frontiers | Frontiers in Political Science

TYPE Policy and Practice Reviews
PUBLISHED 20 October 2025
DOI 10.3389/fpos.2025.1605619

OPEN ACCESS

EDITED BY
Ben Wagner,
Delft University of Technology, Netherlands

REVIEWED BY
Omobolaji Olufunmilayo Olateju,
University of Ibadan Research Foundation, Nigeria
Marcel Robeer,
Utrecht University, Netherlands

*CORRESPONDENCE
Maximilian Zocholl
✉ maximilian.zocholl@europol.europa.eu

RECEIVED 03 April 2025
ACCEPTED 24 September 2025
PUBLISHED 20 October 2025

CITATION
Zocholl M, Stampouli D, Wittfoth M and Mounier G (2025) Fundamental considerations for the use of explainable AI in law enforcement.
Front. Polit. Sci. 7:1605619.
doi: 10.3389/fpos.2025.1605619

Fundamental considerations for the use of explainable AI in law enforcement

Maximilian Zocholl*, Dafni Stampouli, Mark Wittfoth and Gregory Mounier

Innovation Lab, Europol, Den Haag, Netherlands

Explainable AI (XAI) methods have the potential to make the use of AI in law enforcement more understandable, and ultimately more trustworthy. We argue that explanation requirements differ strongly between use cases and between stakeholders ranging from law enforcement officers to affected persons. While no currently known XAI method provides a guarantee to fully reflect the functioning of an AI model, XAI methods are currently the most promising means to bridge the gap between human and AI after increasing the human's AI literacy. Even though the benefits of XAI vary strongly with the accuracy of the AI system and need to be balanced against incurring risks, like automation bias, we argue that not using XAI implies larger risks than exploring the technologies' benefits and further developing it. In order to overcome existing shortcomings, we advocate for more collaborations between law enforcement agencies, academia, and industry.

KEYWORDS

explainable AI, XAI, law enforcement, transparency, trustworthy AI

1 Introduction

AI systems are becoming increasing prevalent and are now integral to many aspects of people's daily lives. Given the growing complexities law enforcement authorities are faced with in the context of large and complex data sets, AI is critical in the context of criminal investigations, too. In order to make the use of AI in the domain of law enforcement as transparent as possible, it is necessary for AI experts, law enforcement officers, lawyers, judges, and affected persons to understand what is happening within the AI system. Explainable AI (XAI) is one way to come closer to this insight. With the range and diversity of stakeholders involved in a specific use case, also the range of using an understanding of the AI systems functioning varies strongly. Explainable AI can be used to assess the AI system's reliability, and robustness, to improve the system's performance, and to enable affected persons to appeal an AI-based decision. Ultimately, explanations can improve public trust in the law enforcement.

Explanations of a single AI system or its outputs can be as varied as the number of existing XAI methods, and as diverse as the different information needs of those seeking explanations. Currently, there is no universal solution for explaining an AI model—let alone all AI models—and delivering satisfying explanations at the push of a button remains elusive. Each explainability method reveals specific details about the model's inner workings. Often, these methods are used together in an iterative process to provide a multi-dimensional view of the AI system.

Gaining a deeper understanding of an AI system is typically seen as a positive development. However, there are also risks related to the use of XAI (Carli et al., 2022). XAI can inadvertently foster excessive trust, where users blindly rely on an AI system's output simply because it has been explained. Other risks include adversarial attacks or the potential for intellectual property theft, where explanations could be misused. As a result, explanations must be tailored to their

specific context and carefully evaluated—there is no guarantee they will always yield a benefit.

However, the risks of inaction may be even greater. AI models are becoming increasingly sophisticated, and their outputs are informing more and more decisions. With explainability often lagging behind these advancements, an increasing number of decisions are being made based on unexplained AI system outputs. Even when decisions are not directly made by AI systems, their suggestions can have a significant impact on human decision-makers. In the case where human input is minimal, these AI-generated suggestions are legally treated as automated decision-making, as confirmed by the European Court of Justice's ruling C-634-21. Under the AI Act, individuals impacted by decisions from high-risk systems are entitled to explanations, and those providing human oversight must be able to accurately interpret the AI system's output. Failing to comply with these legal obligations jeopardises public trust in law enforcement and wastes a valuable opportunity to embrace technological change responsibly and in a commendable way.

The best way to mitigate these risks is through collectively strengthening law enforcement's initiatives on explainable AI systems across Europe. In the following we are focussing on supporting efforts towards XAI in law enforcement by addressing key questions about explainability.

2 Bridging the gap between AI and law enforcement

Explainable AI (XAI) aims to bridge the gap between humans and AI. To make this goal more achievable, it is essential to minimise this gap from the start. While transparency in AI systems helps narrow the gap on the AI side, educating humans about AI is just as important to close the gap from the human side. Law enforcement personnel should be trained to understand how AI systems work, increasing their AI literacy (Leslie et al., 2024), including the different models and components involved, and how these elements interact. It is also crucial for humans to grasp the difference between correlation— a principle on which most AI models are based—and causations, as this distinction is key to understanding how AI models function. Furthermore, users need to be aware of the limitations of AI systems, which can vary across different applications. Understanding the risks associated with AI, such as automation bias and anti-automation bias, is essential (Europol, 2025). With AI systems' natural language capabilities approaching or even surpass human levels, it becomes even more important to remind users that they are interacting with an AI system, and importantly, that machines cannot be held accountable for their decisions. This responsibility stays with the human (Doshi-Velez et al., 2017).

2.1 What is explainable AI (XAI)?

XAI is an active field of research, as well as an umbrella term for a large range of methods that aim at making AI models and systems understandable to humans. At least four types of methods can be distinguished (Barredo Arrieta et al., 2020). First, interpretable models are models which can be understood by humans as their level of complexity is low and the functional relationship between output and input is transparent. While interpretable models, also known as ante-hoc XAI methods, can be understood with a sufficient level of AI literacy, they lack accuracy when performing complex tasks, as discussed in Section 2.5. Second, input perturbation methods are based on the concept of changing parts of the AI system's input repeatedly in an experimental fashion, observing variations in the AI system's output and inferring more and less relevant parts of the input, based on the changes of the AI system's output. While these methods come with the advantage of being applicable to all types of AI systems, they are computationally expensive. Third, output back-propagation methods use the AI system's intermediate results to trace its output back to the input. These methods are computationally cheap but rely on the access to the AI system's internal parameters. Input perturbation and output back-propagation methods are also referred to as post-hoc XAI methods. Fourth, documentation-based methods are based on the idea that an AI system can explained on a general level by information about its training data, the evaluations performed on it, its intended purpose, etc. While offering fundamental information about the AI system, document-based methods do not cover individual decisions of an AI system. With all methods having different strengths and weaknesses, there is no 'best' method to choose for every use case. Using different XAI methods does not only come with the advantage of combining the strength of these methods, but allow to shed light on the AI system's interna from different angles, increasing the overall understanding of the AI system's functioning.

2.2 Why do we need explanations in law enforcement?

Today's AI models often contain billions of parameters and learn complex patterns from vast datasets, with little human guidance during critical stages of the training process. While this enables the automation of learning, it also introduces the risk that AI models may pick up irrelevant or non-meaningful patterns. This can lead to highly accurate results for certain datasets, but poor performance for others. For instance, a classifier might learn to distinguish deepfakes from real images based on image size and format rather than actual content, simply because the training dataset contained deepfakes with different sizes and formats than the real images, an example of the so-called Clever Hans effect. This risk can be reduced by closely examining the training dataset, but biases within the data or discrepancies between training and operational data may be subtle and undetectable to humans. Even when the training and testing datasets are representative of the intended use case, this does not guarantee that the AI model has actually learned the desired concepts as intended by developers, users, or other stakeholders. This highlights the importance of understanding what the AI model has learned during training and why it produces specific outputs during deployment. The following examples illustrate three key benefits of XAI methods.

- Explainability can be used to identify root causes for low accuracy of an AI system, both for False Positive (FP), as depicted in Figure 1, and False Negative (FN) results.
- Explainability can be used to increase AI systems' robustness for True Positive (TP) and True Negative (TN) results. Non-meaningful representations learned during training can be identified more easily, the training dataset can be improved, and the AI model can be retrained, as exemplified in Figures 2A,B.

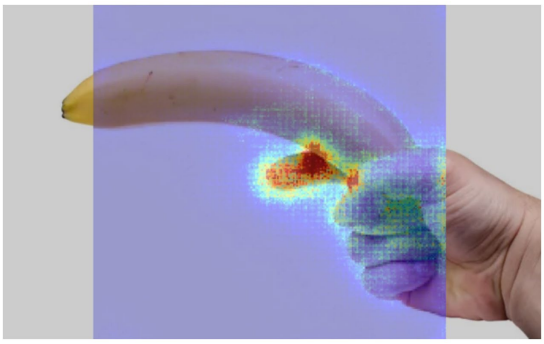

FIGURE 1
XAI (VarGRAD) highlights parts of the image that can be used to investigate why a banana was misclassified as handgun, an example for a FP.

- Explainability can be used to identify biases learned from datasets (Lapuschkin et al., 2019; Selvaraju et al., 2017). By looking at one explanation, biases within the training dataset that may not have been found when inspecting all examples of the dataset individually can become obvious to the human eye.

Explaining an AI system's outputs may not always come with significant benefits, but the benefits arguably increase with the risk exhibited by the AI system's output, cf. (Matulionyte and Hanif, 2021). This idea is reflected by more demanding explainability requirements for high-risk use cases in the AI Act. In particular, Article 86(Carli et al., 2022) asks for "(…) clear and meaningful explanations of (…) the main elements of the decision taken (…)" in order for a person affected by the decision of a high-risk system to be able to exercise their right to challenge this decision. Additionally, Article 14(4)(c) requires the natural person performing human oversight over such an AI system to be enabled to interpret the AI system's output correctly, e.g., with the help of "(…) interpretation tools and methods available." As the approach of the AI Act is purely risk-based, explainability requirements are not limited to a specific subset of law enforcement tasks, but may be applicable to prevention, investigation, detection and prosecution of criminal offences equally, e.g., to explain the risk assessment of a natural person becoming the victim of domestic violence, to explain the result of an emotion recognition system detecting physical violence, or to explain the evaluation of the reliability of forensic evidence in the investigation and the prosecution of a drug trafficking case.

2.3 What is a good explanation?

An explanation provides details or reasons to clarify something and make it easier to understand, cf. (C. U. P. & Assessment, 2024). In the context of AI, an explanation should help clarify the output of an AI system and may also encompass the human decision-making process. It can involve either partially or fully automated methods and is intended for a clearly defined group of stakeholders (High-Level Expert Group on Artificial Intelligence (HLEG), 2019; Gyevnar et al., 2023; Panigutti et al., 2023). An explanation could refer to a specific input and how it is transformed into a particular output by the AI model. It may also describe the general function of an AI model or a class of similar examples and their outputs. An explanation is considered effective if it meets the following criteria:

- Simple (Selten et al., 2023): An explanation should be easy to understand, ideally more so than the AI model itself. What is considered simple depends on the recipient's level of AI literacy. However, over-simplification should be avoided, cf. (Bernardo, 2023).
- Contrastive (Selten et al., 2023; Atkinson et al., 2020): Explanations can be most effective when they present contrasting examples of AI model inputs and outputs. This helps users better grasp the underlying logic by highlighting significant differences along decision boundaries.
- Selective (Atkinson et al., 2020): An explanation should focus on the information needed to understand a particular aspect of the AI model, such as a specific decision, a representative group of decisions, or the general logic behind the system.
- Explicit (Selten et al., 2023): An explanation should be comprehensive, including all relevant information and avoiding references to details that may not be accessible or understandable to the recipient.
- Deterministic (Atkinson et al., 2020): Although many AI models are based on statistical or probabilistic methods and may exhibit randomness, good explanations should provide consistent, repeatable results to aid human understanding, as people often struggle with uncertainty in systems.
- Social (High-Level Expert Group on Artificial Intelligence (HLEG), 2019; Selten et al., 2023; Atkinson et al., 2020): Explanations should consider the social context of the recipient, including their level of expertise, language, and how the information is presented, e.g., visually, acoustically, textually. This makes the explanation more effective and tailored to the individual.
- Meaningful (Gyevnar et al., 2023; Phillips et al., 2021): An explanation should help the recipient interpret and understand the AI model's output correctly. This requires the explanation to accurately reflect the model's reasoning while remaining understandable to the user.
- Accurate (Gyevnar et al., 2023; Phillips et al., 2021): An explanation should accurately represent the inner workings of the AI system to a predefined level of precision.
- Causal (Atkinson et al., 2020; Olsen et al., 2019): A good explanation should include the reasoning behind why a specific output was generated by the AI system, if possible.
- Timely (High-Level Expert Group on Artificial Intelligence (HLEG), 2019): Explanations should be provided to the recipient within a reasonable time frame after a request is made.
- Actionable (Hacker and Passoth, 2022): An explanation should empower the recipient to make decisions or take actions, such as accepting or rejecting the AI model's output.

Depending on the use case, and the explanation recipient, the same AI model may require different XAI methods to address different explanation needs. While both affected persons and law enforcement personnel may require explanations on the level of individual AI model output, both stakeholder groups have very different background knowledge which may require different information included in the explanation as well as different representations of the explanation. To create a good explanation, a multi-disciplinary approach is recommended to address the diverse information needs of all

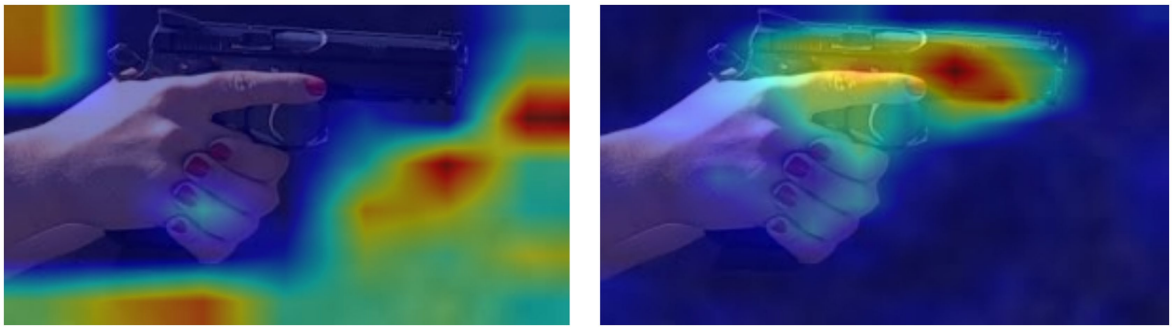

FIGURE 2
Making AI models more robust for "pistol" classification: old AI model with non-meaningful representation **(A)** and new AI model **(B)**. Both explanations were generated with gScoreCAM.

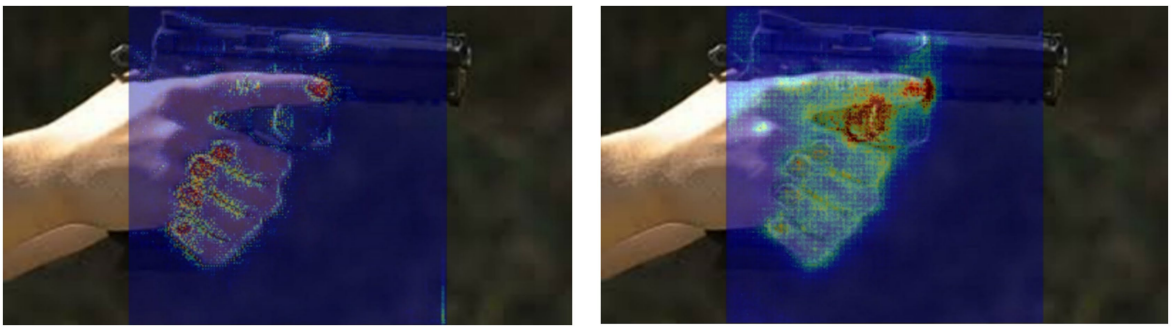

FIGURE 3
Results of output-back-propagation algorithms. Guided back-propagation **(A)** and SquareGRAD **(B)**.

stakeholders. It is important to note that an explanation can still be accurate even if it highlights the AI model's limitations. In such cases, the explanation may help identify flaws in the AI model's behaviour, unlocking its potential to improve the system.

However, explanations are not always inherently beneficial and can carry certain risks when provided, e.g., correct sounding verbal explanations or impressive visual explanations may entice a person performing human oversight to accept the explanation without challenging them even though they may not be correct, resulting in automation bias. Additionally, explanations that simply confirm the expectations of the explanation recipient may be accepted more easily than contrarian explanations and can lead to confirmation bias. If explanations fail to address biases like automation bias or confirmation bias, they may offer little value or, worse, create a false sense of trust in the AI system's decisions. Important to keep in mind is that the output of an XAI system is influenced both by the AI model, and the explanation method. If an explanation reveals that meaningless features are being used by the model, this could point to an issue with the AI model itself, the XAI method, or a combination of both.

To reduce the risk of poor explanations, it is crucial to keep the recipient engaged (Leslie et al., 2024). This can be achieved in various ways, such as by presenting multiple system outputs that require the user to think critically and piece together the explanation. For example, offering independent sub-explanations that must be combined to form a complete explanation can encourage deeper reflection. Additionally, gathering feedback from recipients can help tailor explanations to their specific information needs.

2.4 Is there a single explanation?

The output of an AI model can be explained using various XAI methods. Figures 3, 4 illustrate the results for the same AI model, the same AI model output, and the same image but four different XAI methods. Warm colours in the heatmap highlight areas with high relevance for the AI model's output while cold colours indicate areas with little or no relevance. Figure 3 depicts the results of output-back-propagation algorithms, in particular of guided back-propagation (A), and SquareGRAD (B), while Figure 4 depicts the results of two input-perturbation algorithms Rise (A) and Occlusion (B).

Figure 3 presents the outcomes of model-specific XAI methods, which use the model's parameters to trace the path from the classification back through the AI model to the input features.

Figure 4 shows the results of model-agnostic XAI methods, where the AI model is repeatedly fed modified versions of the input image to observe how changes in the image affect the classification. The differences in the explanations across the four images are significant. Not only can clear differences be observed between model-specific and model-agnostic methods, but there are also noticeable variations between different methods within the same XAI method category. This anecdotal finding is an example for the Rashomon Effect in XAI which states that there are multiple explanations for the same AI model, its input, and its output. This effect frequently results in the disagreement of XAI methods, which can be quantified with different metrics, and which constitutes the major obstacle to the practical use of XAI methods (Müller et al.,

2023). While there is no theoretical solution to the disagreement problem in sight, practitioners cope with the problem by using different XAI methods, or by favouring one XAI method based on its mathematical properties or personal preference (Krishna et al., 2022).

2.5 Does more interpretability mean less performance?

Often, AI accuracy and interpretability are seen as competing requirements. Figure 5 depicts qualitatively different types of AI models according to their performance, e.g., in terms of accuracy, and their interpretability. A higher level of interpretability may or may not translate into a lower level of model performance depending on the task to be solved.

As an example, let us consider two AI models M1 and M2 of different type, for example a Neural Networks (NNs) M1, and a Decision tree M2. For simpler tasks, both models may deliver predictions with similar accuracy, with Decision Trees offering a higher level of interpretability. In such cases, increasing interpretability may not come at the expense of performance.

Let us now consider a more complex task, such as an image generation tasks or a classification task involving noisy, high-dimensional data. These tasks can be performed by NNs but not with Decision trees. In these instances, opting for a more interpretable model might not yield satisfactory results. Therefore, interpretability

FIGURE 4
Results of input-perturbation algorithms. Rise **(A)** and Occlusion **(B)**.

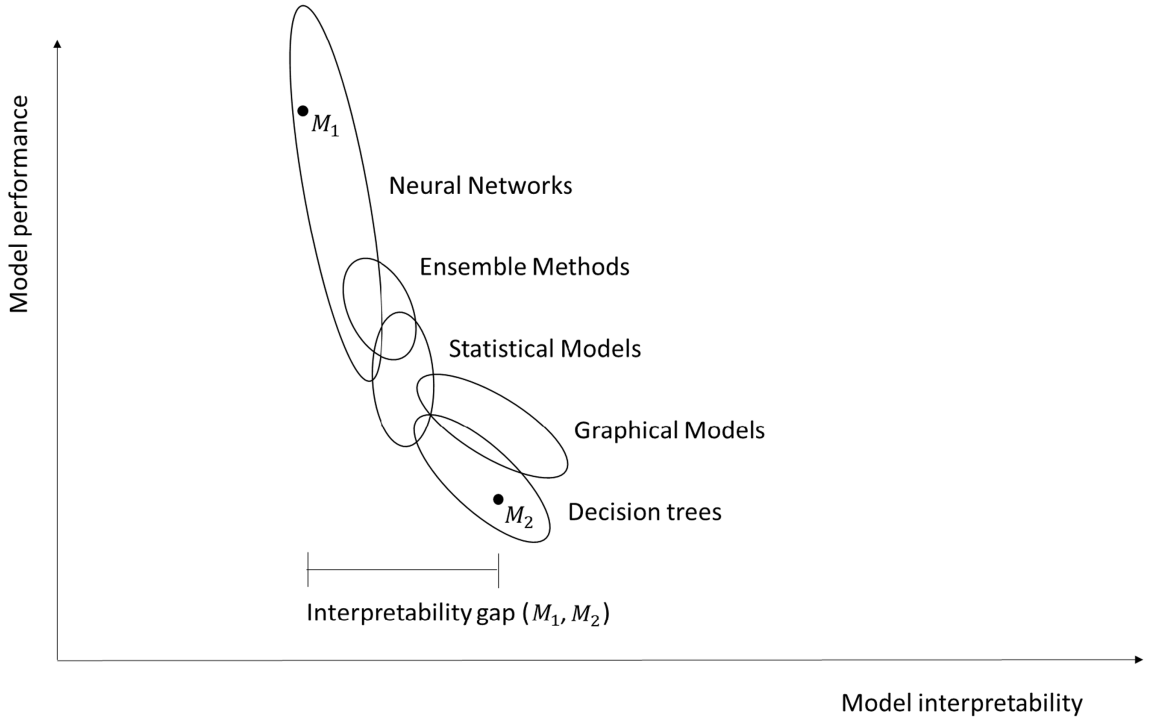

FIGURE 5
Qualitative trade-off: performance vs. interpretability, cf. (Gunning and Aha, 2019, Figure 1).

and performance do not always have to be mutually exclusive, as lower-performing models may not be viable alternatives.

However, in scenarios where different models can achieve the same task and the dataset is complex enough to push the limits of a more interpretable model's representational capacity, a trade-off between interpretability and performance can occur. This trade-off involves weighing the potential reduction in risk from increased interpretability against the possible increase in risk due to a decrease in accuracy (Hacker and Passoth, 2022).

As the complexity of the dataset grows, an interpretability gap between two models, M1 and M2, may become noticeable between ante hoc and non-ante hoc explainable AI models. At this point, post-hoc XAI methods become valuable, as they can help reduce or even close the interpretability gap. While the trade-off between performance and interpretability is well-documented in the literature, introducing explainability adds additional items to be considered, such as increased computational costs, licensing fees, intellectual property concerns, and trade secrets (Hacker and Passoth, 2022). Balancing benefits and costs, opportunities and risks of XAI needs to be done on a case-by-case basis.

2.6 What is the benefit of an explanation?

Understanding an AI system or its results does not always provide the same benefit. When balancing performance and explainability in AI systems, the value of understanding the AI model and the quality of its outcomes need to be quantified in the context of a specific use case. Take the case as an example where a fugitive is stopped for an in-person identity verification based on the output of a remote biometric identification system (TP).

The fugitive who knows they are wanted, may find an explanation for an identity verification to be of little value, whereas an individual who is not wanted may find it more useful to understand why their identity was deemed to be important to be verified (FP). In contrast, if the fugitive is not detected by the AI system (FN), he is unlikely to request an explanation for why the AI did not prompt an identity verification. Similarly, someone who has not committed any wrongdoing and was not flagged by the AI system (TN) would likely not ask for an explanation either. This imbalance in the benefits of explanations, shown in Table 1, highlights that AI systems' explanations are particularly valuable in FP scenarios. This simplified comparison suggests that more accurate AI systems reduce the need for explanations, thereby increasing the overall benefit.

TABLE 1 On the dependency of benefits of explanations on the correctness of AI system's output.

Affected person	True AI system output	False AI system output
Non-wanted person	No identity verification: No benefit from an explanation (TN)	Identity verification: High perceived benefit from an explanation (FP)
Wanted person	Identity verification: Low benefit from an explanation (TP)	No Identity verification: No benefit from an explanation (FN)

2.7 Is explainability a guarantee for improving human-AI team performance?

A collaborative decision-making process that incorporates both human and AI input forms part of a complex socio-technical system. Introducing XAI enables humans to understand aspects of the AI system that were previously opaque. This interaction between humans and AI has the potential to deliver performance that surpasses both human-only and AI-only decision-making. However, XAI is not a cure-all for the challenges in human-AI collaboration. Two key biases that can emerge from this interaction are confirmation bias and automation bias. In law enforcement, confirmation bias occurs when explanations are only accepted if they align with the recipient's pre-existing beliefs (Europol, 2025; Selten et al., 2023). Explanations that contradict the recipient's assumptions are often ignored or rejected, even when they are correct (Selten et al., 2023). For situations where explanation recipients do not have preconceived notions it can be observed that misinterpretations of explanations may result in decisions that contradict both the AI model's output and its explanation (Herrewijnen et al., 2024). On a technical level it is therefore recommended to keep explanations for law enforcement as simple as possible, to use natural language explanations where possible, to use a limited amount of numbers and words, to use only one numerical metric, to ensure the correct interpretation of explanations by case specific studies, and to flank the use of XAI with AI and XAI training for law enforcement personnel (Herrewijnen et al., 2024). Additional actions relying on strategic backing includes in-house development to ensure the use of domain-specific language of explanations, the creation of interdisciplinary teams of stakeholders for the development and testing of XAI methods, as well as a reinforcement of collaboration with academia, industry, and other law enforcement agencies in the areas of research, standardisation, and auditing (Walke et al., 2023).

3 Conclusion

To achieve transparent and trustworthy AI-driven decision-making processes in law enforcement, explainable AI (XAI) holds significant potential in addressing the interpretability challenges of quickly evolving AI systems.

To bring explainability closer to law enforcement experts and decision-makers, who are increasingly required not just to work with, but also to explain AI systems, this article addresses common questions related to XAI methods in the law enforcement context. It presents criteria for good explanations, discusses the benefits and risks of XAI methods, and argues that greater accuracy does not necessarily lead to reduced transparency in AI systems. Importantly, the article stresses that there are multiple explanations for a single AI model output. This implies that more research is needed to ensure explanations that reflect the AI system's functioning.

However, XAI methods provide crucial insights into understanding the inner workings of AI models and their individual outputs—an essential aspect for law enforcement use. Given the heightened risks to fundamental rights, greater transparency and mindfulness are needed to uphold and strengthen public trust. Noting that no specific technical requirements are outlined by AI Act or the

Law Enforcement Directive, high-risk AI systems cannot be deployed without additional transparency efforts, both at the model level and in individual decision-making.

While XAI methods are particularly valuable for increasing transparency in high-performing but non-interpretable AI models, their application does not need to be limited to these models. XAI methods are a valuable tool for evaluating AI systems beyond traditional metrics like recall or precision, unlocking the potential for more robust and meaningful performance in increasingly complex tasks.

At present, XAI methods are the best way to bridge the gap between humans and AI, enhancing the efficiency of human-AI collaboration. However, this gap should be minimized from the outset by investing in AI literacy and by enhancing law enforcement officers' competency to engage with academia, industry, and other stakeholders in the internal security domain. This will help ensure that technical progress aligns with the needs of law enforcement. As AI technologies continue to evolve rapidly, law enforcement must accelerate the customization and adoption of existing technologies to stay ahead of criminal activity. Only by investing in emerging and disruptive technologies tailored for law enforcement needs can we ensure that both accuracy and transparency requirements, are met, ultimately contributing to make Europe safer.

Author contributions

MZ: Writing – original draft, Writing – review & editing. DS: Writing – review & editing. MW: Writing – review & editing. GM: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Atkinson, K., Bench-Capon, T., and Bollegala, D. (2020). Explanation in AI and law: past, present and future. *Artif. Intell.* 289:103387. doi: 10.1016/j.artint.2020.103387

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Bernardo, V. (2023). TechDispatch on explainable artificial intelligence: EDPS.

C. U. P. & Assessment, (2024). Cambridge Dictionary.

Carli, R., Najjar, A., and Calvaresi, D., (2022). "Risk and exposure of XAI in persuasion and argumentation: the case of manipulation," in *International workshop on explainable transparent autonomous agents and multi-agent systems*, Springer International Publishing.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., et al. (2017). Accountability of AI under the law: The role of explanation.

Europol (2025). AI Bias Report.

Gunning, D., and Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* 40, 44–58.

Gyevnar, B., Ferguson, N., and Schafer, B., (2023). "Briding the transparency gap: what can explainable AI learn from the AI act?" in *European conference on artificial intelligence, Krakow, Poland*.

Hacker, P., and Passoth, J.-H. (2022). "Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond" in xxAI - beyond explainable AI (Vienna, Austria: Springer), 343–374.

Herrewijnen, E, Loerakker, MB, Vredenborg, M, and Woźniak, PW (2024). "Requirements and attitudes towards explainable ai in law enforcement", in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*.

High-Level Expert Group on Artificial Intelligence (HLEG), (2019). Ethics guidelines for trustworthy AI.

Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., and Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: a practioner's perspective. *Trans. Mach. Learn. Res.* arXiv preprint arXiv:2202.01602.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever Hans predictors and assessing what machines really learn. *Nat. Commun.* 10:1096.

Leslie, D., Rincón, C., Briggs, M., Perini, A. M., Jayadeva, S., Borda, A., et al. (2024). AI Explainability in practice. London, United Kingdom: The Alan Turing Institute.

Matulionyte, R., and Hanif, A., (2021). "A call for more explainable AI in law enforcement," in *IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*.

Müller, S., Toborek, V., Beckh, K., Jakobs, M., Bauckhage, C., and Welke, P. (2023). "An empirical evaluation of the Rashomon effect in explainable machine learning" in Joint European conference on machine learning and knowledge discovery in databases (Switzerland: Springer Nature), 462–478.

Olsen, H. Palmer, Slosser, J. Livingston, Hildebrandt, T. Troels, and Wiesener, C., (2019). What's in the box? The legal requirement of explainability in computationally aided decision-making in public administration, iCourts Working Paper Series, 162.

Panigutti, C., Hamon, R., Hupont, I., Llorca, D. F., Yela, D. F., Junklewitz, H., et al. (2023). "The role of explainable AI in the context of the AI act," in *ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA*.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., et al. (2021). Four principles of explainable artificial intelligence. Gaithersburg, United States: NIST.

Selten, F., Robeer, M., and Grimmelikhuijsen, S. (2023). 'Just like I thought': street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Adm. Rev.* 83, 263–278. doi: 10.1111/puar.13602

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., (2017). "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *In Proceedings of the IEEE international conference on computer vision*.

Walke, F., Bennek, L., and Winkler, T. J. (2023). "AI in government: a study on Explainability of high-risk AI-Systems in law enforcement and police service" in International conference on Wirtschaftsinformatik (Switzerland: Springer Nature), 393–407.