Transparency of AI Systems

Oliver Müller[1] [2]
Federal Office for Information Security (BSI), Saarbrücken, Germany

Veronika Lazar[1]
Federal Office for Information Security (BSI), Saarbrücken, Germany

Matthias Heck
Federal Office for Information Security (BSI), Saarbrücken, Germany

Abstract

Artificial Intelligence (AI) has now established itself as a tool for both private and professional use and is omnipresent. The number of available AI systems is constantly increasing and the underlying technologies are evolving rapidly. On an abstract level, most of these systems are operating in a black box manner: only the inputs to and the outputs of the system are visible from outside. Moreover, system outputs often lack explainability, which makes them difficult to verify without expert knowledge. The increasing complexity of AI systems and poor or missing information about the system make an assessment by eye as well as assessing the system's trustworthiness difficult. The goal is to empower stakeholders in assessing the suitability of an AI system according to their needs and aims. The definition of the term transparency in the context of AI systems represents a first step in this direction. Transparency starts with the disclosure and provision of information and is embedded in the broad field of trustworthy AI systems. Within the scope of this paper, the Federal Office for Information Security (BSI) defines transparency of AI systems for different stakeholders. In order to keep pace with the technical progress and to avoid continuous renewals and adaptions of the definition to the current state of technology, this paper presents a technology-neutral and future-proof definition of transparency. Furthermore, the presented definition follows a holistic approach and thus also takes into account information about the ecosystem of an AI system. In this paper, we discuss our approach and proceeding as well as the opportunities and risks of transparent AI systems. The full version of the paper includes the connection to the transparency requirements in the EU AI Act of the European Parliament and council.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases transparency, artificial intelligence, black box, information, stakeholder, AI Act

Digital Object Identifier 10.4230/OASIcs.SAIA.2024.11

Category Practitioner Track

Related Version A full version of the paper is available at the BSI website:
German version: `Transparenz von KI-Systemen`
English version: `Transparency of AI systems`

Acknowledgements We would like to thank our colleagues from the Central Office for Information Technology in the Security Sector (ZITiS) as well as from the Federal Office for Information Security (BSI) for their critical comments and for proofreading the full version of the paper.

[1] These authors contributed equally to this work.
[2] Corresponding author.

1 Introduction

In this paper, we present a definition of transparency for information technology systems that have integrated artificial intelligence (AI). The aim of this publication is to develop a common understanding of the term transparency and to highlight the relevance of transparency for various stakeholders and the BSI. Therefore, the paper is addressed to all stakeholders of AI systems and is intended to show, among other things, that different stakeholders may also have different transparency requirements.

2 Definition

▶ **Definition 1.** *Transparency of AI systems is the provision of information about the entire life cycle of an AI system and its ecosystem. Transparency promotes accessibility to information that enable an assessment of the system with regard to different needs and objectives for all stakeholders.*

2.1 Elements

The above definition is based on the presentation of the transparency concept in [5] and [2]. It is compliant with the transparency requirements in the EU AI Act (see full version of the paper for details) and represents the position of the BSI. In the following subsections, the individual elements of the definition are described in more detail.

2.1.1 AI system

The EU AI Act governing the regulation of artificial intelligence, defines an AI system as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (cf. Article 3 EU AI Act). In its definition, the BSI explicitly formulates the hardware component and defines AI systems as software and hardware systems that utilise artificial intelligence in order to behave "rationally" in the physical or digital dimension. Based on their perception and analysis of their environment, these systems act with a certain degree of autonomy in order to achieve certain goals [3]. The technology integrated into these systems, known as AI, consists of different disciplines like machine learning, inference and robotics. These include expert systems and neural networks used in AI systems. This is not an exhaustive list of the techniques used, but is intended to show the abundance of different techniques. An equally wide range becomes clear when considering the different functionalities of AI systems, ranging from simple to highly complex tasks. AI systems can perform tasks such as pattern recognition, classifications, forecasts, recommendations, natural language processing or computer vision and can also be combined with each other in a variety of ways. In addition, AI can be implemented in the systems in different ways depending on the requirements and objectives of the application. On the one hand, it can be developed and used as a separate application and represent the primary function of the system, as is the case, for example, in chatbot applications. On the other hand, it can also be integrated into existing systems, for example in order to expand their functionality and/or increase performance in background processes. Considering the way AI systems are implemented, the degree of automation can also vary greatly. While some systems only use the output of AI as recommendations and

require humans as the final decision-making authority, other (sub-)systems autonomously implement the decisions and classifications of AI without further human action. Overall, it can be summarised that there is not one AI system, but rather a wealth of different techniques, functionalities and forms of implementation.

2.1.2 Ecosystem

In this paper, the term ecosystem refers to the context in which an AI system is developed, deployed and operated. The information regarding the ecosystem of an AI system goes beyond the actual AI system and should, for example, also include details about the provider (e.g. location, contact details) or the development process of the system. The term should also include the entire supply chain of the AI system. The decision to include information about the ecosystem of an AI system in the definition of transparency is based on the fact that there is a (conditional) dependency between the actual AI system and its ecosystem. For example, if the AI system is developed and operated outside the European Union in a third country, corresponding questions and challenges arise with regard to the underlying level of IT security and data protection level. This meta-information can support a well-founded assessment of the situation as well as an informed decision by stakeholders.

2.1.3 Information

Information is the basis for the knowledge needed by stakeholders to form an assessment of the AI system and its ecosystem. They must be disclosed and made accessible so that they are available for this purpose. In addition, information must be relevant and appropriate for gaining knowledge. The full version of the paper explains the transparency requirements in the EU AI Act and sets out the minimum information to be disclosed by providers and operators of certain AI systems.

2.1.4 Life cycle

According to [1], the life cycle of an AI system comprises various phases:

- Planning and conception phase
- Design, development and validation phase
- Commissioning and application phase
- Continuous evaluation phase
- System updates
- Decommissioning

A brief desciption of the phases in the context of the concept of transparency is given in the full version of the paper.

2.1.5 Needs and objectives

These are individual and can be quite different in varying applications. This term is intended to reflect the fact that transparency is not intended to enable access to specific information, but rather to provide information that enables the respective stakeholders to make an assessment. What is specifically assessed by the respective stakeholder is individual and contextual. Stakeholder needs and objectives vary depending on the application scenario. The aim is to cover a wide range of desirable system information enabling the system to be assessed in terms of the specific needs of stakeholders.

2.1.6 Stakeholder

The term "stakeholder" refers to all parties who are either indirectly (e.g. through impacts) or directly (e.g. through application) affected by an AI system or who interact with the system (e.g. developers). These can be individual persons or groups of people. A stakeholder does not have to play an "active" role. While consumers and users usually only use an AI system, it is possible that experts, developers and companies/organisations provide an AI system in addition. Indirectly affected persons/third parties do not provide an AI system, nor do they use it. Nevertheless, they can be affected by the impact and thus become (passive) stakeholders. The presented list of different stakeholders does not claim to be exhaustive and can be refined as desired. However, the chosen representation is sufficient to show that there may be different interests with regard to an AI system, which can be reflected, among other things, in different requirements for transparency – such as the type or level of detail of the information provided – of an AI system. Therefore, the various stakeholders must be taken into account when defining the concept of transparency.

3 Discussion

3.1 Approach and procedure

The existence of already published definitions of the concept of transparency raises the question of the need for a further definition. The sheer breadth of the topic of transparency on the definition market ultimately reflects the different requirements for transparency depending on the stakeholders and the area of application. In order to provide a basis for further work of the BSI with a focus on broad stakeholder groups and generic areas of application, it was decided not to use existing definitions. In addition, the speed at which technologies in the AI sector are developing is enormous. This harbours the risk that definitions, once established, may lose their validity, especially if they are too specific. In order to keep pace with technical progress and to avoid constant renewal and adaptation of the definition to the current state of the art, the definition of transparency presented in this paper is as technology-neutral and future-proof as possible. On the one hand, it should be easy to understand, cover all relevant aspects of transparency and at the same time be open enough to allow individual interpretation depending on the respective stakeholder and the AI technology used. On the other hand, it should serve as a generic basis for future work of the BSI in this area. Furthermore, a holistic approach was taken to the definition: transparency includes both the provision of information about the AI system itself and about its ecosystem, such as the supply chain of the AI system or details about the provider. The weighting of the information provided is the responsibility of the respective stakeholder.

3.2 The aim of transparency

By promoting the transparency of AI systems, the aim is to strengthen the autonomy of stakeholders and enable them to decide for themselves whether the use, modification or provision of an AI system is appropriate and justifiable for them. It is not enough to simply describe the capabilities of the AI system. The limitations of the system must also be analysed and made transparent. Only in this way can a holistic assessment be made by the stakeholders, e.g. whether an AI system is suitable for a particular purpose or not. In the area of digital consumer protection, this should help to ensure that consumers can recognise and use safe and trustworthy AI systems despite increasing digitalisation. Companies and organisations should also be enabled to develop and operate their own AI

systems transparently. Transparent information from the ecosystem of an AI system should also enable third parties affected by impacts to recognise how they can assert their rights in the event of damage. The transparency of AI systems thus serves to empower stakeholders.

3.3 Opportunities through transparency

The use of transparent AI systems can promote the traceability of decisions and the assessment of the appropriateness of systems. Transparency can also help to protect against misuse by enabling potential risks and undesirable effects to be recognised at an early stage. In order to react appropriately to problems, it is important to know, for example, whether the output of an AI system is free from discrimination or whether it violates licence conditions. In terms of consumer protection transparency can also act as a support tool. Transparency provides the basis for a correct assessment of the appropriateness of the system used. In order to be able to make such an assessment at all, information about the system must be accessible. A valid assessment of the adequacy of the AI system forms the basis for positive trust and acceptance processes. Initial publications show, for example, higher download numbers for transparent AI models, which could be an indication of better acceptance of these systems among developers [4]. Lack of transparency complicates the valid assessment of the adequacy of a system, and thus the assessment of its trustworthiness. The latter is a prerequisite for establishing and maintaining a positive relationship of trust with the system and the related output. In addition, transparency can enable users to exercise rights more easily by requiring transparency accompanied by clearer definitions of legal responsibilities and identification of those responsible for the use of AI systems. The listed aspects of traceability, abuse protection, acceptance and trustworthiness as well as legal responsibility show the relevance of transparency in the use of AI systems. This relevance is also reflected in regulatory and legal requirements (see full version of the paper). Transparency can, on the one hand, contribute to the security of AI systems and, on the other hand, promote the safe use of AI applications. In this way, transparency can enable the identification of possible problems and vulnerabilities, make undesirable system behaviour visible and contribute to problem identification and the prevention of misuse. In the context of IT security, transparency also provides the basis for the disclosure and assessment of risks associated with the use of the system. The identification of roles and responsibilities in the event of damage and adverse events as part of transparency requirements can also help stakeholders to detect faulty system behaviour, reduce response times and thus mitigate possible consequential damage. If transparency is practised in the early phases of the life cycle of an AI system, inconsistencies can be avoided within the development team from the outset, sources of error minimised and training phases shortened. AI systems are both developed and increasingly used as development tools – e.g. in AI-supported programming for the automatic generation of program code – for new systems. In the early development phases, it is also important from a developer's point of view to know where the training/test/validation data come from, how they are obtained and whether they are free of bias – e.g. to avoid discrimination. This information is important in order to be able to prepare the data correctly before training/testing/validating (pre-processing). Current development trends also show that pre-existing models are often used, which makes the presence and accessibility of all safety-critical information on existing models particularly relevant. In the absence of this information, there is a risk of implementing the security risks of the basic models into your own products. Both systems are then interdependent, and any lack of transparency in the underlying system is transferred to the system built on top of it. The inheritance of the security risks described above from different areas leads to an increased overall risk in the examples mentioned, which underlines once again

the explosiveness of transparency for security-relevant aspects when using AI systems. In addition, as discussed in the previous section, transparency can contribute to better user empowerment in assessing AI systems. A correct assessment of the application, suitable application scenarios and possible problems and security risks can promote safe use by users.

3.4 Dangers of transparency

So far, the positive aspects of transparency have mainly been presented. Increasing/improving the transparency of AI systems can also have unintended negative effects. For example, the provision of information on the functionality or architecture of an AI system can reveal new attack vectors that attackers can exploit to misuse or compromise the system. Information about limitations or excluded areas of application of an AI system could also be deliberately exploited by attackers, e.g. to deliberately generate erroneous behaviour or destructive output. Conversely, attackers can also abuse the trust that transparency is supposed to create in order to deliberately provide incorrect information. For example, in reality, safety-critical applications can be presented as uncritical. In addition, non-transparent systems can also be marked as transparent. Such pseudo-transparency can be used for marketing purposes of the own product and lead to a wrong assessment of the system by consumers if the transparency label is not checked. Therefore, questions about the trustworthiness of the information disclosed/provided must also be answered in the future. An official transparency label and verifiable transparency criteria could provide a remedy here. The transparency of AI systems is therefore a double-edged sword and should hence be used with caution. The goals and problems are sometimes contradictory and cannot be solved simultaneously. Answering the key questions "What information does a stakeholder need to make a decision?" and "What information is not relevant?" can be helpful. Similar to the EU General Data Protection Regulation (GDPR) the principle of data minimisation is also recommended here, which is answered separately for each individual use case: as much information as necessary, but no more than strictly required, should be disclosed. This "need-to-know" principle applies especially to safety-critical information. The goal should be an appropriate level of transparency, which is sufficient, and at the same time aspects such as security should not be too disadvantaged.

4 Conclusions

Due to the black box properties of many AI systems, data and information are processed in a way that is not transparent to users leading to an unverifiable decision being produced. A lack of knowledge about the AI system goes hand in hand with a lack of traceability and verifiability of system outputs. It is difficult to assess whether system outputs are correct and appropriate. Similarly, questions about responsibility, liability or fairness cannot be answered, if there is a lack of information about the system and its ecosystem. Ultimately, non-transparent AI systems can lead to a loss of trust and a rejection of the system. The implementation of AI components into existing systems and the combination of different systems can also increase complexity and makes it even more difficult to access relevant information. The problems caused by a lack of system insight and a lack of information about the system are manifold and represent a major challenge. Transparency addresses this problem and aims to make AI systems more comprehensible by increasing accessibility to system information and enable a valid assessment of the systems. For these reasons, transparency plays a crucial role for all stakeholders of an AI system. The challenge is to serve all stakeholders with their individual and different transparency requirements. The overall

project and future work on the transparency of AI systems are aimed at all BSI stakeholders. The relevance of the topic for society as a user of the systems is reflected in the expected higher traceability, better protection against abuse, more valid acceptance and trustworthiness processes as well as a more binding legal responsibility. The transparency measures are intended to contribute directly to the empowerment of end users by increasing their trust and autonomy regarding the choice and use of AI systems. Overall, this empowerment of end users aims to democratise the use of AI systems. In addition, the work in the field of transparency and the derivation of concrete criteria and measures should contribute to the overarching goal of the trustworthy use of AI systems. For companies involved in the development of AI systems, the relevance of the topic and the observance of measures in the development and operation of AI systems should be accelerated. Guidelines and positions are to be made available as guidance for stakeholders from the economic environment who want to use third-party AI systems in their organisations or implement them in their systems and products. These guidelines are intended to make it easier for companies to identify suitable, secure and high-performance systems. This work is also intended to provide guidance for public authorities wishing to use AI systems. In addition to their own use, the daily new safety-relevant findings on AI systems, which have to be addressed, pose the challenge of ensuring a technically qualified and adequately positioned staffing level for public sector stakeholders and administrations. This and future work in the field of transparency can be used to facilitate and accelerate permanent and adequate (post-)training of staff. In addition, the establishment of transparency criteria hoped for by this and future work can facilitate the development of meaningful and reliable quality seals by public authorities. With regard to the expected further increasing prevalence and widespread roll-out of AI systems in many areas of life, the relevance of AI systems to society as a whole is steadily increasing. In order to be able to make competent and valid assessments of these systems in the future, the establishment of transparency criteria is indispensable. For providers and operators of certain AI systems – such as general-purpose AI systems or emotion recognition systems – transparency obligations are already defined in the EU AI Act (cf. Article 50 EU AI Act). These are one of the prerequisites for these systems to be marketed and used in the European Union. Transparency criteria can strengthen the autonomy of the stakeholders of an AI system by making informed decisions possible. Therefore, transparency can and should be considered from the outset (transparency by design).

References

1 ISO/IEC 22989:2022. Information technology-artificial intelligence-artificial intelligence concepts and terminology, July 2022.

2 BSI. Ai cloud service compliance criteria catalogue (aic4), 2021. URL: `https://www.bsi.bund.de`.

3 BSI. Safe, robust and comprehensible use of ai - problems, measures and needs for action, 2021. URL: `https://www.bsi.bund.de`.

4 Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. What's documented in ai? systematic analysis of 32k ai model cards. *CoRR*, February 2024. URL: `http://arxiv.org/abs/2402.05160`, `doi:10.48550/arXiv.2402.05160`.

5 OECD. Recommendation of the council on artificial intelligence, 2019. URL: `https://legalinstruments.oecd.org`.