Peertechz

ENGINEERING GROUP

ANNALS OF
Robotics and Automation OPEN ACCESS JOURNAL

DOI: https://dx.doi.org/10.17352/ara

Mini Review

Artificial intelligence: Explainability, ethical issues and bias

Alaa Marshan*

BSc, MSc, PhD, Lecturer and Researcher, Department of Computer, Brunel University London, College of Engineering, Design and Physical Sciences Science, Public University in Uxbridge, England

Received: 22 July, 2021
Accepted: 31 July, 2021
Published: 03 August, 2021

***Corresponding author:** Alaa Marshan, BSc, MSc, PhD, Lecturer and Researcher, Department of Computer, Brunel University London, College of Engineering, Design and Physical Sciences Science, Public University in Uxbridge, England, Tel: +44(0) 1895 266853; E-mail: alaa.marshan@brunel.ac.uk, alaa_marshan@hotmail.com

ORCID: https://orcid.org/0000-0001-6764-9160

Keywords: Artificial intelligence; Explainable AI; Black box problem; Ethical artificial intelligence; Data Visualisation

https://www.peertechzpublications.com

Check for updates

Abstract

There is no doubt that Artificial Intelligence (AI) is a topic that is attracting increasing attention from different communities, business and academic. AI adoption and implementation is faced by the difficulty of interpreting and trusting the outcomes of AI algorithms. Several ethical issues related to AI adoption such as algorithms and data bias are among the factors that hinder AI adoption by the business world. This study aims to highlight and classify the most important research that have been published on AI explainability and ethical issues. The main finding from this research refer to the necessity of forming proper comprehension of advantages and disadvantages offered by Explainable AI techniques. This work concludes that the interpretability of AI models needs to be investigated using innovative approaches such as data visualisation in conjunction with the requirements and constraints associated with data confidentiality and bias as well as the auditability, fairness and accountability of the AI model.

Introduction

Artificial Intelligence (AI) is a topic of growing significance for businesses as well as academic researchers. Its applications encompass many domains such as healthcare [1], finance [2] and manufacturing [3]. Artificial intelligence is represented in general purpose smart technologies that give the machines the ability to imitate human intelligence and perform complex tasks. Communicating with the machine using natural language, operating autonomous and adaptive assembly lines, and predicting supply chain demand stock market fluctuations are all examples of AI implementation in industrial contexts. Investigating the literature and despite AI's broad range of applications in several industry domains, till now, there is no agreement on a unified definition of artificial intelligence. For instance Simmons and Chappell [4], define AI as the term that "denotes behaviour of a machine which, if a human behaves in the same way, is considered intelligent". Also Kumar, et al. [5], describe AI as the "A system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation". Consequently, it can be inferred that AI is not confined to limited number of applications, but rather it is considered as a pervasive economic, societal, and organizational phenomenon.

Contrary to the point of view that sees AI dominating every aspect of our lives, AI does actually bring benefits to our daily lives by improving human health, safety and productivity [6]. The wide adoption of AI, however, didn't come without negative impacts and concerns regarding its explainability, bias and other ethical related issues. The ability to understand the results produced by AI-enabled systems is still under investigation; formulating the "Black Box" problem [7-9]. Many researchers have also highlighted that AI adoption and implementation can cause a number of ethical issues such as dealing with the consequences of the bias associated with AI algorithms. Bias is one of the major issues that AI suffers from, considering that it is embedded in the AI system we design and employed by governments and businesses to make decisions using biased-embedded AI models and data [7,10]. Considering the above issues, this paper aims to explore these issues and presents research directions that could support AI researchers in the future.

Citation: Marshan A (2021) Artificial intelligence: Explainability, ethical issues and bias. Ann Robot Automation 5(1): 034-037.
DOI: https://dx.doi.org/10.17352/ara.000011

In presenting this work, the rest of the paper is structured as follows: the second section will discuss the black box issue associated with the ability to interpret the results of AI algorithms. Section three discusses the ethical issues that accompany the adoption and implementation of AI; stressing the bias problem with AI implementation. Finally, section four will reflect on the preceded discussions; highlighting important future research directions.

Explaining the Black Box problem and the need for explainable AI

The terms Artificial Intelligence (AI) and Machine Learning (ML) are frequently used interchangeably in the literature [11]. Despite this overlapping, however, AI refers to the wider range of intelligent tasks such as resembling human cognitive ability to support learning, reasoning, and self-correction [12]. ML, on the other hand, represent a subset of AI algorithms that aim to detect patterns in the data that support classification and prediction tasks among other supervised and unsupervised learning task [8,13]. ML algorithms take input such as structured or unstructured, data and provide output such as prediction or classification results. In terms of explainability, these algorithms can be classified into those who have simple structure and generate inherently interpretable outcomes such as Decision Trees (DT), Support Vector Machine (SVM), Bayesian classifiers, additive models, and spare linear models. Their interpretability is due to simple restricted ML model's internal components such as weight of a feature in a linear model, a path in a decision tree, or a specific rule [8,14]. On the other hand, deep learning or Deep Neural Network (DNN) models with their variations (e.g., convolutional neural networks) are characterised with opaqueness or the lack of transparency of the way the results are generated; formulating the "Black Box" problem, which can be described as the lack of a clear mathematical mapping between the input and the outcome of the algorithms or the inability to traverse back from the outputs to the original data [9,13]. Such lack of understanding of how AI models work raise trust issues with the results generated from the DNN-based models.

Inevitably, calls for explainable AI started to mount; calling for accountable, auditable and transparent AI systems [15]. Explainable AI (XAI) is a reference to the post-hoc interpretability techniques that are capable of approximate deep-learning black-box models with simpler interpretable models [8,15]; converting the "Black Box" into "Glass Box" [8,16]. Additionally, Barredo Arrieta, et al. [15], Carvalho, et al. [13], Rai [8], have classified XAI models based on their generalisability and scope as well as the time of model creation: pre-model, in-model, and post-model [13]; resulting in the following classification:

- ***Model-specific global explanation***: Which aims to embed interpretability constraints (e.g., sparsity, monotonicity and semantic meaningfulness) into the structure and learning rules and parameters of deep learning models.

- ***Model-specific local explanation***: Which has the goal of providing an explanation for a specific instance in the deep learning model by utilising specific mechanisms to focus on specific features among high-dimensional input.

- ***Model-agnostic global explanation***: Which seeks to develop interpretable global alternative models that can map the input of the model to its output based on the association between them and the importance of the features utilised by the original "Black Box" model. For instance, approximating deep learning model with an interpretable decision tree that resemble the IF-THEN logic would offer a justification of the relative importance of the factors that affect customer response to marketing campaign.

Model-agnostic local explanation: Which has the objective generating model-agnostic explanations for a specific instance or for the vicinity of a specific instance. In this category, Ribeiro, et al. [16] have developed Local Interpretable Model-Agnostic Explanation (LIME) technique that generate an explanation of a model's behaviour in the neighbourhood of an instance.

Reviewing the literature shows that many researchers have proposed several models and framework to address the issue of AI explainability (see for instance Ribeiro, et al. [17] and Zednik [18]. Despite these efforts, however, it can be concluded that interpretability is a subjective concept that depends on its accuracy, understandability and efficiency, and, thus, hard to formalize in a way that fit all stakeholders involved in the implementation and interpretation of the results of an AI system [13].

Ethical issues with artificial intelligence

Connected to the transparency and accountability of ML algorithms, ethical AI is another major concern that attracts attention from researchers in AI domain who argue that ethical decisions should be one of the main drivers in AI development and adoption [19]. Researchers reported several cases where AI algorithms have demonstrated racial bias [7,13], such as imposing stricter jail sentences on black defendants [20] or demonstrating racial discrimination against non-white mortgage applicants [21]. Driverless cars are one of the prominent examples that urge researchers to call for principles that govern how AI make decisions since people's lives depend on them [19]. Moreover, the lack of transparency and accountability as well as the systematic invasion of people's privacy are other examples that highlight the issues and the need for ethical AI [22].

The advocates for ethical AI argue that for AI models to be Responsible, these algorithms and models should consider fairness, transparency, and privacy main components of their design [15]. In accordance with this goal, Leslie [22] discusses the FAST principles that should be considered while developing AI project. These principles represent fairness, accountability, sustainability and transparency. Fairness relates to algorithms as well as data and pertain to features of humans must be designed to meet the discriminatory non-harm. Accountability is concerned with developing AI systems that are that can answer questionable decisions generated by

Citation: Marshan A (2021) Artificial intelligence: Explainability, ethical issues and bias. Ann Robot Automation 5(1): 034-037.
DOI: https://dx.doi.org/10.17352/ara.000011

the AI algorithms. Sustainability is the principle that ensure that the AI-enabled systems have transformative effects on individuals and society. Finally, transparency offers the bases for the AI system to explain, in a simple language, the factors that were considered while behaving in a specific way, and to justify the ethical permissibility, the discriminatory non-harm and the public trustworthiness of the outcomes and the process behind them [22]. These principles illustrate how AI model interpretability must be addressed while considering requirements and constraints related to data privacy, model confidentiality, fairness and accountability. It is argued that by Barredo Arrieta, et al. [15] that in order to achieve a responsible development, adoption and implementation of AI methods by developers and organisations these principles must be studied jointly.

Conclusion and future work

The term Artificial Intelligence (AI), is credited to John McCarthy who coined this name in the mid-1950s when he and other twentieth century pioneers of AI – such as Marvin Minsky, Herbert Simon, Alan Turing, Allen Newell and John Clifford Shaw – paved the path to technologies that could emulate aspects of human intelligence, but at magnitudes of speed, resilience, reach and processing power beyond human capability [23]. It is increasingly being perceived as a possible panacea to address some of the world's most challenging social problems [24]. Governmental organisations and technology giants, such as Google and Microsoft, foster the perception of a future incorporating benevolent AI. Extending beyond the needs of society, organisations large and small are exploring and exploiting the potential of their data and AI to foster innovation and deliver value. This optimistic perspective, however, opposes the innate risk of AI, which like many human inventions has the potential to reflect the darker side of humanity.

Exploring the nature of how AI algorithms work as well as the potential for and the existence of 'dark' characteristics of this non-sentient emulation of human intelligence, which is not necessarily encumbered by the ethical and moral constraints of its sentient creators, is an important mission that researchers need to focus on. Inherited in its understandability, data visualisation could be one possible approach to provide better understanding of how AI and more specifically deep learning models work and justify the results emerging from these "black box" algorithms to address the need for the "Glass Box". Furthermore, the argue here is that AI, in its multitude of physical and virtual manifestations and with its preternatural capabilities and questionable transparency, has the exponential capacity for darkness, even when designed for good; especially if the design or use of the intelligence artefact - that is AI, is imbued with any amoral or immoral values and intent its makers foster. Considering the previous discussion, this research is motivated by the question of: what effect does artificial intelligence have if its explainability as well as ethical issues are not considered? The aim of the study is to motivate the researchers in AI domain to uncover the nature of AI and explore its negative impacts on organisations and society.

References

1. Hanson CW, Marshall BE (2001) Artificial intelligence applications in the intensive care unit. Critical Care Medicine 29: 427–435. **Link:** https://bit.ly/2V6dKzT

2. Bahrammirzaee A (2010) A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. Neural Computing and Applications 19: 1165–1195. **Link:** https://bit.ly/3ldYSKi

3. LI B, Hou B, Yu W, Lu X, Yang C (2017) Applications of artificial intelligence in intelligent manufacturing: a review. Frontiers of Information Technology and Electronic Engineering 18: 86–96. **Link:** https://bit.ly/3rKyh94

4. Simmons ASAB, Chappell SG (1988) Artificial Intelligence-Definition and Practice 13: 14–42.

5. Kumar V, Rajan B, Venkatesan R, Lecinski J (2019) Understanding the role of artificial intelligence in personalized engagement marketing. California Management Review 61: 135-155. **Link:** https://bit.ly/3xh40Qa

6. Dignum V (2018) Ethics in artificial intelligence: introduction to the special issue. Ethics and Information Technology 20: 1–3. **Link:** https://bit.ly/3rRUO3W

7. Favaretto M, De Clercq E, Elger BS (2019) Big Data and discrimination: perils, promises and solutions. A systematic review. Journal of Big Data 6. **Link:** https://bit.ly/3BZcmzw

8. Rai A (2020) Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science 48: 137–141. **Link:** https://bit.ly/3xh3x0m

9. Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, et al. (2019) Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11839 LNAI, 563–574. **Link:** https://bit.ly/3lh6uM3

10. Wigan MR, Clarke R (2013) Big data's big unintended consequences. Computer 46: 46–53. **Link:** https://bit.ly/3ja30bA

11. Agrawal A, Gans J, Goldfarb A (2019) Artificial Intelligence and Behavioral Economics. Economics of Artificial Intelligence 587–610. **Link:** https://bit.ly/3faSrEb

12. Beetz M, Buss M, Wollherr D (2007) Cognitive technical systems - What is the role of artificial intelligence? Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 4667 LNAI 19–42. **Link:** https://bit.ly/3zXdwcW

13. Carvalho DV, Pereira EM, Cardoso JS (2019) Machine learning interpretability: A survey on methods and metrics. Electronics (Switzerland) 8: 1–34. **Link:** https://bit.ly/3ikxg4F

14. Doshi-Velez F, Kim B (2017) Towards A Rigorous Science of Interpretable Machine Learning, (Ml) 1–13. **Link:** https://bit.ly/3rJzehO

15. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, et al. (2020) Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58: 82–115. **Link:** https://bit.ly/2TLdMfX

16. Holzinger A (2018) Explainable AI (ex-AI). Informatik-Spektrum 41: 138–143. **Link:** https://bit.ly/37q1iO9

17. Ribeiro MT, Singh S, Guestrin C (2016) "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135–1144. **Link:** https://bit.ly/3rOhRMP

18. Zednik C (2019) Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. Philosophy and Technology 1–29. **Link:** https://bit.ly/3feNGJL

19. Etzioni A, Etzioni O (2017) Incorporating Ethics into Artificial Intelligence. Journal of Ethics 21: 403-418. **Link:** https://bit.ly/3ieJgUZ

20. Caplan R, Donovan J, Hanson L, Matthews J (2018) Algorithmic Accountability: A Primer. **Link:** https://bit.ly/3j04aXe

21. Bostrom N, Yudkowsky E (2014) The Ethics of Artificial Intelligence. The Cambridge Handbookof Artificial Intelligence. **Link:** https://bit.ly/3BSusmO

22. Leslie D (2019) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute 6: 97. **Link:** https://bit.ly/3C32aGc

23. Moor J, Minsky M, Shannon C (2006) Artificial Intelligence Conference : The Next Fifty Years. AI Magazine 27: 87–91.

24. Chui M, Harryson M, Manyika J, Roberts R, Chung R, et al. (2018) Applying AI for social-good: Discussion paper 52. **Link:** https://mck.co/3rKyJUO