UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Ingeniería

https://revistas.udistrital.edu.co/index.php/reving/issue/view/1221

DOI: https://doi.org/10.14483/23448393.21583

Research

Explainable Artificial Intelligence as an Ethical Principle

Inteligencia artificial explicable como principio ético

Mario González-Arencibia[1], Hugo Ordoñez-Erazo[2], and Juan-Sebastián González-Sanabria[3]*

[1]Universidad de las Ciencias Informáticas, Habana, Cuba
[2]Universidad del Cauca
[3]Universidad Pedagógica y Tecnológica de Colombia

Abstract

Context: The advancement of artificial intelligence (AI) has brought numerous benefits in various fields. However, it also poses ethical challenges that must be addressed. One of these is the lack of explainability in AI systems, *i.e.*, the inability to understand how AI makes decisions or generates results. This raises questions about the transparency and accountability of these technologies. This lack of explainability hinders the understanding of how AI systems reach conclusions, which can lead to user distrust and affect the adoption of such technologies in critical sectors (*e.g.*, medicine or justice). In addition, there are ethical dilemmas regarding responsibility and bias in AI algorithms.
Method: Considering the above, there is a research gap related to studying the importance of explainable AI from an ethical point of view. The research question is *what is the ethical impact of the lack of explainability in AI systems and how can it be addressed?* The aim of this work is to understand the ethical implications of this issue and to propose methods for addressing it.
Results: Our findings reveal that the lack of explainability in AI systems can have negative consequences in terms of trust and accountability. Users can become frustrated by not understanding how a certain decision is made, potentially leading to mistrust of the technology. In addition, the lack of explainability makes it difficult to identify and correct biases in AI algorithms, which can perpetuate injustices and discrimination.
Conclusions: The main conclusion of this research is that AI must be ethically explainable in order to ensure transparency and accountability. It is necessary to develop tools and methodologies that allow understanding how AI systems work and how they make decisions. It is also important to foster multidisciplinary collaboration between experts in AI, ethics, and human rights to address this challenge comprehensively.

Keywords: artificial intelligence, AI, ethics, ethical principles, explainability, transparency

Article history

Received: 24th/Nov/2023
Modified: 17th/Jan/2024
Accepted: 9th/Apr/2024

Ing, vol. 29, no. 2, 2024, e21583

*Correspondence: juansebastian.gonzalez@uptc.edu.co

Resumen

Contexto: El avance de la inteligencia artificial (IA) ha traído numerosos beneficios en varios campos. Sin embargo, también plantea desafíos éticos que deben ser abordados. Uno de estos es la falta de explicabilidad en los sistemas de IA, *i.e.*, la incapacidad de entender cómo la IA toma decisiones o genera resultados. Esto plantea preguntas sobre la transparencia y la responsabilidad de estas tecnologías. Esta falta de explicabilidad limita la comprensión de la manera en que los sistemas de IA llegan a ciertas conclusiones, lo que puede llevar a la desconfianza de los usuarios y afectar la adopción de tales tecnologías en sectores críticos (*e.g.*, medicina o justicia). Además, existen dilemas éticos respecto a la responsabilidad y el sesgo en los algoritmos de IA.
Método: Considerando lo anterior, existe una brecha de investigación relacionada con estudiar la importancia de la IA explicable desde un punto de vista ético. La pregunta de investigación es *¿cuál es el impacto ético de la falta de explicabilidad en los sistemas de IA y cómo puede abordarse?* El objetivo de este trabajo es entender las implicaciones éticas de este problema y proponer métodos para abordarlo.
Resultados: Nuestros hallazgos revelan que la falta de explicabilidad en los sistemas de IA puede tener consecuencias negativas en términos de confianza y responsabilidad. Los usuarios pueden frustrarse por no entender cómo se toma una decisión determinada, lo que puede llevarlos a desconfiar de la tecnología. Además, la falta de explicabilidad dificulta la identificación y la corrección de sesgos en los algoritmos de IA, lo que puede perpetuar injusticias y discriminación.
Conclusiones: La principal conclusión de esta investigación es que la IA debe ser éticamente explicable para asegurar la transparencia y la responsabilidad. Es necesario desarrollar herramientas y metodologías que permitan entender cómo funcionan los sistemas de IA y cómo toman decisiones. También es importante fomentar la colaboración multidisciplinaria entre expertos en IA, ética y derechos humanos para abordar este desafío de manera integral.

Palabras clave: inteligencia artificial (IA), ética, explicabilidad, principios éticos, transparencia.

Table of contents

1. Introduction

Algorithms based on artificial intelligence (AI), especially those using deep neural networks, are transforming the way in which humans approach real-world tasks. Day by day, there is a significant increase in the use of machine learning (ML) algorithms to automate parts of the scientific, business, and social workflow (1). This is partly due to increasing research in a field of ML known as *deep learning* (DL), where thousands – or even billions – of neural parameters are trained to generalize when performing a particular task (2).

As AI systems become increasingly complex and powerful, a fundamental concern arises: *how can we understand and explain the decisions and actions of these systems?* Explainability is a fundamental ethical principle that seeks to address the challenge of understanding and explaining how AI systems make their decisions (3). It is seen as vital to ensuring transparency, distributed accountability, and trust in AI (4).

Explainable artificial intelligence (XAI) is a field of AI that seeks to generate explanations understandable to humans with regard to how AI makes decisions, rather than simply accepting answers or conclusions without clear explanations. The term was coined by the Defense Advanced Research Projects Agency (DARPA) in 2016 to address the lack of transparency and understanding in AI systems, especially in military applications (5). XAI promotes a set of technological tools and algorithms which can generate high-quality explanations that are interpretable and intuitive for humans (6,7).

However, in the ethical dimension of AI, explainability implies that a system must not only be able to explain how it reached a certain conclusion, but also ensure that it is not biased, discriminatory, and unfair. From this point of view, explainability constitutes a fundamental ethical principle.

There are several reasons why AI explainability can be considered to be a strong ethical principle. One of them is the need for responsibility and accountability. If an AI system makes decisions that significantly impact people's lives, it is essential for those affected to understand how these decisions were made and the logic behind them. This allows affected individuals to challenge unfair or erroneous decisions and hold those in charge of their implementation accountable.

AI explainability is important to avoid bias and discrimination. AI models often learn from historical datasets that may be biased and contain harmful patterns. If the results of an AI system are not explainable, it becomes difficult to identify and address these biases, which can result in biased decisions that perpetuate injustice and discrimination. By making decisions and the processes involved transparent, any bias or discrimination that may arise can be examined and corrected.

On the other hand, AI explainability is important to avoid negative impacts on areas such as privacy and security. If AI systems make decisions without clear explanations, there is a risk that incorrect conclusions may be drawn, or that the results may be abused to obtain sensitive information from individuals. By ensuring explainability, users can understand and evaluate how their data are being used and make informed decisions about their privacy.

This leads to gaps in understanding how transparency and explainability can affect public trust in AI systems and how issues such as algorithmic bias and injustice can be addressed. In the field of ethics in particular, there are disagreements on how to tackle these problems (1). According to (21), explainability should be a mandatory requirement for all AI systems, while other scholars advocate for more flexible approaches that balance explainability with other values such as efficiency or privacy (3).

This theoretical contradiction poses challenges in developing clear policies and standards for AI, and it extends to the methodological and practical levels. How can AI explainability be defined and measured? What are the best approaches to this effect? How can ethical concepts be translated into concrete practices and policies? These questions still need clear and consensual answers. Based on these concerns, this research aims to analyze the importance of explainable AI as an ethical principle in the development and application of technologies, striving to ensure transparency, accountability, and trust in AI systems.

Analyzing the importance of explainable AI as an ethical principle is fundamental in today's society. Understanding how decisions are made and how the underlying algorithms work provides transparency and accountability in their use. This is essential to ensuring that decisions made by AI are fair, ethical, and comprehensible to users and society at large.

2. Methodology

This work was carried out while following a qualitative approach, using document research techniques and content analysis. For data collection, a bibliographic review of academic databases was conducted (*i.e.*, Xplore, ACM Digital Library, Google Scholar, and arXiv), looking for relevant literature on XAI published in the last five years. The search keywords were .[ex]plicable artificial intelligence", ïnterpretable machine learning", and "transparency in AI".

A systematic review was conducted, obtaining a representative sample of the available literature. Keyword sampling and expert experience were applied, the former to search for documents specifically related to the ethical component of XAI, and the latter to identify leading institutions in the field, aiming to ensure the quality and relevance of the sample. 36 documents were selected, including scientific articles, technical reports, and case studies on the applications of XAI in different domains, all of them in English. As a selection criterion, the sources had to constitute a representative sample of the available literature on the ethical component of XAI.

The analysis was carried out using content coding. Key concepts, definitions, problems, solutions, and documented applications of XAI were identified. Data triangulation was used, contrasting the findings of different sources. Finally, integrative synthesis was employed to consolidate the results. The reported approaches were compared, and relationships between key concepts were established, thus obtaining an updated state of the art on XAI.

3. Results

3.1. AI explainability as an ethical principle

As an ethical principle, the aforementioned concept refers to the ability to understand and justify the decisions made and actions taken by AI systems. In an ethical context, it is critical to explain and understand how AI algorithms generate results, especially when they significantly impact people's lives. This is significant in the field of ethical AI because it allows stakeholders to assess the reliability and fairness of the decisions made by their systems.

XAI allows understanding how the data, algorithms, and models used to reach a certain conclusion have been interpreted. It allows users or affected parties to identify and rectify possible biases or discrimination inherent in AI processes. Moreover, it contributes to accountability and transparency in the development and use of AI systems, enabling developers, users, and other stakeholders to audit and understand AI decision-making processes while ensuring that systems follow some crucial standards, namely:

- *Transparency:* the algorithms and processes employed should be clear and understandable to users and stakeholders (9).
- *Justice:* XAI must be impartial and avoid any kind of bias or discrimination. This involves considering different user groups and ensuring that outcomes are equitable for all.
- *Protection of privacy and personal data:* XAI must respect user privacy and ensure data security (1). To comply with privacy regulations, appropriate measures and policies must be implemented.
- *Accountability* implies taking responsibility for any errors or damages that may arise due to the use of AI. It is important to implement mechanisms to correct and prevent these issues (10).
- *Collaboration and participation* are necessary to the development of XAI. This involves including ethics experts, researchers, users, and civil society organizations the process to ensure a more diverse view of ethical impacts and challenges.

3.1.1. Approaches

There are different approaches and types of XAI in the academic literature (11–14). These approaches include the use of rules and logic, *i.e.*, AI systems use logical rules and deductive reasoning algorithms to provide explanations based on logical inferences. Another perspective is the use of interpretable models, such as decision trees or linear regressions (15).

These intrinsically interpretable models allow AI systems to generate explanations based on their characteristics and coefficients. Moreover, these systems can also keep track of all inferences and reasoning steps taken during the decision-making process, allowing to generate retrospective explanations through inference logs.

In the case of neural networks (NNs), techniques such as the visualization of neural activations or feature elimination can be used to make systems more interpretable and explainable. Furthermore, there are approaches such as meta-explainers, which use separate AI systems to explain the decisions and actions taken by other systems.

To measure and attain explainability in AI systems, there are different approaches. One of them is the transparency approach, where the goal is for the system to provide clear and detailed information about the data used, the algorithms applied, and the processing steps implemented. This allows users to understand how decisions are reached and facilitates the identification and correction of possible errors or biases.

Some researchers are also exploring approaches based on simulation or case-level explanations, where the goal is for an AI system to explain its decisions through concrete examples by relating them to similar, previous cases (34).

3.1.2. Objective, features, and functions

According to the above, the main goal of XAI is to address the issue of opacity in AI systems, *i.e.*, the inability to understand how they work and why they make certain decisions. By providing clear and understandable explanations, the aim is to increase transparency and trust in AI systems, both for users and for decision-makers who rely on them.

The key characteristics of XAI systems include transparency, interpretability, and the ability to provide clear and understandable explanations (14, 16). Therefore, XAI systems must be able to show how they arrived at a specific decision or conclusion, what data they used, and how weights or rules were applied in the decision-making process.

These characteristics are especially relevant in critical applications such as medicine, where it is vital to understand how AI systems arrived at a diagnosis or treatment recommendation. They are also important in applications such as security and justice, where decisions can have a significant impact on people's lives.

The primary function of XAI is to provide explanations that are understandable to humans (17). This involves developing techniques that allow AI systems to elucidate their actions and provide evidence for them.

3.2. Ethical challenges and concerns associated with the lack of explainability

The lack of explainability in AI can lead to unfair and harmful decisions in several ways (4,9,10,16, 18,19), as shown in Fig. 1.

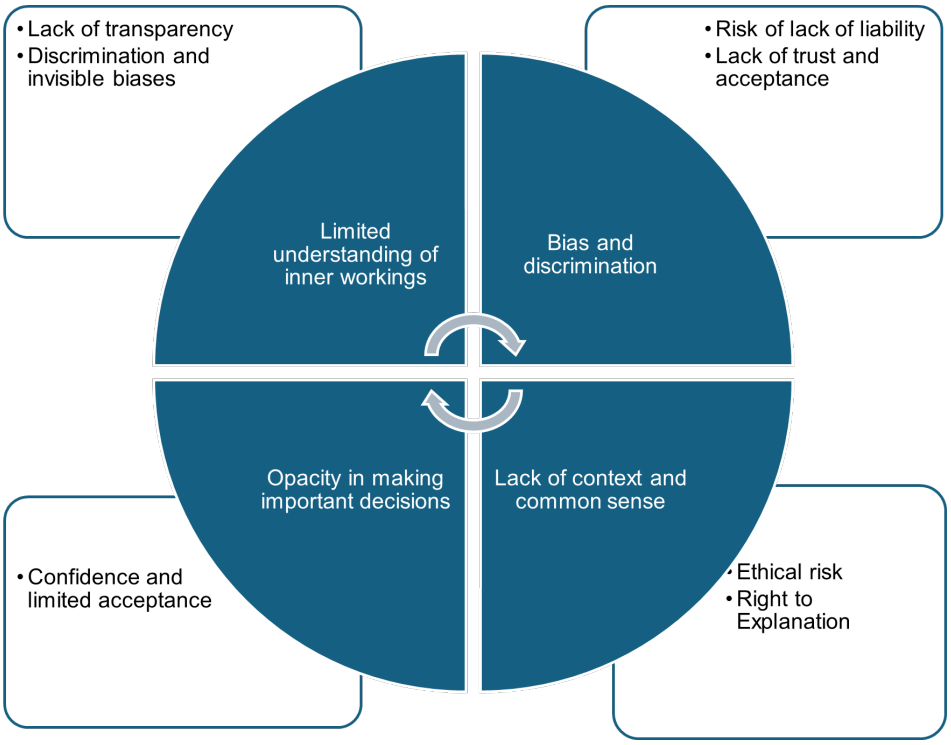

Figure 1. Challenges and ethical concerns related to the lack of explainability in AI

Fig. 1 presents different components related to the ethical challenges and concerns arising from the lack of AI explainability. One of these components is the lack of transparency, which implies the absence of a clear or understandable explanation of how a system or algorithm works. This lack of transparency can lead to discrimination and invisible biases, as the results obtained may disadvantage certain groups or individuals without an easily identifiable cause.

The lack of explainability poses a risk to accountability: if a certain decision or result is not understood, it is difficult to assign responsibilities in the case of issues or errors. This may also entail a lack of trust and acceptance, as people may be hesitant to use or trust opaque systems that cannot be adequately explained. If users do not have full knowledge about the internal functioning of a system or algorithm, their trust in it is likely to be limited. This poses some threats, including the ethical risk of making wrong or unjust decisions, as well as violating people's rights, such as the right to explanation.

Fig. 1 also refers to discrimination, which can be caused by the lack of transparency and explainability in important decision-making processes. This opacity in decision-making can have negative ethical consequences and affect people's rights.

AI has been implemented in vital areas such as medicine, criminal justice, and autonomous driving to improve the efficiency and accuracy of existing systems, with various effects and ethical concerns (10, 18, 20) (Table **I**).

Table I. Applications of AI in vital areas

Vital areas	Ethical concerns
Medical diagnostics	AI systems in medicine may face distrust due to lack of explainability. This raises ethical concerns regarding responsibility and accountability in the event of errors or harm to patients. In addition, discrimination and algorithmic bias have been observed in the development of these systems, such as racial biases in the diagnosis of diseases.
Criminal justice: recidivism prediction	The lack of AI explainability in criminal justice generates distrust and carries ethical risks. Algorithms are regarded as 'black boxes', which limits the ability to challenge AI-based decisions and can perpetuate discrimination and bias in the criminal justice system. This raises ethical concerns regarding justice, equity, and human rights, as algorithms may reproduce and amplify biases in the data used to train them.
Autonomous vehicles	In the context of autonomous vehicles, the lack of transparency carries ethical and legal risks. When an accident caused by an autonomous vehicle occurs, it is vital to be able to understand the decision that led to it. However, the algorithms for autonomous driving are often considered to be black boxes, and they hinder the establishment of liability and guilt. In the field of autonomous driving, ethical concerns encompass the safety of pedestrians and other drivers, as well as responsibility and accountability in accidents – *who* is responsible if an autonomous vehicle is involved in an accident?

One of the main ethical advantages of using XAI is confidence in their results (21). When systems are understandable and explainable, users can comprehend how their results were obtained and rely on their validity and accuracy. That trust is essential, especially in critical applications such as healthcare or justice, where AI-based decisions can have a significant impact.

AI explainability allows detecting and addressing possible biases in data and algorithms. AI systems may be influenced by biases in the training data or introduced by developers (18). The

ability to understand how decisions are made enables the identification and correction of unjust or discriminatory biases, ensuring impartiality and fairness in the decision-making process.

Moreover, explainable AI systems also promote ethical accountability. When they can understand and reflect upon the decisions made by AI tools, people can identify errors, biases, or unintended consequences (10). This facilitates the identification of potential ethical risks and the adoption of measures to mitigate them. Explainability is vital in several cases (22), as presented in Table **II**.

Table II. Cases where AI explainability is vital

Case	Explanation
Disease detection	In disease detection, explainability is vital to ensuring trust and acceptance of the results by healthcare professionals and patients. It allows clinicians to understand how certain conclusions were reached and which patient characteristics influenced the decision. In addition, explainability helps identify possible flaws in the models or data used, and it is relevant when deciding which algorithm to use in clinical practice.
Judicial decision-making	The explainability of AI systems in judicial decision-making is vital to ensuring fairness and transparency. It allows judges and lawyers to understand and question the basis of AI systems' recommendations or conclusions, especially in cases where there may be bias or discrimination. Explainability helps to identify and correct these biases, fostering impartiality and non-discrimination.
Cybersecurity	In cybersecurity, explainability is critical to understanding and addressing cyberthreats and cyberattacks. Security experts need to understand how an attack was made and what vulnerabilities were exploited in order to take corrective action and prevent future incidents. In addition, when it comes to the attribution of responsibility, explainability is needed to identify the culprit and take appropriate legal action.

There are several challenges associated with AI explainability. Some of them are shown in Table **III** (8,23).

Now, from an economic perspective, explainability can have significant implications. On the one hand, the lack of transparency can undermine consumers' trust in companies and in the use of their data. This can lead to reduced user engagement and, ultimately, to a negative impact on the revenue of businesses that rely on data collection and analysis. On the other hand, transparency can be a

Table III. Challenges associated with the explainability in AI

Challenge	Explanation
Model complexity	Many AI models, such as deep NNs, are highly complex and difficult to understand. Therefore, explaining how a model makes decisions and what features are most relevant to it is a complex task.
Privacy policy	To understand how an AI model makes decisions, it is necessary to analyze the training data used. However, these data may contain sensitive or confidential information. It is necessary to find a balance between explainability and data privacy.
Coherence and consistency	AI models can be inconsistent in their decisions, making it difficult to explain why a particular decision was made at a given time. Ensuring coherence and consistency in decisions is a challenge in ensuring explainability.
Bias and discrimination	AI models can be affected by biases in the training data, which can lead to discriminatory results. It is important to address these biases and ensure explainability, in order to identify and correct any potential discrimination.

competitive advantage for companies, as it can generate greater trust among consumers and therefore increase their loyalty and willingness to share data.

At the policy level, explainability can be vital to ensuring accountability and fair decision-making. If the algorithms used to make decisions – such as those related to loans, employment, or justice – are opaque or biased, there may be negative consequences for certain groups or individuals. Transparency in data collection and use, as well as accountability in cases of incorrect or biased decisions, are key elements to ensure fairness and avoid discrimination.

In cultural and ethical terms, explainability can influence how we perceive data use and automated decision-making. System opacity can lead to mistrust and concerns about a lack of control or knowledge with regard to how our data are used. Likewise, transparency can provide a greater sense of empowerment and autonomy by allowing to understand and question the decisions that affect us.

From a legal perspective, the absence explainability may affect the protection of privacy and individual rights. Transparency in data collection and use can help to ensure compliance with existing laws and regulations, such as the European Union's General Data Protection Regulation (GDPR). In addition, in cases of incorrect or biased decisions, liability may have legal implications and lead to lawsuits or penalties.

3.3. Approaches to improving explainability

So far, the question could be *how to improve AI explainability?* XAI is an important topic from an ethical standpoint. There are different approaches to improving explainability in AI models, including interpretable models, inference rules, and visualization methods (17,24,27).

One of the most common approaches to improving explainability is the use of interpretable models. These models, also known as *white-box models*, are those that can be easily understood and explained by humans. Examples of interpretable models include decision trees, linear regression, and decision rules. Another approach is inference rules. These rules allow establishing logical connections between premises and conclusions within a reasoning process. Inference rules can be used to explain predictions by AI models. For example, they can be used to explain why a certain decision was made by a recommendation system.

Visualization methods can also be used to improve the explainability of AI models. These methods allow graphically and comprehensibly representing the decision-making process of a given model. Some examples include flowcharts, tree diagrams, and heat maps. Visual representations help users to more intuitively understand how a certain prediction was reached.

The ethical importance of model interpretability lies in transparency and accountability when using algorithms and models in decision-making processes that can affect people. This includes a) understanding how a certain prediction or decision was reached, b) contributing to increased accountability in the organizations or institutions using AI systems, and c) facilitating trust in the technology. By understanding how algorithms make decisions, people can feel more secure and trust that relevant aspects have been considered, thus avoiding arbitrary decisions.

There are several reasons for the necessity of establishing standards and regulations that promote AI explainability.

Firstly, explainability is crucial for ensuring accountability. If a decision made by an AI system has a negative impact on a person, it is necessary to be able to evaluate and understand the process followed, in order to determine if there has been any bias or discrimination. Without explainability, AI systems cannot be effectively held accountable for their actions.

Secondly, explainability is essential for fostering public trust in AI. Many people are uncomfortable with relying on AI systems whose decisions cannot be understood or explained. The lack of transparency and clarity in AI decision-making can erode trust in this technology and generate resistance to its adoption.

AI explainability is important to upholding ethical principles and people's rights. Automated decisions made by AI systems can affect fundamental rights such as privacy, non-discrimination, and equal opportunities. If we cannot understand how these decisions are made, we cannot ensure that these rights and ethical principles are being respected.

3.4. Solutions to the ethical challenges posed by the lack of AI explainability

To address these challenges, several solutions have been proposed:
Techniques are being developed to enable a better interpretation and understanding of AI models. This includes methods such as the generation of explanations or justifications for decisions made by AI systems and the visualization of the influence of different characteristics on the predictions made. These techniques allow users and stakeholders to understand how a decision is made and to assess its validity and fairness.

Research into approaches to identify and mitigate biases in AI systems is ongoing. This includes systematically analyzing the data used to train the models, identifying potential biases, as well as implementing techniques to reduce their impact. AI model auditing and evaluation can help to ensure that decisions are fair and equitable for all groups involved.

Regulations and standards are being proposed to promote explainability in AI systems. These legal and ethical frameworks can establish clear requirements regarding the transparency and accountability of AI systems, in addition to ensuring fairness and non-discrimination in their implementation.

Approaches such as user-centered design and stakeholder engagement can ensure that decisions on the use of AI are transparent, and that ethical concerns are taken into account. Engaging people affected by AI decisions and enabling human feedback and control can help to avoid unfair or harmful outcomes.

3.5. Existing regulatory initiatives

This section outlines some current initiatives aimed at consigning XAI into valid regulations.

One of the main regulatory initiatives that promotes AI explainability is the European Union's General Data Protection Regulation (GDPR) (26). The GDPR states that individuals have the right to explanations about automated decisions that affect them, including algorithm-based decision-making processes. This regulation requires organizations to provide clear and understandable information about the criteria used to make automated decisions, as well as the consequences and scope of such decisions.

Another regulatory initiative is presented in *Big data: A report on algorithmic systems, opportunity, and civil rights*, a document published by the US White House in 2016. This report emphasizes the importance of explainability in AI systems, especially in critical areas such as healthcare and finance. It proposes regulatory measures that promote transparency, accountability, and explainability in the algorithms used in these sectors.

Another example is the California Consumer Privacy Act (US). This law requires AI systems used by government agencies to be transparent and provide clear and understandable explanations of how decisions affecting citizens are made.

Furthermore, the European proposal for a legal framework on AI (AI Act) aims to establish a comprehensive regulatory framework for AI in the European Union. It includes specific explainability requirements, such as the obligation to provide clear and understandable information about the functioning of AI systems and the expected consequences.

The Institute of Electrical and Electronics Engineers (IEEE) has also developed explainability principles (the IEEE Explainable and Trustworthy Artificial Intelligence initiative) that promote transparency and accountability in AI systems. These principles include generating understandable explanations and documenting the AI decision-making process.

Finally, the UK government has established the Artificial Intelligence Explainability Initiative to promote explainability in AI systems used in the public sector. This initiative involves developing standards and guidelines that foster transparency and allow understanding the decisions made by AI.

The effectiveness of these regulatory initiatives in promoting explainability can be assessed from different perspectives, *e.g.*, the impact they have had on public awareness and widespread concern about the explainability of AI systems. These initiatives have sparked debate and put explainability on the political and business agenda.

Overall, while existing regulatory initiatives represent an important step towards promoting AI explainability, there is a need for a multifaceted approach that combines regulation with technological research and involves multiple actors, such as data scientists, engineers, and policymakers.

3.6. Education and awareness: actions towards implementing training programs

Educating users, developers, and policy makers about this issue is critical to ensuring that AI systems are trustworthy, ethical, and accountable. It is imperative to educate users about AI explainability, as they must understand how such a system works and how decisions are made (1). This will allow them to assess whether these decisions are appropriate and ethical. In addition, users should be informed about the possible bias and discriminatory actions of AI systems and how they can be mitigated.

As for developers, it is critical for them to be aware of the importance of AI explainability from the design and development stages (13). They must acquire skills and knowledge to develop XAI systems. This involves using techniques and algorithms to track and explain the reasoning of their models. Developers must also be trained in ethics and accountability, in order to ensure that AI systems are used fairly and transparently.

Finally, policymakers have a critical role in promoting AI explainability. They must understand its importance and legislate accordingly (27). This implies establishing regulatory frameworks that require the explainability of AI systems in certain contexts, such as automated decision-making in the legal or healthcare fields. They should also encourage research and training and promote collaboration between the public and private sectors, in order to guarantee the responsible development of AI. To implement

training and outreach programs on XAI, different actions can be undertaken (Table **IV**).

Table IV. Actions for training and dissemination regarding AI explainability

Actions	Contents
Courses and workshops	Training programs addressing the fundamentals of XAI, existing techniques, and related ethical and legal challenges can be designed for users, developers, and policymakers.
Conferences and seminars	Educational materials can be developed, such as guides, manuals, or infographics that explain the concepts and techniques related to XAI in a clear and accessible way. These resources could be distributed online or in print.
Research promotion	Research into XAI could be encouraged by funding research projects and programs. This will facilitate the development of more explainable AI techniques and models.
Interdisciplinary cooperation	Collaboration between experts in AI, ethics, law, and other relevant disciplines can be promoted to comprehensively address the implications of XAI.

3.7. The need for multidisciplinary approaches

XAI demands a multidisciplinary approach due to its complexity and its implications for different aspects of society. In this vein,

- *Sociology* can analyze the impact of AI on social structures by studying how it influences work organization, income distribution, and new inequalities. In addition, it can investigate the effects of AI on social interactions and the formation of virtual communities (28).
- *Psychology* can contribute to understanding how people interact and relate to AI systems. Studies on users' perception of AI and how it affects their trust and acceptance could be carried out. Likewise, the possible biases and prejudices of AI models could be analyzed (1).
- *Ethics and law* are fundamental to establishing the principles and regulations guiding the development and use of AI (10, 18, 29). Aspects such as privacy, responsibility, and ethical values must be considered in decision-making processes related to this type of systems.
- *Economics* can assess the economic impacts of AI, focusing on growth and productivity. It can also evaluate potential labor displacements and the restructuring of industries. It is important to analyze how AI can drive innovation and generate economic opportunities, but also how it can lead to inequalities and challenges in the labor market (30).

- *Culture, history, and demographics* can provide a socio-cultural context for AI (19). These disciplines can analyze how these systems can reflect and influence the values, beliefs, and practices of a particular society, as well as the demographic changes that may arise due to their implementation.

3.8. Evaluation and follow-up

Ethical metrics and assessments to measure AI explainability are critical to ensuring transparency and reliability. Some proposals reported in the literature are presented below (18,31,32) (Table **V**).

Table V. Metrics and assessments

Metrics	Assessment
Explainability metrics	These metrics quantify an AI system's degree of explainability. Some common metrics include clarity, comprehensibility, and transparency. An example of this is measuring the amount of information provided by the system to justify its decisions, or its ability of the system to explain the rules or algorithms used.
Model interpretability assessment	This type of assessment focuses on the ability of an AI model to be interpreted and understood by users. Techniques such as data visualization, significant feature extraction, and pattern identification can be used to this effect.
Assessment of social and societal impact	This type of assessment focuses on the impact of an AI system on society in terms of equity, bias, or discrimination. Measures such as differential accuracy can be used in order to assess whether the system shows fair and equitable treatment towards different population groups.

It is important to mention that these metrics and evaluations must be applied ethically, considering all relevant values and ethical principles. Ethical monitoring and updating mechanisms are required to ensure that AI systems remain explainable as they evolve and are updated. Some proposed mechanisms include those shown in Table **VI** (10,21,25).

The importance of the elements in Table VI lies in their contribution to ensuring that AI systems are ethical and meet the established standards. For example, periodic AI auditing processes are essential for regularly evaluating the explainability and ethical impact of AI systems. This enables the identification of potential ethical issues and the timely implementation of corrective actions. Furthermore, this constant review helps to improve the understanding of how a system works, which in turn allows explaining its decisions and actions to users and ethics experts.

Table VI. Ethical requirements and indicators

Indicator	Ethical requirement
AI auditing processes	These processes should be carried out on a regular basis to ensure that AI systems meet the established ethical standards. This involves regularly reviewing and evaluating their explainability and ethical impact.
Continuous feedback	Feedback and communication channels should be established with users and ethics experts in order to gather feedback and concerns about the explainability and ethical impact of AI systems. This feedback can help identify potential issues and improve explainability.
Updated regulatory frameworks and policies	Up-to-date policies and regulatory frameworks that require AI explainability should be established. These policies should consider both technical and ethical aspects.

4. Conclusions

As an ethical principle of AI, explainability is vital due to several reasons:

- First, it promotes transparency and builds trust in how technology works. Users and stakeholders can understand how AI systems make decisions, avoiding the 'black-box' feeling and reducing uncertainty.
- Second, AI explainability makes it possible to ensure accountability and liability for systems and their creators. It is essential to be able to explain and justify the reasons behind the decisions made by AI systems, especially when they have an impact on people's lives. Explainability facilitates the identification of biases, errors, or bad practices. Thus, these issues can be corrected, or the actors involved can be held accountable.
- Third, explainability helps to identify and mitigate biases inherent in the training data or algorithms used in AI systems. Understanding how these biases influence system decisions allows taking steps to ensure fairness and impartiality in the field of AI.
- Fourth, when using AI, it is also critical to ensure compliance with ethical and legal standards. By understanding how decisions are made, one can assess whether relevant ethical and legal principles such as privacy, non-discrimination, and fairness are being met. In addition, explainability facilitates the auditing and monitoring of AI systems.
- Fifth, explainability contributes to improving the acceptance and adoption of AI by society. When people understand how AI works and why it makes certain decisions, they are more likely to trust the technology and feel comfortable using it.
- Finally, explainability reduces the perception of AI as a 'mysterious black box' and aids in overcoming cultural and trust barriers. In short, explainability is essential to ensuring the transparency, accountability, fairness, and acceptance of AI.

5. CRediT author statement

Mario González Arencibia: conceptualization, investigation, methodology, writing (original draft).
Hugo Ordoñez-Erazo: investigation, writing (review and editing).
Juan-Sebastián González-Sanabria: validation, visualization, writing (review and editing).

References

[1] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint*, 2017. arXiv:1702.08608. ↑3, 4, 5, 13, 14

[2] M. Huang, V. K. Singh, and A. Mittal, *Explainable AI: interpreting, explaining, and visualizing deep learning*. Berlin, Germany: Springer, 2020. ↑3

[3] A. Hanif, X. Zhang, and S. Wood, "A survey on explainable artificial intelligence techniques and challenges," in *IEEE 25th Int. Ent. Dist. Object Computing Workshop (EDOCW)*, 2021, pp. 81-89. https://doi.org/10.1109/EDOCW52865.2021.00036 ↑3, 4

[4] M. Coeckelbergh, "Artificial intelligence, responsibility attribution, and a relational justification of explainability," *Sci. Eng. Ethics*, vol. 26, no. 4, pp. 2051-2068, 2020. https://doi.org/10.1007/s11948-019-00146-8 ↑3, 7

[5] T. Izumo and Y. H. Weng, "Coarse ethics: How to ethically assess explainable artificial intelligence," *AI Ethics*, vol. 2, no. S1, pp. 1-13, 2021. https://doi.org/10.1007/s43681-021-00091-y ↑3

[6] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," *arXiv preprint*, 2020. arXiv:2006.11371. ↑3

[7] G. Adamson, "Ethics and the explainable artificial intelligence (XAI) movement," *TechRxiv*, Preprint, 2022. https://doi.org/10.36227/techrxiv.20439192.v1 ↑3

[8] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fus.*, vol. 58, pp. 82-115, 2019. https://doi.org/10.1016/j.inffus.2019.12.012 ↑9

[9] A. Weller and E. Almeida, "Principles of transparency, explainability, and interpretability in machine learning," *Cogn. Technol. Work*, vol. 3, pp. 1-14, 2020. ↑5, 7

[10] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389-399, 2019. https://doi.org/10.1038/s42256-019-0088-2 ↑5, 7, 8, 9, 14, 15

[11] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *IEEE 5th Int. Conf. Data Sci. Adv. Analytics*, 2018, pp. 80-89. https://doi.org/10.1109/DSAA.2018.00018 ↑5

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, 2016, pp. 1135-1144. https://doi.org/10.1145/2939672.2939778 ↑5

[13] A. Weller and H. Aljalbout, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *JAIR*, vol. 68, pp. 853-863, 2020. ↑5, 13

[14] Z. Lipton, "The mythos of model interpretability," *arXiv preprint*, 2018. arXiv:1606.03490. ↑5, 6

[15] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, 1-42, 2018. https://doi.org/10.1145/3236009 ↑5

[16] A. Weller and S. V. Albrecht, "Challenges for transparency," in *Proceedings of the AAAI/ACM Conf. AI Ethics Soc.*, 2019, pp. 351-357. ↑6, 7

[17] H. Nissenbaum, *Privacy in context: Technology, policy, and the integrity of social life.* Stanford, CA, USA: Stanford University Press, 2009. https://doi.org/10.1515/9780804772891 ↑6, 11

[18] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," BD&S, vol. 3, no. 2, e2053951716679679, 2016. https://doi.org/10.1177/2053951716679679 ↑7, 8, 14, 15

[19] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *BD&S*, vol. 3, no. 1, e2053951715622512, 2016. https://doi.org/10.1177/2053951715622512 ↑7, 15

[20] A. D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," *Ford. Law Rev.*, vol. 87, e1085, 2018. https://doi.org/10.2139/ssrn.3126971 ↑8

[21] A. Weller and L. Floridi, "AIEthics Manifesto," *Min. Mach.*, vol. 29, no. 3, pp. 371-413, 2019. ↑4, 8, 15

[22] V. Dignum, "Responsible artificial intelligence: How to develop and use AI in a responsible way," *ITU J.* (Geneva), vol. 1, no. 6, pp. 1-8, 2021. ↑9

[23] L. Floridi and M. Taddeo, "What is data ethics?" *Phil. Trans. R. Soc. A*, vol. 376, no. 2128, e20180083, 2018. ↑9

[24] Z. Lipton, "The mythos of model interpretability," *arXiv preprint*, 2016. arXiv:1606.03490. ↑11

[25] C. Molnar, *Interpretable machine learning*, 2019. [Online]. Available: https://christophm.github.io/interpretable-ml-book/ ↑15

[26] Unión Europea, *Reglamento general de protección de datos (GDPR)*, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/ ↑12

[27] European Commission, *Ethics guidelines for trustworthy AI*, 2019. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai ↑11, 13

[28] A. Brynjolfsson and A. McAfee, *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York, NY, USA: WW Norton & Company, 2014. ↑14

[29] B. Green and S. Hassan, "Explaining explainability: A roadmap of challenges and opportunities of machine learning interpretability," in *24th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, 2019, pp. 2952-2953. ↑14

[30] L. Floridi and J. W. Sanders, "On the morality of artificial agents," *Min. Mach.*, vol. 14, no. 3, pp. 349-379, 2004. https://doi.org/10.1023/B:MIND.0000035461.63578.9d ↑14

[31] A. B. Arrieta, V. Dignum, R. Ghaeini, A. López, V. Murdock, M. Osborne, and A. Rathke, "Transparent AI: An overview,"*Art. Inte.*, vol. 290, pp. 1-43, 2020. ↑15

[32] L. Liao, A. Anantharaman, and K. Pei, "On explaining individual predictions of machine learning models: An application to credit scoring," *arXiv preprint*, 2018. arXiv:1810.04076. ↑15

Mario González-Arencibia

He holds a Bachelor's degree in Political Economy from Universidad del Oriente, a Master's degree in International Economics from Universidad de La Habana, and a PhD in Economic Sciences from Universidad del Oriente. Additionally, he holds a PhD in Economics from Universidad Complutense de Madrid. With 38 years of experience in higher education, he currently serves as a full professor at Universidad de Ciencias Informáticas in Havana, Cuba. He has been involved in digital-era research for over 20 years and has authored over 200 publications, including more than 20 books and articles, covering areas such as economics, politics, and ethics.
Email: mgarencibia@uci.cu

Hugo Armando Ordoñez-Erazo

He is a systems engineer from Fundación Universitaria San Martín. He holds an MSc in Computing and a PhD in Telematics Engineering from Universidad del Cauca. He served as a full-time professor in both undergraduate and graduate programs, and as a researcher at the Department of Computer Science of Universidad de San Buenaventura. He currently works as a full-time professor at Universidad del Cauca. He is a co-author of three books, five JCR articles, and over thirty SJR and SciELO articles. He has supervised 11 graduate students. His research interests include data mining, data analysis, machine learning, information retrieval, and software engineering. His main international collaborations are with Spain, Mexico, Cuba, Ecuador, and Brazil. He is classified as an associate researcher by the Colombian Ministry of Science, Technology, and Innovation (Minciencias).
Email: hugoordonez@unicauca.edu.co

Juan Sebastián González-Sanabria

He's a Systems and Computing Engineer from Universidad Pedagógica y Tecnológica de Tunja (UPTC). He holds two specializations: one in Databases from UPTC and another one in Scientific and Technological Information Management from Universidad Nacional de La Plata. Additionally, he holds a Master's degree in Software and Information Systems from Universidad Internacional de La Rioja. He has attended training courses on Big Data (MIT), Pedagogy for Virtual Education (UTP), and Scientific Journal Editing (Diploma, Universidad de Negocios y Ciencias Sociales). He has taught courses in the Database and Research line, and he served as the director of UPTC's School of Systems and Computing Engineering in the 2014-2016 and 2021-2022 periods. He has been a Technology and Innovation Advisor for UPTC since 2017, and an editor of UPTC's Pensamiento y Acción Journal since 2019. He has authored more than a dozen presentations and scientific articles in various journals, mainly in the areas of data analysis and development.
Email: juansebastian.gonzalez@uptc.edu.co