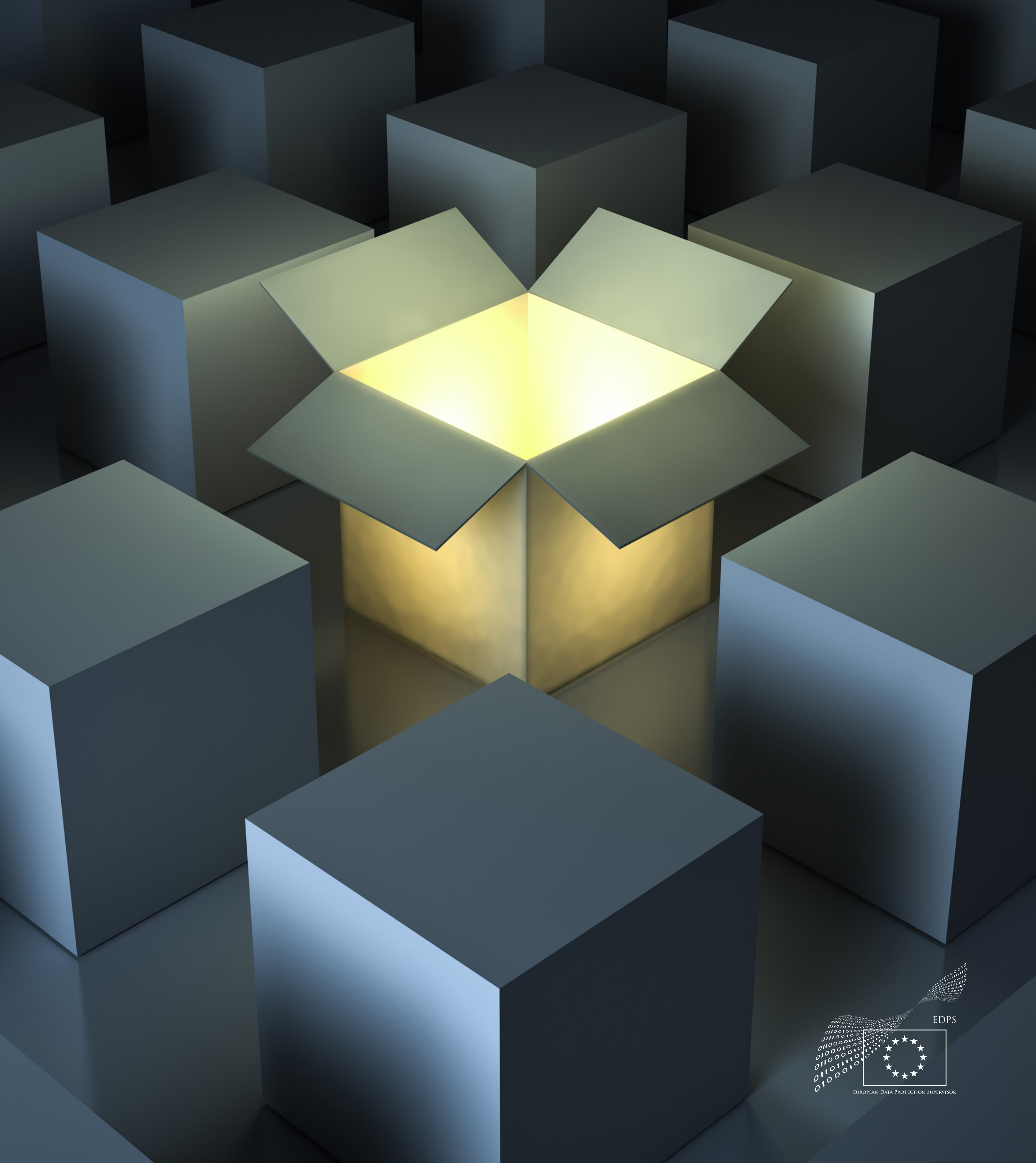
TechDispatch

EXPLAINABLE ARTIFICIAL INTELLIGENCE



EDPS

European Data Protection Supervisor

HTML	ISBN 978-92-9242-715-3	ISSN 2599-932X	doi: 10.2804/132319	QT-AD-23-002-EN-Q
PDF	ISBN 978-92-9242-716-0	ISSN 2599-932X	doi: 10.2804/802043	QT-AD-23-002-EN-N

The “black box” effect

The adoption of artificial intelligence (AI) is rapidly growing in sectors such as healthcare, finance, transportation, manufacturing and entertainment. Its increasing popularity in recent years is largely due to its ability to automate tasks, such as processing large amounts of information or identifying patterns, and its widespread availability to the public.[1] Large language models[2] (LLMs), like *ChatGPT*, or text-to-image models,[3] like Stable *Diffusion*, are two examples of AI that have gained large popularity in recent years.[4]

However, despite the growing use of AI, many of these systems operate in ways that are opaque to both those providing AI systems (‘providers’), those deploying AI systems (‘deployers’), and those affected by the use of AI systems. In the complex realm of AI systems, even the providers of these systems are often unable to explain the decisions and outcomes of the systems they have built.

This phenomenon is commonly referred to as the **“black box” effect**.

1. The risks of opaque AI systems

AI systems such as machine learning (ML) or deep learning (DL) use algorithms learned by their own process of training,[5] rather than by explicit human programming.

During the process of training, AI models can discover new correlations between certain input features (e.g., clinical symptoms) and can make decisions or predictions (e.g., medical diagnoses) based on highly complex models involving a large number of interacting parameters (possibly millions), making it difficult even for AI experts to understand how their outputs are subsequently produced (Peters, 2023).

In these situations, reasons why systems have made certain decisions may be unclear, both to the users of the systems and to those affected by the systems. The resulting "black box" effect could lead to either misplaced trust or over-reliance on AI systems, both of which could have negative consequences for individuals.

It can be argued that, in today's society, users do not need to understand how a particular technology works in order to use it, and often do not fully comprehend the way certain technologies work. For example, how many drivers can actually describe how an automatic transmission works? To this end, one should bear in mind that AI technology is often implemented for automated decision-making (or decision support), including by public authorities. In this context, transparency and accountability are, in most cases, essential legal requirements.

It is therefore unacceptable to have a "black box" effect that hides the underlying logic of decisions made by AI.

Another difference, this time from a technological point of view, taking into account the car's automatic transmission, sometimes, unlike the case of the car's functioning, AI engineers themselves may not have a full understanding of what is going on beneath the surface".

This opacity not only makes decisions more difficult to understand, but it can also have direct impact on individuals since it can hide deficiencies in AI systems, such as the existence of bias,[6] inaccuracies, or so-called "hallucinations".[7]

Poorly designed, developed or tested algorithms can produce results that are potentially discriminatory or harmful to individuals.

For instance, when AI is used to select job applicants, systems might inadvertently favour candidates from certain demographics or backgrounds due to biased training data. If the system is a "black box", it could be difficult to understand why certain candidates have been rejected or selected, making it harder to identify and address bias.[8]

Another example is the use of AI models for medical diagnosis that may disproportionately misdiagnose or miss certain conditions for certain demographic groups due to biased training data. When the model is a "black box", it becomes difficult for healthcare professionals to understand the reasoning behind the decisions, hindering their ability to address potential biases.[9]

Discriminatory outcomes are not the only problem with "black boxes". The lack of transparency itself may hinder the ability of those affected by automated decisions to understand the underlying logic and its potential impact. This could be the case, for example, with AI models used for credit approval, where bank customers might have no insight into the automated decisions that affect their financial lives.[10]

More importantly, individuals may be affected by automated decision-making systems used by governments, the operation or capabilities of which may not be entirely clear or well defined in existing legislation.[11]

2. What is explainable artificial intelligence?

Explainable Artificial Intelligence (XAI) is the ability of AI systems to provide clear and understandable explanations for their actions and decisions. Its central goal is to make the behaviour of these systems understandable to humans by elucidating the underlying mechanisms of their decision-making processes.

However, many efforts to improve explainability often lead to explanations that are primarily tailored to the AI researchers themselves, rather than effectively addressing the needs of the intended users. This places the responsibility for defining a satisfactory explanation for complex decision models in the hands of AI experts who have a detailed understanding of these models (Miller T. H., 2017).

Ideally, XAI should include the ability to explain the system's competencies and understandings, explain its past actions, ongoing processes and upcoming steps, and disclose the relevant information on which its actions are based (Gunning, 2019).

Transparency, interpretability and explainability

The concepts of transparency, interpretability and explainability in the context of AI have no formal definition, and are sometimes used interchangeably.
In the present document, the different concepts are interpreted as follows:

- **Transparency** refers to the ability for a specific model to be understood. In the strictest sense, **a model is transparent if a person can contemplate the entire model at once**. Transparency can be considered at the level of the entire model, at the level of individual components (e.g., parameters), and at the level of a particular training algorithm. A second and less strict notion of transparency might be that each part of the model (e.g., each input, parameter, and computation) admits an intuitive explanation (Lepri, 2018).

A transparent AI system enables accountability by allowing stakeholders to validate and audit its decision-making processes, detect biases or unfairness, and ensure that the system is operating in alignment with ethical standards and legal requirements.

- **Interpretability** refers to **the degree of human comprehensibility of a given "black box" model or decision** (Lisboa, 2013) (Miller T. H., 2017). Poorly interpretable models "are opaque in the sense that when presented with the resulting decision, rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs (Burrell, 2016).

Interpretable AI models allow humans to estimate what a model will predict given an input, and understand when the model has made a mistake.

- **Explainability** in AI concentrates on providing **clear and coherent explanations for specific model predictions or decisions.** It aims to answer questions like "Why did the AI system make this particular prediction?" by offering human-understandable justifications or reasons for a specific out-come. Explainability requires interpretability as a building block but also looks to other fields and areas, such as human-computer interaction, law, and ethics (Thampi, 2002).

Explainability is particularly important in critical applications where human lives or sensitive information are at stake, as it helps users, regulators, and stakeholders comprehend the rationale behind AI-generated outcomes.

Explainability is important to build trust in AI systems. However, it may not be necessary if systems are sufficiently interpretable. This is easier to achieve with certain (less complex) types of AI.

Rule-based or expert systems, for instance, are subsets of AI that use rules and specific expert knowledge to provide advice or diagnosis, and are commonly used in the sectors of healthcare, logistics and finance. When providing sufficient transparency and interpreted by experts with the appropriate do-main knowledge, these systems can obviate the need to implement explainability mechanisms for the users.

However, XAI may be necessary for non-experts in the field to fully understand the systems.

3. Possible approaches to explainable AI

The possible approaches to AI explainability can be divided into two categories: self-interpretable models, which means that interpretability is built into the design of the systems; and post hoc explanations, where the behaviour of the system is first observed and then explained.

Self-interpretable (or "white box") models feature easy-to-understand algorithms that show how data inputs influence outputs or target variables. "Black box" models, on the other hand, are not explainable by themselves. The lack of explainability might result from an intentional obfuscation from the system designer (Xu, 2018), or from the complexity of the model.

"White box" approach: Self-interpretable models

In "white box" models, the algorithms used are straightforward to understand and it is possible to interpret how the input features are transformed into the output or target variable. The most important features for predicting the target variable can be identified, and these features are understandable (Thampi, 2002).
Interpretability can be provided at different levels: for the whole model, for individual components (e.g., input parameters), or at the level of a particular training algorithm.

Two examples of "white box" models are decision trees, and linear regression.[12]

An example of a decision tree model could be an email classification system that automatically determines whether incoming emails are spam or not. The model is first trained on a dataset of emails labelled as "spam" and "not spam", and recursively partitions the data based on features to create a tree-like structure. At each node, the tree selects the feature that provides the most information gain in terms of classifying emails. The resulting decision tree can be visualised as a flowchart like structure. Each node represents a condition (e.g. "Does the email contain the word "free"?") and each branch represents a possible outcome based on that condition. The leaves of the tree represent the final classification ("spam" or "not spam").

However, certain types of AI present specific difficulties due to their inherent complexity and lack of interpretability. Examples of more complex architectures include neural networks, which consist of multiple layers of interconnected artificial neurons, with each layer performing computations and passing signals to the next layer. Another example of complex architectures are deep learning algorithms, which are neural networks consisting of more than three layers. In many situations, the representations needed to illustrate the internals of the model could become as complex to understand as the models themselves (Lipton, 2018).

This suggests that it would be unrealistic to expect models to be self-interpretable at all times. The post hoc approach therefore seems more appropriate for complex systems.

"Black box" approach: Post hoc explanations

In a post hoc approach, explanations are generated after the model decision has been made, and can be classified as either global or local.

Global explanations provide an overall understanding of the behaviour and decision-making process of an AI model, and aim to capture patterns, general trends, and insights that apply broadly to the model's behaviour (e.g. how does the system select the best candidates for a job vacancy?).

An example of a global explanation technique is "feature importance" (Breiman, 2001), which identifies the most influential features or variables in the model's decision-making process, to help understand which input factors have the greatest impact on the model's predictions or classifications. For instance, elements like the user's listening history, genre preferences, and song metadata can be features significant for a music recommendation system.

Another global explanation technique is "rule extraction" (Craven, 1996), which generates human-readable rules or decision trees that mimic the behaviour of a complex model. These rules provide a global understanding of the decision process and enable interpretability. In a medical diagnosis model, for instance, rules can be extracted indicating specific combinations of symptoms, test results, and patient characteristics that lead to a particular diagnosis (e.g., "if patient age > 50 and blood pressure is high, then diagnose hypertension").

Local explanations, on the other hand, focus on the decision-making process of an AI model for a specific output (e.g. "why my application for a job vacancy has been refused?"). Rather than providing a global explanation that applies to the entire model, local explanations aim to clarify the model's behaviour for a particular case, and understand why a particular prediction or decision was made.

Two examples of techniques for local explanations are LIME and SHAP.

LIME, which stands for Local Interpretable Model-agnostic Explanations (Ribeiro, 2016), is a technique that creates perturbations (manipulates) input data, creating a series of artificial data changing the values of only a part of the original attributes and observes the output of the model. From that observation, LIME creates interpretable "surrogate" models to help explain them. The surrogate models are simpler and more interpretable, allowing users to understand how the input features contribute to the model's decision.

For example, LIME could be used to determine whether an applicant would have been approved for a loan based on various characteristics such as income, credit score and employment history. In such a scenario, LIME could show that the model approved the loan because the applicant's high credit score and stable employment history had the most significant positive impact on the decision. By taking into account the inputs and outputs, LIME would be able to generate a simpler (surrogate) model that could explain which features had more weight in the assessment.

SHAP, or Shapley Additive Explanations (Lundberg, 2017), is a method based on cooperative game theory[13] that assigns values to each feature in a model. It calculates the contribution of each feature to the prediction for a specific instance, considering all possible feature combinations. This technique provides a unified measure of feature importance and helps explain the model's decision at a local level.

To illustrate, consider a machine learning model that predicts house prices based on features, such as square footage, number of bedrooms, and distance to the city centre, and there is need to understand why a particular house received a certain price prediction. SHAP can be applied to the house characteristics, which helps to understand how much each feature contributed to the difference between the model's prediction for this house and the model's average prediction across all houses. This insight can help understand which factors are driving the predictions and how they interact with each other.

There is, however, research that shows potential weaknesses in LIME, SHAP and other perturbation-based post-hoc explanation methods (Slack, 2020) (Lakkaraju, 2020). Because the perturbations injected by these methods are distinguishable from normal input data, models can tell them apart and a malicious developer could potentially create a highly biased and discriminatory model that would provide seemingly "unbiased" output when detecting perturbation-based inputs.

In fact, several studies suggest that post-hoc explanatory methods should not be considered reliable. According to (Vale, 2022) "the use of post-hoc explanatory methods is useful in many cases, but these methods have limitations that prohibit reliance as the sole mechanism to guarantee fairness of model outcomes in high-stakes decision-making". Different research, (Bordt, 2022)refers that "from a technical and philosophical point of view these explanations can never reveal the "unique, true reason" why an algorithm came to a certain decision". It concludes that, "in the worst case, the explanations may induce us into falsely believing that a "justified", or "objective" decision has been made even when this is not the case".

Therefore, the limitations of "black box" approaches should be considered when trying to assess the fairness of the models.

4. XAI and personal data protection

The ability of XAI to provide transparent insights on AI decisions can contribute to ensure compliance with several personal data protection principles, notably transparency, accountability and fairness.

Transparency

Transparency of the data processing is a core principle of data protection. Personal data should be processed in a lawfully, fairly and **transparent manner** in relation to the data subject. Additionally, the controller shall take appropriate measures to provide any information relating to the processing of information of the data subject in a concise, **transparent**, intelligible and easily accessible form, **using clear and plain language**.

The controller is also required to provide the data subject, with information about the existence of automated decisions, including profiling, and with meaningful information about the logic applied, at the time of collection of personal data.[14]

Explainable AI can provide insight into how AI systems process data and arrive to their conclusions, providing an understanding of the "reasoning" that led to the conclusions/ decisions. It can also empower deployers and individuals affected by these systems by increasing their opportunities to understand and interact with the decision-making process.

In general, transparency should foster trust and confidence in the use of AI systems, in addition to being a legal requirement in some cases, in particular when supporting decision making by public administrations, which are legally obliged to justify their decisions.[15]

Data controller accountability

Organisations have a responsibility to ensure that the processing of personal data is carried out in a lawful and transparent manner. This includes the need to implement mechanisms that not only comply with the data protection principles, but also enable effective oversight and auditability of processes.

Greater accountability and understanding of the systems will also lead to a better assessment of the risks that data controllers need to carry out (e.g., when performing data protection impact assessments).

Properly implemented, XAI can facilitate audits and play a key role in holding organisations accountable for their AI-driven decisions, promoting responsible AI development, fostering public trust in these technologies, and ensuring AI is used according to regulatory criteria where applicable.[16]

Data Minimisation

The principle of data protection by design and default emphasises the need to apply technical and organisational measures to implement data protection principles, such as data minimisation. XAI's ability to reveal the most influential factors and features within AI decision-making processes can directly contribute to the reduction of data collection, storage, and processing.

XAI can help organisations comply with data protection regulations by identifying which data points are critical to decision-making. The insights offered by XAI can lead to more focused and targeted data collection efforts, minimising the intrusion into individuals' privacy whilst still achieving accurate and effective AI-driven outcomes.

Special Categories of data

AI training may involve the use of special categories of data, which can pose a high risk to privacy if mishandled or misused. The opacity of AI algorithms can raise concerns about the processing of special categories of data and the potential impact on outcomes if, for example, a particular category of data such as religion or sexual orientation can be inferred from the training data.

AI systems, such as machine learning models, can identify correlations between certain attributes and information related to the data subjects - these are known as proxy attributes. In certain situations, proxy attributes might be used to infer specific categories of data about individuals.

For example, in some cities there may be a strong correlation between the postcode and the ethnicity of the population, making the postcode attribute a proxy for ethnicity. An AI system could identify this correlation during its training and make decisions based on this proxy attribute when it is used, for example, to make credit reliability decisions.

However, there is a risk that such inferences about individuals may be completely wrong. XAI can help developers and users to identify proxy attributes that may be linking decisions to particular categories of data.

It is important to emphasise that the implementation of XAI **does not automatically lead to compliance with data protection regulations.**
However, XAI may be a technical measure that is useful for the controller to demonstrate that data processing activities have been conducted according to data protection principles, having regard to the purpose, nature, context and scope of the data processing activities and the probability of a serious risk to the freedoms and rights of natural persons.

5. Risks associated with the implementation of XAI

Whilst XAI has the potential to promote transparency and trust in AI systems, its adoption can create risks for controllers, developers, engineers, and data subjects. When implementing explainability, precautions should be taken to mitigate the following risks.

Misinterpretation

Depending on how it is implemented, XAI can lead to explanations that are too complex or technical for the audience to understand, or oversimplified in a way that does not capture the full complexity of AI models (as described in the section **"Black box approach: Post hoc explanations"**). In either case, this could lead to misinterpretation by individuals.

Information relating to the processing should be provided to the data subjects in a concise, transparent, intelligible and easily accessible manner, using clear and plain language. Explanations should therefore, be presented in an understandable way, avoiding jargon and technical complexity.

In order to reduce the risk of misinterpretation, organisations should first identify the different **stakeholders** to which they want to provide explanations. Then, for each audience, the level of detail of the explanations should be adjusted. Explanations should be provided in clear and plain language to bridge the gap between the complexity of the subject matter and the individual's level of understanding. The explanation process can be facilitated by the use of user-friendly interfaces with graphical representations, however this should not lead to an oversimplification of the systems. Careful validation and testing of the XAI methods is essential to ensure that explanations accurately reflect the behaviour of the AI system, and that users are not misled by incomplete or inaccurate explanations.

In addition, organizations must ensure that XAI explanations are not only clear, but also **neutral**, avoiding any reinforcement of existing biases.

Potential exploitation of the systems

Organisations need to implement appropriate technical and organisational measures to ensure a level of security appropriate to the risks to the individuals, including the confidentiality, integrity, and availability, resulting from the processing of their data.

In the context of XAI, this means avoiding the risk of disclosing personal data or details that could be used to exploit the AI system and potentially affect individuals.

The paper (Kuppa, 2021) presents several different types of attacks, including membership inference attacks and adversarial attacks, made against an anti-virus system using information provided by XAI. According to the authors, *"counterfactual explanation methods can help attackers to find quicker ways to find adversarial samples, instead of solving a hard-to converge "black-box" optimization problem in input space. Attackers can simply use counterfactual explanations to optimize their attack path"*.

This requires a careful balance between transparency and protection of sensitive components of the system.

Disclosure of trade secrets

Similarly, XAI raises the issue of the potential risk of loss of business competitiveness for the provider (and/or for the deployer) of the AI system, due to the disclosure of proprietary information or sensitive business strategies.

The principles of accountability and data protection by design and by default are relevant here.

From the moment that the means of processing are determined, organisations should build mechanisms into XAI's implementation to ensure that the explanations are informative, in particular for the individuals possibly affected by the use of these systems. This can be done without undue disclosure of proprietary algorithms, trade secrets or other commercially sensitive details.

Over-reliance on the AI system by deployers

Explanations can increase the likelihood that humans will 'blindly' accept AI's recommendations (automation bias) regardless of their correctness. Human's critical engagement is indeed a necessary component for a successful 'human-AI interaction', especially where the cost of error is high, such as in the field of healthcare (Gajos, 2022).

Organisations need to ensure that XAI implementations include mechanisms that allow individuals to, at a minimum, obtain human intervention from the controller, express their point of view, and challenge the decision. This is in line with the right of individuals not to be subjected to a decision based solely on automated processing, including profiling, which has legal effects or otherwise significantly affects them. More generally, XAI aims to ensure that those responsible for making a decision remain in control of the decision, maintain a balanced perspective, and do not become overly dependent on AI systems.

To address the risk of over-reliance on AI systems, organisations should actively promote human involvement and human oversight on decisions that have significant consequences, particularly risks to physical or economic harm, or risks to the rights and freedoms of individuals and groups.

Clear communication about the limitations of AI is needed to ensure that technological progress is translated into responsible and socially acceptable decisions. People affected by the use of these systems should be encouraged to seek human intervention where necessary (and should have easy and timely access to human support).

In essence, whilst XAI offers significant potential, it should be accompanied by a comprehensive understanding of both its importance in enhancing the trustworthiness of the AI and of its limitations. This requires comprehensive and in-depth risk assessments, continuous monitoring of the functioning of the AI systems, and collaborative efforts between data protection authorities and the competent sectorial oversight authorities (e.g., labour inspectorate, health-care oversight body, financial oversight authority, etc.) to ensure responsible and secure implementation.

5. The importance of the human factor

Whatever the approach to explaining AI systems, it is essential that the human aspect is taken into account, as explanations must ultimately be relevant and meaningful to people. Human beings perceive and process information in different ways from one another, based on a number of different factors: humans preferences for contrastive explanations, their selectiveness, their trust in the explanations, and their ability to contextualise explanations.

Aspects to consider when providing explainability

HUMANS PREFER CONTRASTIVE EXPLANATIONS

Beyond wanting to know “Why?” people tend to ask “Why event P happened instead of Q?”. Contrastive explanations simplify complex decision-making processes by highlighting the key differences between options and provide a basis for individuals to learn from past choices and refine their own decision strategies. (Miller T., 2019)

HUMANS ARE SELECTIVE

When faced with complex explanations, individuals may selectively focus on the most strinking or relevant aspects, while filtering out the detaills that they consider to be less important. They may also tend to gravitate towards explanations that align with their existing knowledge. (Mittelstadt, 2019)

HUMANS MUST TRUST THE EXPLANATIONS

Trust can be considered in terms of the accuracy and reliability of the system, but also in terms of how much individuals trust the explanations given. Mistrust of the whole system can result from explanations that are too complicated, incomplete or inaccurate. (Ribeiro, 2016)

EXPLANATIONS ARE CONTEXTUAL

XAI systems should be able to explain their capabilities and understandings, however every explanation is set within a context that depends on the task, abilities, and expectations of the user of the AI systems (Gunning, 2019).

EXPLANATIONS ARE SOCIAL

Explanations are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer’s beliefs about the explainee’s beliefs. They are influenced by individual vs. group behaviour, norms and morals, etc. (Miller T., 2019)

In addition, the "intended audience" needs to be taken into account when providing explanations.

For example, supervisory authorities would probably need more detailed explanations when auditing the system to verify compliance with the legislation applicable to the activity in relation to which the AI is put into service (e.g., health and safety conditions at work, in the case of an AI work management system).

Conclusion

To be trustworthy, AI should be, amongst others, transparent, accountable and ethical,[17] and explainable AI can play an important role in meeting these specific requirements.

From the perspective of the EDPS, the concept of XAI embodies the commitment to develop 'human-centred' AI. By explaining the "why" behind AI decisions, it enables individuals to participate in the digital landscape in a meaningful way, for example by filing complaints if their rights are violated by AI-driven decisions.

XAI empowers individuals with understandable insights into how their personal data is being handled by revealing the rationale behind AI decisions. The transparency provided by XAI not only strengthens trust between organisations and users, but is also in line with core data protection principles.

In addition, XAI would enable data subjects to identify and challenge unfair decisions, thereby enhancing fairness in decision-making, promoting equal treatment and the principle of non-discrimination. But, XAI is not just a step towards transparency, it is a leap towards bridging the gap between machine-driven processes and the human search for justification, trust, and fairness. The role of XAI goes beyond mere explanation; it embodies the recognition that human understanding, ethical judgement and empathetic consideration are integral facets in the use of AI technologies. The human dimension remains essential. As we embrace the transformative potential of artificial intelligence, it is crucial to remember that the use of AI systems can have profound consequences on individuals as well as on society as a whole. In this context, AI goes beyond technology to combine AI's potential with human need for understanding, responsibility and ethical oversight.

The adoption of XAI contributes to a future where AI should be defined not only by its technical capabilities, but also by humanity's collective responsibility to uphold human rights, ethics, and accountability.

However, the advent of XAI also presents a number of potential challenges that require careful attention.

As highlighted above: complex or oversimplified explanations could lead to misinterpretation; XAI, if used incorrectly or maliciously, could degenerate into "persuasion exercises" to justify system behaviour; and over-protection of trade secrets could hinder transparency. Additionally, the financial cost of XAI should also be taken into account, as emphasised by (Adadi, 2018): *"Furthermore, making AI systems explainable is undoubtedly expensive; they require considerable resources both in the development of the AI system and in the way it is interrogated in practice."*

As the capability (and public demand) of AI systems grows, so does the risk that AI developers may cut corners and disregard ethical considerations in pursuit of new breakthroughs.

As a society, we have a responsibility to demand and ensure that this development takes place in a way that safeguards fundamental rights, notably the fundamental rights to privacy and to the protection of personal data.[18]

As the data protection authority supervising EU Institutions, bodies, offices and agencies, it is our responsibility to ensure that the use of AI systems is in line with data protection principles and complies with the rules laid down in the applicable legislation.

References

Adadi, A. &. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6.

Antoniadi, A. M. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Applied Sciences, 11(11).

Bellotti, V. &. (2001). Intelligibility and accountability: human considerations in context-aware systems. Human–Computer Interaction, 16(2-4), 193-212.

Bordt, S. F. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 891-905.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big data & society, 3(1).

Craven, M. W. (1996). Extracting comprehensible models from trained neural networks. The University of Wisconsin-Madison.

Gajos, K. Z. (2022). Do people engage cognitively with AI impact of AI assistance on incidental learning. In 27th International Conference on Intelligent User Interfaces, 794-806.

Gunning, D. S. (2019). XAI—Explainable artificial intelligence. Science robotics, 4(37), eaay7120.

Kazim, E. &. (2020). Explaining decisions made with AI: a review of the co-badged guidance by the ICO and the Turing Institute. Available at SSRN 3656269.

Kuppa, A. &.-K. (2021). Adversarial XAI methods in cybersecurity. IEEE transactions on information forensics and security, 16, 4924-4938.

Lakkaraju, H. &. (2020). "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 79-85.

Lepri, B. O. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. Philosophy & Technology, 31, 611-627.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 31-57.

Lisboa, P. J. (2013). Interpretability in machine learning–principles and practice. International Workshop on Fuzzy Logic and Applications. Cham: Springer International Publishing., 15-21.

Lundberg, S. M. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.

Miller, T. H. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547.

Mittelstadt, B. R. (2019). Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency, 279-288.

Peters, U. (2023). Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque. AI and Ethics, 3(3), 963-974.

Rengasamy, D. M. (2022). Feature importance in machine learning models: A fuzzy information fusion approach. Neurocomputing, 511, 163-174.

Ribeiro, M. T. (2016). "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135-1144.

Slack, D. H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society , 180-186.

Thampi. (2022). Interpretable AI: Building explainable machine learning systems. Simon and Schuster, 14-15.

Vale, D. E.-S. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. AI and Ethics, 2(4), 815-826.

Xu, H. S. (2018). Deepobfuscation: Securing the structure of convolutional neural networks via knowledge distillation. arXiv preprint arXiv:1806.10313.

End notes

[1] According to a McKinsey 2022 report, the adoption of AI doubled between 2017 and 2022, and the level of investment in AI increased alongside its rising adoption (available at **https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review**).
[2] A large language model is a trained deep learning model that understands and generates text in a human-like way. ChatGPT is available at **https://chat.openai.com**
[3] A text-to-image model is a machine learning model that takes an input natural language description and produces an image that matches that description. Stable Diffusion is available at **https://stablediffusionweb.com**
[4] The trend towards making AI technologies more affordable and user-friendly for a wider range of users is sometimes referred to as the "democratization of AI".
[5] In AI, training is the process of providing data to an AI model so that it can learn from that data, which typically requires large datasets of examples or labelled information.
AI models are software programs trained to perform specific tasks such as pattern recognition and prediction.
[6] A phenomenon that occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions during the algorithm development.
[7] "Hallucination" in AI refers to the generation of outputs that may sound plausible but are either factually incorrect or unrelated to the given context. There can be several reasons for "hallucinations". They can be related to a lack of real-world understanding of the AI models, too much specialisation (overfitting), or poor-quality training data.
[8] An example of this was the system that Amazon developed between 2014 and 2015 to screen the CVs of job candidates. The aim was to automate the search for top talent, but the development team found that the system was teaching itself to favour male candidates. Reportedly, Amazon's system had been trained to screen applicants by observing patterns in resumes submitted to the company over a 10-year period, most of which were from men. It penalised resumes that contained the word "women's", as in "women's chess club captain". Amazon scraps secret AI recruiting tool that showed bias against women, **https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G**
[9] In 2017, an investigation published by STAT News revealed potential problems with IBM's Watson for Oncology system. The investigation reported that physicians around the world had complained that the technology frequently recommended cancer treatments that were not suitable for their patients. A document reportedly obtained during the investigation stated that the "inadequacy of the training cases" for Watson for Oncology "undermined" the technology's effectiveness. IBM's Watson recommended 'unsafe and incorrect' treatments for cancer patients, investigation reveals, **https://www.advisory.com/daily-briefing/2018/07/27/ibm**
[10] A famous case concerns the Dutch government's system for detecting various forms of fraud, including social benefits, allowances and tax fraud, the Systeem Risico Indicatie (SyRI). In February 2020, the District Court of The Hague delivered a judgment in a case about SyRI. On paragraph 6.90 the Court refers that "The foregoing results in the inability to verify how the simple decision tree, to which the State refers, is generated and of which steps it is comprised. Consequently, **it is difficult to comprehend how a data subject could be able to defend themselves against the fact that a risk report has been submitted about him or her.** It is just as difficult to see how a data subject whose data were processed in SyRI but which did not result in a risk report, can be aware that their data were processed on correct grounds." In paragraph 6.86 the Court mentions that "Considering the principle of transparency, the principle of purpose limitation and the principle of data minimisation – fundamental principles of data protection – the court holds that the SyRI legislation is insufficiently transparent and verifiable to conclude that the interference with the right to respect for private life which the use of SyRI may entail is necessary, proportional and proportionate in relation to the aims the legislation pursues."
The English version of the decision can be found in **https://uitspraken.rechtspraak.nl/#!/details?id=ECLI:NL:RBDHA:2020:1878**
[11] See the **Statement by the Consumer Financial Protection Bureau (CFPB), CFPB Issues Guidance on Credit Denials by Lenders Using Artificial Intelligence. Consumers must receive accurate and specific reasons for credit denials**, 19 September 2023, *"Technology marketed as artificial intelligence is expanding the data used for lending decisions, and also growing the list of potential reasons for why credit is denied,"* said CFPB Director Rohit Chopra. *"Creditors must be able to specifically explain their reasons for denial. There is no special exemption for artificial intelligence."*

[12] Decision trees are graphical representations of decision-making processes that resemble a tree structure, where each internal node represents a decision or test on an attribute, and each branch represents a result of the test. Linear regression is a statistical and machine learning technique used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, i.e. it tries to find the best fitting straight line (a linear equation) that represents the data.

[13] SHAP is based on the concept of Shapley values, an expression that was coined after the mathematician Lloyd Shapley for his work in the field of game theory. The concept can be explained using a football analogy: in a coalition of players who cooperate and achieve a certain total gain from that cooperation, the Shapley values can be used to determine how important each player is to the overall cooperation, and what payoff he or she can reasonably expect.

[14] See **WP29 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679**, adopted on 3 October 2017, as last Revised and Adopted on 6 February 2018.

[15] Having regard to automated systems to be used by public sector bodies, a useful assessment tool is provided by the European Law Institute (ELI) **Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration ('Model Rules')**. These Model Rules provide for an impact assessment of those algorithmic decision-making systems used by public authorities which are likely to have significant impacts on persons concerned by their use. Though tailored to public administrations, the Model Rules provide a methodology and a set of questions that are in most cases also applicable in case of use of algorithmic decision-making systems by private entities. It can be noted that **transparency of the system and explainability of its decisions** is a core element of this Impact Assessment, see questions for Impact Assessment at page 38: "transparency of the system and explainability of its decisions; 32. How will you inform persons concerned and the public about the existence and functioning of the system? 33. Can you explain the decision(s) of the system to the persons concerned? 33.1 Do you continuously survey the persons concerned to determine whether they understand the decision(s) of the system?"

[16] On accountability as an essential requirement to ensure the "alignment" of the data processing by the controller with the applicable legal requirements, and on the resulting limitations on the use of machine learning AI systems, see Judgment of the Court of Justice of the European Union (Grand Chamber) of 21 June 2022, C-817/19, Ligue des droits humains, ECLI:EU:C:2022:491, paras. 194-195: "194. As regards the criteria that the PIU may use to that end, it should be noted, first, that according to the very wording of Article 6(3)(b) of the PNR Directive those must be "pre-determined" criteria. As noted by the Advocate General in point 228 of his Opinion, that requirement precludes the use of artificial intelligence technology in self-learning systems ("machine learning"), capable of modifying without human intervention or review the assessment process and, in particular, the assessment criteria on which the result of the application of that process is based as well as the weighting of those criteria. 195. It is important to add that use of such technology would be liable to render redundant the individual review of positive matches and monitoring of lawfulness required by the provisions of the PNR Directive. As observed, in essence, by the Advocate General in point 228 of his Opinion, given the opacity which characterises the way in which artificial intelligence technology works, it might be impossible to understand the reason why a given program arrived at a positive match. In those circumstances, use of such technology may deprive the data subjects also of their right to an effective judicial remedy enshrined in Article 47 of the Charter, for which the PNR Directive, according to recital 28 thereof, seeks to ensure a high level of protection, in particular in order to challenge the non-discriminatory nature of the results obtained."

[17] On 8 April 2019, the High-Level Expert Group on AI presented ethical guidelines for trustworthy artificial intelligence. Among the seven key requirements that AI systems should meet to be considered trustworthy are transparency, accountability, and fairness. See more in **https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai**

[18] Articles 7 and 8 of the Charter of Fundamental Rights of the European Union.

This publication is a brief report produced by the Technology and Privacy Unit of the European Data Protection Supervisor (EDPS). It aims to provide a factual description of an emerging technology and discuss its possible impacts on privacy and the protection of personal data. The contents of this publication do not imply a policy position of the EDPS.

Issue Authors: Vítor BERNARDO
Editors: Massimo ATTORESI, Xabier LAREO, and Luis VELASCO
Contact: **techdispatch@edps.europa.eu**

To subscribe or unsubscribe to TechDispatch publications, please send a mail to **techdispatch@edps.europa.eu**. The data protection notice is online on the **EDPS website**.

edps.europa.eu

@EU_EDPS

EDPS

European Data Protection Supervisor

@EDPS@social.network.europa.eu

@EDPS@tube.network.europa.eu