 2016) 754 (US).
[46] Julia Angwin and others, 'Machine Bias' (*ProPublica*, 23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 3 December 2024.
[47] JUSTICE, 'AI in our Justice System' (2025) 40. <https://files.justice.org.uk/wp-content/uploads/2025/01/29201845/AI-in-our-Justice-System-final-report.pdf>

Ethical Dilemmas

Another obstacle in ensuring that context-based measures reliably uphold explainability standards is the ethical challenge of balancing the conflicting demands of completeness and interpretability. Interpretability seeks to make AI systems comprehensible to humans by simplifying their operations into concepts or terms that align with human knowledge. [48] Completeness, by contrast, aims to offer a full and accurate depiction of a system's functionalities.[49] As Gilpin posits, attaining both at the same time is fundamentally challenging because the most accurate explanations are difficult for people to understand; and conversely, the most interpretable descriptions often lack predictive power. [50] The most accurate explanations are often too complex to understand, while the most interpretable ones risks producing persuasive rather than transparent explanations.[51] As Herman points out, this tension creates grave ethical dilemmas and poses two questions: "*When is it unethical to manipulate an explanation to better persuade users? And how do we balance concerns of transparency and ethics with our desirability for interpretability?*"[56] These questions highlight the dangers of oversimplified explanations that prioritise persuasion over credibility, which can lead to deceptive explanations that erode trust and accountability. To conclude, the balancing of completeness and interpretability presents a challenge in achieving useful explainability.

[48] Leilani Gilpin and others, 'Explaining Explanations: An Overview of Interpretability of Machine Learning' (2018) Massachusetts Institute of Technology Cambridge.
[49] ibid.
[50] ibid.
[51] ibid.
[56] ibid.

Furthermore, the possibility of misguiding users is another ethical dilemma. Wachter[52] notes that counterfactual explanations can reveal when an algorithmic decision was influenced by protected characteristics such as gender or race, exposing potential discrimination. However, these explanations also carry inherent limitations.[53] Defending his stance, Wachter explains that counterfactual reasonings merely describe dependencies between a decision and external factors, but not necessarily causal links.[54] This means that while a counterfactual explanation might show that altering a certain factor leads to a different conclusion, it does not disclose the underlying causal connection or the algorithm's internal logic. This limitation arises because counterfactual explanations only lay out specific dependencies, and not the full set of circumstances influencing the decision. Wachter illustrates this by showing that, for example, if a counterfactual reasoning demonstrates that a "Black" individual's race influenced a decision, it does not imply that the race of "White" individuals was treated as irrelevant.[55] Thus, this could lead to a superficial understanding of why a decision was made and therefore may mislead user's intuitions. In conclusion, this reveals the limits of counterfactual explanations as tools for achieving explainability as it only provides partial insights into the algorithm's behaviour.

The Way Forward

Addressing these challenges of explainability in algorithmic decision-making is a complex and ongoing issue. However, a promising temporary solution lies in fostering inclusivity in the

[52] Sandra Wachter, Brent Mittelstadt and Chris Russel, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31(2) Harvard Journal of Law & Technology 841, 853854.
[53] Sandra Wachter, Brent Mittelstadt and Chris Russel, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31(2) Harvard Journal of Law & Technology 841, 853854.
[54] ibid.
[55] ibid.

process of designing these tools. Both Turner[56] and Al Olama[62] highlight the importance of involving diverse voices from governments and citizens through public consultations and online questionnaires, ensuring AI's societal impact is shaped collectively and not controlled by a technocratic elite. By involving individuals in these processes, trust in these systems can be cultivated, leading to greater confidence of their use. Building on this, decision-makers should also bear the responsibility of explaining AI-influenced outcomes, which would incentivise greater scrutiny and critical evaluation of these tools.[63] Moreover, given the current limitations in achieving full explainability, implementing an "opt-in or optout" mechanism offers a practical interim solution which allows users the autonomy to engage with such systems.[57] This approach also aligns with the European Ethical Charter, which emphasises the importance of informing in automated decision-making processes.[58] Together, these approaches can foster public trust in algorithmic decision-making and lay the groundwork for more sustainable and ethical use.[59]

Conclusion

In conclusion, the integration of algorithms into the justice system presents significant opportunities for enhanced efficiency, particularly in addressing the rising backlog cases. However, the above discussion underscores the critical importance of explainability in these

[56] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan Cham 2018) 264-272. [62] Dom Galeon, 'An Inside Look at the First Nation With a State Minister for Artificial Intelligence' (*Futurism*, 12 November 2017) <https://futurism.com/uae-minister-artificial-intelligence> accessed 19 December 2024.
[63] Hans de Bruijn, Martijn Warnier and Marijn Janssen, 'The perils and pitfalls of explainable AI: Strategies for
explaining algorithmic decision-making' (2022) 39(2) Government Information Quarterly <https://doi.org/10.1016/j.giq.2021.101666> accessed 19 December 2024.
[57] Han-Wei Liu, Ching-Fu Lin and Yu-Jie Chen, 'Beyond *State v Loomis*: artificial intelligence, government algorithmization and accountability' (2019) 27(2) IJLIT 122, 140.
[58] The European Commission for the Efficiency of Justice, 'European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment' (December 2018) <https://rm.coe.int/ethical-charteren-for-publication-4-december-2018/16808f699c> accessed 13 December 2024.
[59] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan Cham 2018) 278.

systems as highlighted by ethical and policy discussions. The challenges of defining and achieving explainability standards are multifaceted, involving technical, proprietary and ethical dilemmas. Addressing these issues requires a nuanced approach that includes diverse stakeholder involvement and mechanisms for user engagement. By tackling these challenges, the justice system can move towards a more transparent and ethical use of algorithms, fostering public trust and upholding the principles of justice.