transparency

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

Transparency is a key requirement for trustworthy artificial intelligence (trustworthy AI). It obliges the provider of an artificial intelligence system (AI system) to supply the information needed to understand the system's output, limitations, reliability, and intended use. For machine learning (ML) methods, transparency is closely related to explainability, i.e., the extent to which humans can understand the predictions of a learned hypothesis. The EU AI Act makes transparency a binding design requirement: the provider must design a high-risk artificial intelligence system (high-risk AI system) to be sufficiently transparent for its deployers to interpret its output, persons exposed to an AI system must be informed that they are interacting with an automated system, and affected persons may obtain an explanation of decisions based on the output of a high-risk AI system. Transparency also encompasses documentation of the algorithm design, the training datasets, and the intended use of an AI system. Finally, content generated by an AI system, such as a deep fake, must be marked as artificially generated.

Definition

Transparency is a fundamental requirement for trustworthy artificial intelligence (trustworthy AI) (High-Level Expert Group on Artificial Intelligence, 2019). In the context of machine learning (ML) methods, transparency is often used interchangeably with explainability (Gallese, 2023; Jung and Nardelli, 2020). Some ML methods inherently offer this form of transparency: logistic regression quantifies the confidence in a classification via the value $|h(\mathbf{x})|$, the distance of the feature vector from the decision boundary, which indicates how reliable an individual prediction is, and decision trees allow human-readable decision rules (Rudin, 2019). In the broader scope of artificial intelligence systems (AI systems), transparency extends beyond explainability and includes providing information about the system's limitations, reliability, and intended use.

The binding transparency obligations discussed in the following are those of the EU AI Act; other jurisdictions impose related but distinct duties (see explainability for a state law of the United States). The Act distributes its obligations among three roles: the provider that develops an AI system, the deployer that uses it, and the persons who interact with the AI system or are affected by its output (see deployer).

Toward the deployer, the provider must design and develop a high-risk artificial intelligence system (high-risk AI system) so that its operation is sufficiently transparent to enable the deployers to interpret its output and use it appropriately (European Parliament and Council of the European Union, 2024, Art. 13). In medical diagnosis, a clinic acts as the deployer: the provider can meet this duty by designing the system to disclose to the clinician the confidence level for the predictions delivered by a learned hypothesis.

Toward persons who interact directly with an AI system, such as an artificial intelligence (AI)-powered chatbot, the provider must design the AI system so that these persons are informed of that fact (European Parliament and Council of the European Union, 2024, Art. 50(1)). Providers of AI systems that generate synthetic audio, image, video, or text content must mark these outputs in a machine-readable format as artificially generated (European Parliament and Council of the European Union, 2024, Art. 50(2)); watermarking and content provenance techniques implement this marking. For a deep fake, the duty changes roles: its deployer must, in addition, disclose visibly that the content is artificially generated or manipulated (European Parliament and Council of the European Union, 2024, Art. 50(4)).

An affected person, by contrast, need not interact with the AI system at all. A patient whose diagnosis is supported by an AI-based system deals only with the clinician: the deployer must inform the patient that a high-risk AI system is used concerning them (European Parliament and Council of the European Union, 2024, Art. 26(11)). The EU AI Act also grants the affected person a right to explanation: the deployer must, on request, provide a clear and meaningful explanation of the role of the AI system in a decision based on its output (European Parliament and Council of the European Union, 2024, Art. 86). In credit scoring, for example, a loan applicant faced with an adverse automated decision may obtain from the deployer, under this right, an explanation of the contributing factors, such as income level or credit history, and use it to contest the decision.

Transparency also encompasses documentation detailing the purpose and design choices underlying the AI system. For a high-risk AI system, the provider must draw up documentation that covers the algorithm design and the datasets used for training (European Parliament and Council of the European Union, 2024, Art. 11). The provider of a general-purpose AI model (GPAI model) must additionally publish a sufficiently detailed summary of the content used for training (European Parliament and Council of the European Union, 2024, Art. 53). Datasheets for datasets (Gebru et al., 2021) and model cards (Mitchell et al., 2019) help practitioners understand the intended use cases and limitations of an AI system.

Fig. 1 summarizes these transparency obligations as information flows between the provider of an AI system, its deployer, and the persons exposed to its output.

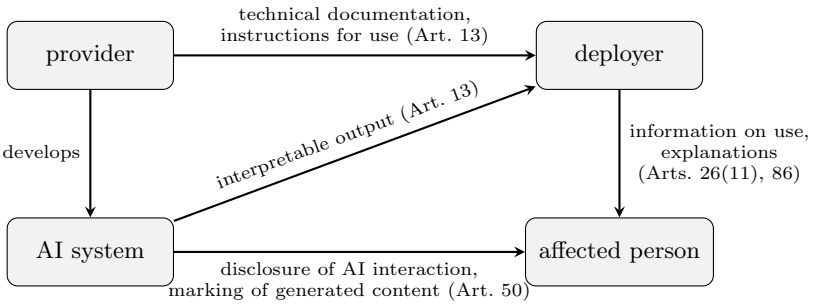

Figure 1: Transparency obligations of the EU AI Act as information flows. The provider of a high-risk AI system supplies technical documentation, covering the algorithm design and the training datasets, along with instructions for use. The deployer must be able to interpret the output of the AI system. Persons who interact directly with the AI system must be informed of that fact, and generated content must carry a machine-readable mark that it is artificial (Art. 50). An affected person who never interacts with the AI system, such as a patient whose diagnosis the system supports, is informed of its use, and owed an explanation of an adverse decision, by the deployer (Arts. 26(11), 86)

Cross References

trustworthy AI, explainability, interpretability, explanation, right to explanation, EU AI Act, high-risk AI system, AI system, GPAI model, provider, deployer, watermarking, deep fake, content provenance

References

1. Mitchell et al. (2019). Model Cards for Model Reporting. In: Proc. Conf. Fairness, Accountability, Transparency, pp. 220–229.

2. Gebru et al. (2021). Datasheets for datasets. Commun. ACM, 64(12), pp. 86–92.

3. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

4. High-Level Expert Group on Artificial Intelligence (2019). Ethics Guidelines for Trustworthy AI.

5. Jung and Nardelli (2020). An Information-Theoretic Approach to Personalized Explainable Machine Learning. IEEE Signal Process. Lett., 27, pp. 825–829.

6. Gallese (2023). The AI Act proposal: A new right to technical interpretability?. SSRN Electron. J.

7. Rudin (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Mach. Intell., 1(5), pp. 206–215.