Algorithmic Transparency:

The Cornerstone of Accountability in AI Systems

Author:

Ahmad El Mallah

Submission Date:

August 2024

Abstract

The foundation of ethical artificial intelligence (AI) is algorithmic transparency, which is essential to guaranteeing that AI systems function justly, responsibly, and in a way that fosters trust between users and society at large. The present study delves into the vital role of transparency in artificial intelligence, specifically emphasizing the drawbacks of widely used explainability methods as SHAP (Shapley Additive Explanations) scores. Although SHAP scores are frequently employed to assign weights to characteristics in AI models, they may offer deceptive justifications, which compromises the impartiality and responsibility of AI systems.

We propose that more formal, rigorous approaches to explainability are needed in addition to the existing emphasis on heuristic approaches like SHAP, based on findings from the paper ”Explainability is Not a Game”. These techniques are required to guarantee that AI systems follow the moral precepts of justice and fairness and to steer clear of the moral traps brought on by incorrect predictions. The ethical underpinnings of algorithmic transparency are also explored in this work via a combination of various kinds of ethical theories, such as utilitarianism, virtue ethics, justice ethics, and deontological ethics.

While utilitarianism emphasizes the wider social advantages of transparent AI systems, deontological ethics promotes the responsibility of transparency as a basic ethical imperative in AI. Virtue ethics promotes truthful and precise justifications while highlighting the significance of moral integrity in the advancement of AI technology. Justice ethics emphasizes how important it is for AI decision-making systems to be equitable, guaranteeing that each person is treated fairly.

We support a move toward formal explainability techniques that offer dependable and credible justifications in light of these moral issues, bringing AI operations into compliance with legal and ethical constraints. By doing this, we want to advance AI systems that respect the greatest ethical standards in addition to their performance, making society more equitable and responsible.

Contents

1 Introduction

Artificial intelligence (AI) has advanced so quickly that it has drastically changed a number of industries, including criminal justice, healthcare, and finance. The usage of AI systems has raised ethical questions, as these systems have a growing impact on important decision-making processes. Algorithmic transparency, or the clarity and openness with which AI system decision-making processes are conveyed to stakeholders, is one of the most urgent ethical challenges.

Algorithmic transparency is a basic ethical value that is necessary for the responsible deployment of AI systems, not only a technological need. By making sure AI systems are held responsible for their choices, transparency helps all parties involved—developers, users, and those impacted by the decisions made—to comprehend how these systems work. This knowledge is essential for spotting and preventing biases, mistakes, and unjust results. Without transparency, AI systems run the risk of turning into opaque "black boxes," where the reasoning behind choices is concealed, which might undermine public confidence and endanger both people and society as a whole.

A major obstacle to accomplishing algorithmic transparency is explainability, or the capacity to offer coherent and intelligible justifications for the choices made by AI systems. In high-stakes situations where decisions can have far-reaching effects, including healthcare, criminal justice, and finance, explainability is especially important. But it's frequently challenging to provide precise and understandable explanations due to the complexity of contemporary AI models, particularly those that rely on machine learning.[6]

Of all the techniques created to improve explainability, SHAP (Shapley Additive Explanations) scores have gained widespread traction. Based on cooperative game theory, SHAP scores aim to assign a feature's relative value in a model's decision-making process. The goal of SHAP scores is to increase the transparency and understandability of AI judgments by disclosing which factors have the most impact on a model's output[6].But even with their widespread use, SHAP scores have drawbacks. According to the paper "Explainability is Not a Game," SHAP scores might occasionally give false explanations by emphasizing unimportant aspects while ignoring important ones. This may result in skewed or inaccurate interpretations, which would eventually undermine the accountability and fairness of AI systems [4].

In this study, we demonstrate that whereas similar heuristic approaches like SHAP scores provide some explainability, they are not enough to guarantee genuine algorithmic transparency. Formal, rigorous methods of explainability must be used in conjunction with these heuristic approaches in order to properly address the ethical problems presented by AI. These approaches, which are based on logic and mathematical reasoning, provide explanations that are more dependable and trustworthy, lowering the possibility of erroneous interpretations and improving the general accountability and transparency of AI systems [6].

This argument's ethical underpinnings stem from a number of important ethical theories. According to deontological ethics, transparency is a duty owed to society by AI developers and users as it is necessary to uphold people's right to knowledge and guarantee that AI systems behave morally. Utilitarianism is in favor of openness because it increases society's well-being by building confidence and guaranteeing that AI systems provide positive results. The moral need of AI developers to seek truth and integrity in their work—including giving truthful justifications for AI decisions—is emphasized by virtue ethics. [1] Lastly, ethics of justice and fairness emphasize how crucial openness is to averting discriminating results and guaranteeing that AI systems treat everyone equally.

This study examines the function of algorithmic transparency in AI from a theoretical and practical standpoint, given the importance of these ethical standards. It analyzes the drawbacks of the explainability techniques that are already in use, including SHAP scores, and makes the case for the adoption of more formal, morally compliant methodologies. By offering a solid foundation for comprehending and improving transparency in AI systems, the article hopes to further the current conversation on AI ethics.

2 Importance of Algorithmic Transparency

2.1 Definition and Purpose

The term "algorithmic transparency" describes how transparent and easily understandable AI systems make the data, reasoning, and procedures they use to make decisions available. This openness is crucial for a number of reasons. First, it makes decision-making processes understandable to all parties involved, including users, developers, and others who are affected by AI choices. This knowledge is essential in areas like criminal justice, healthcare, and finance services where AI systems are used to make choices that have a big influence on people's lives. [6]

One way to stop and lessen biases that might be present in the algorithms is through transparency. Stakeholders can examine the inputs and outputs of AI systems, spot any biases, and seek to address them by making the decision-making process transparent. For example, the COMPAS recidivism algorithm has been criticized in the criminal justice setting for its lack of openness, raising questions about potential racial bias in the algorithm's projections [1]. The ethical concerns of opaque AI systems have been brought to light by this lack of openness, underscoring the need for more understandable and transparent algorithms.

Transparency also promotes confidence between consumers and AI developers. Users are more inclined to believe the conclusions of an AI system when they are aware of the decision-making process used. The broad use of AI technology depends on this trust, especially in delicate fields like healthcare where patients must have faith in the advice given by AI-driven diagnostic tools.[6]

2.2 Challenges in Transparency

Algorithmic transparency is important, yet difficult to achieve. The intrinsic complexity of contemporary AI models, particularly those based on deep learning, is one of the main obstacles. These models frequently serve as "black boxes," making it challenging for even the developers to completely comprehend the reasoning behind particular choices [4]. This opacity is caused by the complex, nonlinear interactions that AI systems describe as well as the large dimensionality of the incoming data. This makes it difficult to present these models in a way that is accurate and understandable to non-experts.

Finding a balance between the need to safeguard private information and transparency is another difficulty. Businesses that create AI systems frequently view the algorithms they use as important pieces of intellectual property. Over revealing the inner workings of these algorithms may jeopardize their competitive edge. For example, disclosing a model's internal workings might leave it vulnerable to reverse engineering or make it simpler for rivals to copy the technology. This puts the corporate necessity for secrecy and the ethical requirement for transparency at odds.

Furthermore, there's a chance that transparency increases might result in oversimplification. Simplifying explanations to make complicated models clear might be tempting, but doing so may leave stakeholders with inaccurate or missing information. This is especially troubling in situations when precise and comprehensive explanations are essential, such when making decisions in the legal or medical domains. For instance, the paper "Explainability is Not a Game" criticizes techniques such as SHAP scores, which, although meant to promote openness, can provide erroneous or too simplistic explanations for a model's behavior. [4]

It takes a sophisticated strategy to address these issues, striking a balance between the complexity of contemporary AI systems, the requirement for openness, and the security of confidential information. This might entail creating novel, rigorous, and understandable explainability techniques so that stakeholders can have confidence in and comprehend AI judgments without jeopardizing the models' integrity.

3 Explainability in AI: The Role and Risks of SHAP Scores

3.1 Overview of SHAP Scores

Shapley Additive Explanations, or SHAP, is a popular technique for deciphering intricate machine learning models. SHAP scores, which are based on Shapley values, a notion from cooperative game theory, are intended to equitably allocate the role of every feature in a model's decision-making procedure[3]. In machine learning, SHAP is especially helpful for comprehending and elucidating predictions in high-stakes domains like healthcare and finance since it helps to quantify the extent to which each attribute effects the model's output. [6]

For example, a machine learning model may be used in a medical application to forecast a patient's likelihood of contracting a specific illness. Which characteristics—such as age, health history, or lifestyle choices—had the biggest influence on that forecast may be shown by looking at SHAP scores. Because of this, SHAP is useful in situations where stakeholders must interpret specific forecasts in addition to comprehending the model's overall behavior. [4]

The distinguishing feature of SHAP above other explainability approaches is its capacity to offer local explanations that account for specific predictions. Because of this, SHAP is especially well-suited for industries that depend heavily on decisions, like the legal or financial sectors, where every choice must be made with reasoned clarity. [3]

3.2 Critique Based on *Explainability is Not a Game*

SHAP scores have drawbacks in spite of their transparency-enhancing advantages. One of the main criticisms of SHAP scores, according to [4], is that they might lead to erroneous interpretations. The SHAP values exhibit significant variability based on the particular dataset and model, potentially resulting in inaccurate attribute attributions. For instance, SHAP could undervalue the significance of important variables or overemphasize the importance of unimportant characteristics, misleading users about how the AI model actually operates.

SHAP's approach of evaluating feature relevance by deleting or modifying features might warp the logic of the model in intricate models with a large number of interacting features. When SHAP's explanation differs from the actual factors affecting the choice, this might lead to a misleading impression of transparency [4]. Marques-Silva and Huang contend further that there is a serious risk associated with this false transparency in areas such as criminal justice, where inaccurate explanations may lead to discriminatory practices or unjust treatment.

For example, it was discovered that SHAP scores in predictive police models incorrectly assigned a considerable weight to geographic parameters that had no relevant correlation with criminal behavior. Systemic disparities were strengthened as a result of discriminatory and unjust policing judgments [4]. Therefore, even if SHAP seems explainable, it doesn't always give complete transparency, which is necessary for morally sound and precise decision-making.

Even though, it's crucial to recognize the substantial drawbacks of SHAP ratings while also recognizing their potential for inaccurate interpretations in complex and highly dynamic models. In situations where there are fewer feature interactions and simpler predictive models, SHAP scores continue to be especially useful. In these situations, they offer quick, easily comprehensible findings that can successfully guide stakeholders toward features that require more in-depth research and inform preliminary studies. Therefore, the objective should be to clearly grasp where and how SHAP ratings can be applied effectively, using them as one part of a larger arsenal of interpretability tools to ensure thorough and morally sound explanations, rather than simply dismissing them completely.

3.3 Examples of Misleading SHAP Scores

There is a real-world risk of erroneous interpretations with SHAP ratings; this is not simply a theoretical concern. Using SHAP in strongly correlated feature models is one of the primary obstacles. Under such circumstances, SHAP may give weight to a characteristic that only marginally influences the result, resulting in erroneous inferences [4].

For instance, SHAP scores may suggest in credit scoring models that a person's geography plays a significant role in determining their creditworthiness. This gives rise to ethical issues regarding the possibility of biased lending practices, even if it may represent trends in the training data [5]. Should the explanation supplied by SHAP mirror preconceived notions ingrained in the data—such as socioeconomic correlations—then the transparency it purports to offer may inadvertently function to rationalize prejudiced judgments.

Similar to this, misleading correlations in the data may cause a patient mortality model in the healthcare industry to place disproportionate weight on the kind or location of the institution. Even if the true predictors are more strongly correlated with the patient's health status, the SHAP score may indicate that these characteristics have a considerable influence on the result [4]. These kinds of deceptive attributions draw attention to the risks associated with depending too much on SHAP scores for explainability without cross-checking them with other interpretability methods or domain expertise.

[4] advise against relying just on SHAP scores to determine explainability in their critique. They contend that in order to guarantee that explanations are truthful and morally sound, explainability techniques need to be put through a rigorous testing process and employed in conjunction with other strategies. In sensitive domains like criminal justice and healthcare, SHAP ratings may result in judgments that reinforce preexisting biases or provide unfair consequences if they are not carefully validated [2].

4 The Necessity of Algorithmic Transparency in Ensuring Accountability

4.1 Regulatory Frameworks

Not only is algorithmic openness morally necessary, but it is also required by law in many places. Regulations requiring openness in automated decision-making processes include the General Data Protection Regulation (GDPR) in the European Union. People have the right to know why choices made by automated systems were made, especially if those decisions affect them legally or in a way that is equally important, according to Article 22 of the GDPR [6]. This legal paradigm emphasizes how crucial openness is to guaranteeing that people are not exposed to arbitrary or opaque decision-making by artificial intelligence (AI) systems.

The notion of accountability, which mandates that enterprises not only develop transparent AI systems but also show that these systems function in accordance with legal and ethical requirements, is the foundation for the GDPR's emphasis on transparency. This is crucial because AI judgments can have significant and far-reaching effects in industries including banking, healthcare, and criminal justice [6].

Beyond GDPR, additional international legal regimes are starting to acknowledge the need for AI openness. For instance, the Algorithmic Accountability Act of 2019 was put out in the US to mandate that businesses evaluate the effects of their AI systems, taking into account their fairness and transparency. By putting transparency at the forefront of the requirements for using AI technology, these policies together advocate for more responsibility [6].

4.2 Issues with Current Practices

It is still difficult to achieve genuine algorithmic transparency in practice, even with legal regulations in place. The intrinsic complexity of contemporary AI systems, especially those built on deep learning and other cutting-edge machine learning methods, is one of the main problems. These systems frequently operate as "black boxes," meaning that even the creators may find it difficult to completely comprehend or justify the reasoning behind particular decisions made [4].

The reason this opacity is problematic is that it compromises the accountability that transparency is meant to uphold. Holding companies responsible for the results of AI systems becomes challenging if the decision-making processes of these systems are not easily understood. Users and stakeholders may believe they are at the whim of unintelligent algorithms as a result of this lack of clarity, which can breed mistrust [6].

Furthermore, explainability techniques that are already in use, such the application of SHAP ratings, frequently fall short of offering the degree of openness necessary to adhere to legal and ethical requirements. As covered in "Explainability is Not a Game," SHAP scores and related techniques can be useful, but they can also provide erroneous or simplistic explanations that fall short of capturing the complexity of AI decision-making [4]. When AI systems seem transparent but are actually still opaque in important ways, this might lead to a false feeling of security.

The conflict between the demand for openness and the security of confidential information is another important problem. Businesses that create AI technology frequently reserve the right to completely reveal the inner workings of these systems, viewing their algorithms as intellectual property [5]. In order to avoid disclosing sensitive information, businesses may simply offer incomplete or broad answers, which can result in a lack of openness. The commercial drive to preserve proprietary assets and the ethical demand for transparency clash in this case.

4.3 Case Studies

4.3.1 Positive Examples

Efforts to increase AI openness are admirable, despite persistent obstacles. The release of GPT-3 by OpenAI, which included with thorough documentation outlining the model's architecture, training data constraints, and possible hazards, is one noteworthy example [5]. This openness facilitates a more responsible and knowledgeable deployment by assisting users and researchers in comprehending the model's strengths and limitations.

IBM's AI Fairness 360 Toolkit is yet another example; it gives developers the means to identify and lessen bias in AI models. By enabling practitioners to systematically evaluate and address ethical challenges, this project promotes the creation of more equitable and transparent systems [2].

Furthermore, Google's Model Cards framework provides organized documentation for AI models that covers ethical issues, intended use cases, and performance metrics across demographic groups. Before deployment, stakeholders can evaluate model behavior critically and morally thanks to this standardized approach, which also improves openness and accountability.

4.3.2 Negative Examples

On the other hand, there are important instances where a lack of openness has raised serious moral questions. The U.S. criminal justice system's COMPAS recidivism prediction program is among the most talked-about examples. Due to the tool's private nature, defendants and legal teams were unable to access or comprehend the methodology used to calculate risk levels. When compared to white defendants with comparable histories, investigations showed that the algorithm disproportionately classified Black defendants as high risk [1]. This lack of openness impeded people's deontological right to know and protest decisions that impact their liberty and went against ethical norms of justice and fairness.

Facebook's ad distribution algorithm is another example; it has been demonstrated to reinforce societal biases even when marketers employ neutral targeting criteria. Men were preferentially shown employment advertisements for technical, high-paying positions, but women were more likely to see adverts for lower-paying positions [4]. Algorithmic optimization based on past engagement data was the source of this bias, which strengthened preexisting stereotypes. Users were not aware of how or why they were being targeted due to the opacity of these algorithmic determinations, which raised ethical questions about informed consent, autonomy, and discrimination.

5 Formal Explainability as a Path Forward

5.1 Formal vs. Informal Explainability

Supporters of informal explainability techniques, such SHAP scores and LIME, claim that these methods are useful options because of their ease of use, accessibility, and simplicity, especially in situations where quick insights are required. Without requiring in-depth technical knowledge, these techniques can give stakeholders clear explanations and intuitive visualizations, facilitating fast initial interpretation. In settings where stakeholders, such as legislators, medical professionals, or the general public, might not have technical knowledge in formal techniques, this kind of accessibility is extremely beneficial.

Nevertheless these benefits, depending only on informal methods could lead to interpretations that are insufficient or unduly straightforward, particularly in intricate and high-stakes situations. When fairness, ethical compliance, and accountability are crucial, formal explainability techniques provide rigor and consistency, as well as strong, verifiable explanations that can stand up to scrutiny.

The difference between formal and informal explainability becomes important in the pursuit of transparency in AI. Heuristics and approximations are used by informal explainability techniques like SHAP scores and LIME (Local Interpretable Model-agnostic Explanations) to shed light on how AI models decide. These techniques have become more well-liked due to their relative simplicity and use, but as Marques-Silva and Huang's assessment makes clear, they frequently fail to offer thorough and trustworthy explanations (2024). These unofficial approaches can provide simplistic explanations that fall short of capturing the actual complexity of AI systems, which raises ethical concerns and may result in misunderstandings [4].

Formal explainability, on the other hand, refers to the use of statistical, mathematical, and logical frameworks to accurate and verifiable explanations of AI decision-making processes. In order for an explanation to be systematically verified against the behavior of the model, it must be both intelligible and verifiable. This is the goal of formal techniques. Because of this rigor, formal explainability is especially useful in high-stakes industries like healthcare, banking, and criminal justice where AI judgments can have far-reaching effects [6].

AI's use of symbolic thinking is one instance of formal explainability. Formal logic is used in symbolic reasoning to express and analyze the information contained in AI models. Symbolic reasoning provides explanations that people can immediately understand and verify by breaking down complicated model predictions into a series of logical assertions. This method guarantees that the justifications are in line with the fundamental ideas controlling the AI system's behavior in addition to improving transparency [4].

Counterfactual explanations are formal techniques that investigate "what-if" scenarios in order to provide an explanation for AI judgments. A counterfactual explanation might, for example, demonstrate how the model's output could be affected by a little modification to the input data (such as changing a single feature). By giving a solid and understandable explanation, this approach gives stakeholders a clear knowledge of how many elements affect the choice [3]. Counterfactual explanations are a potent tool for accountability and transparency because they are especially good at showing the causal relationship between decisions.

5.2 The Importance of Formal Explainability

The desire to solve the shortcomings of informal approaches and guarantee that AI systems are genuinely transparent and responsible is what is driving the movement toward formal explainability. Compared to informal procedures, formal methods have several advantages. To start, their explanations are consistent. Formal techniques yield explanations that are consistent across circumstances and can be repeated and validated by others since they are based in logic and mathematics [4].

Second, by offering explanations that are both practical and comprehensible, formal explainability improves responsibility. Stakeholders are better able to contest or defend AI judgments when they have a comprehensive understanding of the reasoning behind the conclusions. In legal and regulatory environments, where AI choices frequently need to be defended in court or examined by regulatory agencies, this is especially crucial [6]. Formal explainability, for instance, guarantees that businesses may offer understandable and substantiable justifications for automated choices that affect people's rights in the context of GDPR compliance [6].

Formal explainability techniques are also necessary to address fairness and bias in AI systems. Formal approaches enable the discovery and correction of any biases incorporated in the model by offering thorough and comprehensive explanations of the decision-making process. This is essential to guaranteeing that AI systems function fairly and without discrimination, especially in delicate domains like healthcare and criminal justice [1].

Lastly, formal explainability contributes to AI system confidence. An AI system gains the trust of users and stakeholders when they can depend on the explanations it provides. For AI technology to be widely adopted and used ethically, this trust is essential. Marques-Silva and Huang (2024) [4] point out that confidence in AI conclusions is derived from both their comprehension and their conviction that they are supported by reason.

5.3 Future Directions

In the future, formal explainability technique development will be a crucial area of study for artificial intelligence. Integrating formal explainability with automated reasoning techniques is one intriguing avenue that might enable AI systems to provide explanations in real-time while guaranteeing their accuracy and comprehensibility. In this so, transparency would become more scalable and accessible by fusing the rigor of formal approaches with the useful advantages of automation [4].

Using hybrid models—which mix formal and informal explainability methods—is another area of investigation. These models would take use of the advantages of both techniques, utilizing informal ways to give extra insights that could be easier for users to understand and formal methods to provide robust, baseline explanations. These hybrid methods might ensure that AI systems are clear and easy to use by striking a compromise between the requirements for rigor and usability [3].

Furthermore, standardizing explainability techniques will become more and more important as AI develops. To guarantee that AI systems are held to uniform ethical and regulatory norms across many industries and countries, it will be crucial to develop universal standards and frameworks for explainability [6]. Additionally, standardization will make it easier to compare and validate various explainability strategies, fostering best practices and innovation in the industry.

To sum up, formal explainability is an important advancement in the quest for responsible and transparent AI. Formal approaches may improve trust, guarantee justice, and encourage the moral use of AI technology by overcoming the drawbacks of informal methods and offering thorough, verifiable explanations. The creation and acceptance of formal explainability techniques will be crucial in ensuring that AI systems function in a transparent and socially responsible manner as the field of AI develops.

6 Conclusion

Algorithmic openness is urgently needed as artificial intelligence begins to seep into different facets of society. This essay has examined the vital role that openness plays in guaranteeing that artificial intelligence (AI) systems are not just efficient but also moral, responsible, and reliable. The difficulty of explainability—the capacity to offer concise and intelligible justifications for AI judgments—lies at the core of this conversation.

More than just a technological need, algorithmic transparency is a basic ethical condition that supports the responsible application of AI systems. Without transparency, AI systems run the danger of turning into opaque ”black boxes,” with users and stakeholders unable to understand the reasoning behind choices. Significant ethical issues including prejudice, injustice, and a lack of responsibility can result from this opacity. As previously mentioned, laws like the General Data Protection Regulation (GDPR) stress the value of openness by mandating that people impacted by automated judgments be given the opportunity to request an explanation. The aforementioned legislative frameworks underscore the moral and legal requirements for AI transparency.

True transparency in AI, however, is difficult to achieve. Informal explainability techniques, like SHAP scores, can shed light on AI decision-making to some extent, but they are frequently unable to supply the thorough justifications required in high-stakes situations. These techniques, as criticized in ”Explainability is Not a Game,” can occasionally result in erroneous or simplistic explanations, which could hide the actual functions of AI systems and provide moral dilemmas.

Formal explainability, on the other hand, provides a more comprehensive method of transparency. Formal approaches offer accurate, verifiable explanations that can be systematically validated via the use of logical, mathematical, and statistical frameworks. These techniques are especially useful in fields like healthcare, banking, and criminal justice where AI choices have significant ramifications. Formal explainability promotes accountability by allowing stakeholders to contest or defend AI choices with logical, cogent arguments. It also improves transparency.

Furthermore, formal explainability is essential for mitigating prejudice and guaranteeing impartiality in AI systems. Formal approaches enable the discovery and correction of any biases incorporated in the model by offering thorough and comprehensive explanations of the decision-making process. This is necessary to ensure that AI systems are implemented in a way that is morally righteous and consistent with society ideals, as well as to preserve public confidence in these technologies.

Going forward, a crucial area of study in AI is the creation and use of formal explainability techniques. One potential avenue for future research is the combination of automated reasoning techniques with formal explainability, which would allow AI systems to provide explanations in real-time while preserving interpretability and correctness. Furthermore, investigating hybrid models that integrate the benefits of formal and informal approaches might provide a well-rounded strategy for explainability, increasing transparency while maintaining rigor.

Lastly, standardizing explainability procedures will become more and more important as AI develops. To guarantee that AI systems are held to uniform ethical and regulatory norms across various industries and jurisdictions, it will be imperative to establish universal standards and frameworks. By facilitating the assessment and comparison of various explainability methodologies, this standardization will also promote best practices and innovation in the sector.

To sum up, the quest for algorithmic transparency and formal explainability is a moral obligation as much as a technological one. For AI systems to be deployed ethically and to build confidence, it is imperative that they be made transparent, responsible, and equitable. As artificial intelligence (AI) grows more and more integrated into our daily lives, achieving transparency and explainability will be crucial to use AI to its fullest potential in a way that benefits everyone in society.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

[2] IBM Research. Ai fairness 360 toolkit, 2020. `https://aif360.mybluemix.net/`.

[3] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[4] João Marques-Silva and Xiaowei Huang. Explainability is not a game. *Communications of the ACM*, 67(7):66–74, 2024.

[5] OpenAI. Gpt-3 documentation, 2020. `https://beta.openai.com/docs/`.

[6] Giovanni Sartor and Mihalis Kritikos. *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence*. Publications Office of the European Union, 2020.