Requirements for Trustworthy Artificial Intelligence – A review

Davinder Kaur, Suleyman Uslu and Arjan Durresi
Department of Computer and Information Science,
Indiana University and Purdue Universaity, IN, USA
{davikaur,suslu}@iu.edu, adurresi@iupui.edu

Abstract. The field of algorithmic decision-making particularly artificial intelligence has been drastically changing. With the availability of huge amount of data and increase in the processing power, AI systems have been used in vast number of high-stake applications. So, it becomes important to make these systems reliable and trustworthy. Different approaches have been proposed to make theses systems trustworthy. In this paper we have reviewed these approaches and summarized them based on the principles proposed by European Union for trustworthy AI. This review provides an overview of different principles that are important to make AI trustworthy.

1 Introduction

In today's world, algorithmic decision-making and artificial intelligence (AI) are playing crucial role in our day to day lives. The area of automated decision-making using machines is not new. However, decision-making now days is highly data driven and complex. The decisions made by machines leave a profound impact on our society. To give an estimate about the impact, International Data Corporation (IDC) estimates the spending on AI systems will reach $97.9 billion dollars in 2023 and there will be 28.4% increase over the period of 2018-2023 [1]. These numbers show how AI is impacting our society by making decisions in almost every aspect of our life.

Different kinds of statistical tools, artificial intelligence algorithms and machine learning models are used to make decisions in all kind of applications such as healthcare, government, business, judicial and political spheres. These advancements in decision making led to fast growth in every sector. Now decisions made by artificial intelligence and machine learning algorithms can beat some of the best human players, serves as our personal assistant, used in medical diagnostics, used by companies for automated customer support and much more. With enormous applications and their impact, it is especially important to make sure that all these systems on which we are relying on so much are reliable and trustworthy.

Now AI has the power to analyze huge amount of data and make predictions based on that. However, these systems are so complex and opaque that it is difficult to judge and interpret their decisions as fair and trustworthy. And there are no set standards or mechanisms established to govern and test these systems. It is found out that these systems could behave unfair and can lead to dangerous consequences. Recidivism algorithm used across the country has been shown to be biased against black defendants [2]. Recruitment algorithm used by a big corporate company was found to be biased against women [3] and many more. These examples show that decisions made by machines can be rogue and can have life-critical consequences. So, it is important to design, develop, implement and oversee these systems very carefully.

With the growing need to make these systems reliable and trustworthy, researchers have proposed different solutions. Some have proposed by making data used for training AI systems unbiased, some proposed explainable and interpretable methods that will make the AI systems easy to understand and interpret by the users. Some researches have proposed overseeing methods to keep a check on AI systems and other researches have proposed methods that enable collaborative intelligence using both human and machines in decision making. All these proposed solutions involve human at different levels of AI lifecycle. These solutions have one common objective which is ensuring that AI systems should behave as promised and to create a notion of trust towards AI among its users.

In this paper we have discussed different aspects that are important to make AI decisions acceptable and trustworthy, policies and guidelines required to govern the working of these systems and how human intervention is important in this changing era of AI. This paper is organized as follows. Section 2 presents the foundational concepts and preliminary background work in the field of trustworthy AI. Section 3 presents the review of the latest developments in the field of Trustworthy AI. Section 4 discusses technical challenges and future directions. Finally, section 5 concludes the paper.

2 Background and Foundational Concepts

In this section we have discussed problems with traditional AI, key concepts of AI. This section also discusses about the key principles and guidelines that should be considered while designing, developing, implementing and overseeing the system.

2.1 Need of Trustworthy AI

The field of AI has major impact on our day to day lives. With the availability of huge amount of data, high computational power and efficient algorithms, AI has given us many useful solutions that benefit our society. However, with so many benefits AI also raises some concerns. With all these advancements, AI has become complex for the human to understand and control it. There is a need of mechanism to oversee the decisions made by AI to be trustworthy and within the ethnic guidelines. This is only possible if the machines or algorithms making these decisions are fair, understandable by designers designing them, users using them and policy makers making laws to

govern them. All these concerns related to AI creates a fear among users which in turn decrease the trust on the system.
Before looking at the guidelines proposed for making AI trustworthy, let us look at the problems and risks related to present AI systems. [4] Once Stephen Hawking said that "AI impact can be cataclysms unless its rapid development is controlled". AI systems can be dangerous and harmful if strict measures are not taken in designing, developing, implementing and overseeing them. In today's world, almost all the sectors are utilizing the superpowers of AI systems in decision making and in analyzing huge amount of data. But these superpowers of AI not always yield good results. Lot of these AI systems failed and showed dangerous consequences. For example, self-driving car killed a pedestrian because its self-driving system decided not to take any action after detecting pedestrian on the road [5]. AI chat-bot become racist after being corrupted with twitter trolls [6]. COMPAS recidivism algorithm used by judges across the nation has showed biased against black people [2]. These are some of the examples that shows how AI can be untrustworthy and dangerous if their development is not controlled. So, it is important to make sure that these AI systems do not cause any kind of harm to the mankind.

2.2 Requirements to make AI Trustworthy

Artificial intelligence is used to make decisions in high stake applications like healthcare, transportation, judicial system and many more. With the increase in the use of AI systems in decision making it becomes very important to develop guidelines and policies that ensures that AI will not cause any intentional or unintentional harm both to the society and the users using it. AI is designed by us and its our responsibility to make sure it is only for good [7]. Several researchers and experts in this field have proposed different guidelines and policies to make AI trustworthy. [8] European union proposed four ethnic principles (respect of human autonomy, prevention of harm, fairness, explicability) and seven key requirements(human agency and oversight, technical robustness, privacy and data governance, transparency, diversity and fairness, societal well-being and accountability) to make AI trustworthy.[9] considers explainability, integrity, conscious development, reproducibility and regulations important to make AI trustworthy. [10] did a review on all guidelines proposed by different organizations and research institutes to make AI trustworthy. They said despite of so many guidelines available, there is a difficulty in coming to the consensus about what properties make AI ethical and trustworthy. Following are the properties that are important to make AI system trustworthy:

Accuracy and Robustness: Accuracy of the model refers to the model ability to correctly predict the outcomes by generating less false positives and false negatives. Robustness refers to the model ability to perform accurately under uncertain conditions.

Non-Discrimination: Non-Discrimination refers to the model ability to treat all the users equally without discriminating against any section of the society. This means absence of any type of bias and discrimination.

Explainability: Explainability of the model enable the users of the model to correctly understand the working of the model. This property facilitates users to correctly predict the outcomes for given input and the reasons that could lead to model failure.

Transparency: Transparency of the model provides a clear picture of the model to the users. It allow users to clearly understand the model by seeing whether the model have been tested or not, what criteria it has been tested on, if the input output of the model make sense to the users and if the users of the system clearly understand the decisions made by the model.

Accountability: Accountability of the model refers to the model ability to justify the decisions made by it to the users of the system. This includes taking responsibility of all the decisions made whether they are good or caused some errors and unexpected results.

Integrity: Integrity of the model defines that the model should output results or make decisions within set parameters. These parameters can be operational, ethical or technical and can be different for different applications.

Reproducibility: Reproducibility of the model ensures that all the decisions made by the system can be reproduced if the same input parameters and conditions are provided to the system.

Privacy: Privacy of the model means that the model should protect the data on which it is trained on and the identity of the users using it.

Security: Security of the model makes sure that the model is secure from outside attacks, that can change and modify the decisions made by the system.

Regulations: Government and policy makers should develop laws and guidelines to govern the development and working of AI systems.

Human Agency and Oversight: This is the most important property that enforces that AI system should always be in control of humans to prevent harm.

3 Review

Researchers have proposed vast number of solutions focusing on different properties of trustworthy AI. In this review we have mapped these properties to the four principles(principle of respect for human autonomy, principle of prevention of harm, principle of fairness and principle of explicability) introduced by European Union[11] to make AI ethical, lawful and robust. Figure 1 shows this mapping of properties to the principles of trustworthy AI.

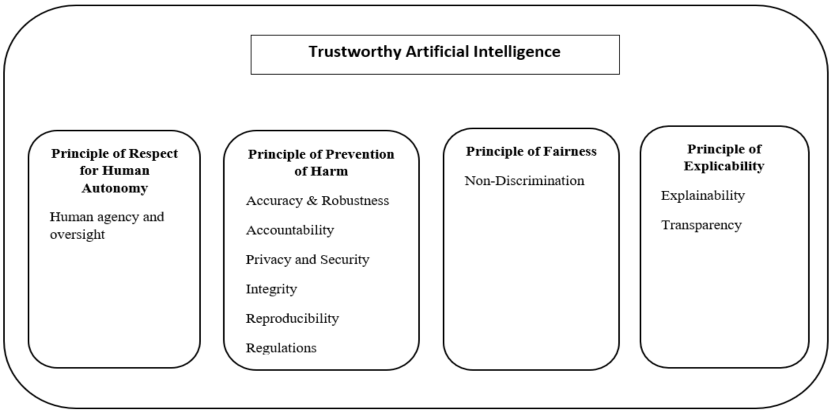
Trustworthy Artificial Intelligence
Principle of Respect for Human Autonomy
Human agency and oversight
Principle of Prevention of Harm
Accuracy & Robustness
Accountability
Privacy and Security
Integrity
Reproducibility
Regulations
Principle of Fairness
Non-Discrimination
Principle of Explicability
Explainability
Transparency

Fig. 1. It shows the mapping of different properties to the principles of trustworthy AI.

3.1 The Principle of Respect for Human Autonomy

This is the most important principle to design trustworthy AI. It ensures that the AI systems should be designed to complement and empower human cognitive ability instead of replacing them. [12] This new era of AI requires collaborative thinking where humans and machines work together towards a common goal. Making humans and machines work together will help to reduce incorrect and undesired results and will help to avoid accidents. The designing of AI systems should be human centric which means that the humans should be involved at different levels of AI lifecycle [11]. They should be involved in planning, designing, development and oversight phase of the AI system based on the application requirements. Humans should be in the center to set limits, flag errors made by machines, override wrong decisions, help to improve AI system by providing feedback. Human involvement is important to keep machine decision within moral and ethical compliance. European Union [13] have proposed some guidelines of how humans should be involved in AI decision making:

- For high stake applications, decisions made by the AI systems should only become effective if it has been examined and authenticated by the human experts. For example, in the medical AI systems where a wrong decision could lead to dangerous consequences, doctors should validate the decision made by the AI system based on their expertise and experience before implementing them.
- For the applications that require AI decisions to be effective immediately, there should be a way by which humans can intervene to review the AI decision and if needed can override the decisions. For example, in the loan approval AI system if the application is rejected by the AI system, it should

be possible for the human/loan expert to review the application again and change the decision if needed.

- And humans should able to oversee the working of AI systems and able to stop and interfere the working if he/she thinks that the AI system is not working appropriately or the decisions made by the AI systems is not safe anymore. For example, in autonomous vehicles if some sensor failed or the vehicle is not operating properly driver should able to overtake the controls of the vehicle.

Several methods have been proposed for human-machine collaboration. These methods have showed how human involvement can increase the trust and accuracy of the AI systems. [14] proposed analyst-in-loop AI system for intrusion detection which take feedback from security analysts to decrease the false-negatives generated by the system. This system improves the detection rate of AI system by three times. [12] showed how AI is screening huge debit and credit card logs to flag questionable transactions that can be evaluated by humans.

Some researchers have proposed a collaboration mechanism for human- machine interactions based on trust framework [15]. [16][17][18] used human feedback based decision-making system for resource allocation in FEW (Food Energy Water) sector. [19] proposed fake user detection system in social networks taking in account community knowledge along with machine learning algorithm. [20] proposed a human-machine collaboration mechanism to govern the interaction between police and machine for crime hotspot detection to facilitate greater accuracy.

All these proposed solutions showed that by combining the superpowers of both humans and machines help to improve accuracy and decrease the harm caused by AI systems which in turn make the AI systems trustworthy. So, this principle enforces that AI systems should empower humans, not replace them.

3.2 The Principle of Prevention of Harm

This principle ensures that the AI system should not cause any unintentional or intentional harm to the humans and the society. It also guarantees that the AI systems operate in safe and secure environment without causing any kind of harm to anybody. AI systems should be reliable in decision making task. Lot of elements should be taken into consideration while designing and implementing AI systems so that they do not cause any type of harm and can behave reliably. These elements are discussed below:

Accuracy and Robustness: This property ensures that the AI systems should have high accuracy and are robust. The accuracy of the system should be above certain threshold for reliable decision making. The system should be robust, that is it should able to work in adversaries and able to handle errors. And the results or the decisions made by the AI systems should be reproducible if provided with same input and similar conditions. Researchers have proposed different methods to deal with adversaries and making AI system accurate. [21] proposed feature squeezing method to decrease the complexity of input space, hence making it less prone to adversaries.[22] proposed inputting adversarial examples in the training set to make system robust.

Accountability: This property deals with the state of being responsible and accountable for all the good and bad decisions made by AI systems [23]. As algorithms cannot take responsibility for their decisions, designers of the AI systems should take responsibility for their working by proper testing, auditing and overseeing framework. It also deals with designing of laws and regulations for the controlled development of AI systems. Several researchers have proposed different auditing and testing methods to prevent harm caused by them. [23] proposed an internal auditing framework for algorithmic auditing to keep check on the development life cycle of AI systems. This auditing can be done by experts on each step of the development process. This method will help to prevent and mitigate the harm before it even occurred. [24] explained the importance of community involvement in designing algorithms to address algorithmic harm.[25] explains different type of accountability and how different level of users can be accountable for the decisions made by the AI system. This paper discussed accountability based on socio-economic aspect of the society.

Privacy and Security: Privacy of the AI system deals with the protection of the data on which AI system is being trained on, identity of users using it, internal working of the system and intellectual property if known can lead to dangerous consequences. Security of the AI system deals with the protection of system from outside attacks that can interfere and disturb the working of the AI system. Different methods have been proposed to enforce privacy and security of the AI systems. [26] proposed a pipeline for data protection when two or more agencies are involved in development of AI system. [27] discuss different type of attacks that can happen on AI system and what measures should be taken to prevent them.[28] discusses different privacy laws for data protection in research.[29] proposed a method of ignoring and forgetting malicious data in training phase so that AI system can reclaim security when attacked.

All these properties will help to prevent the unintentional and intentional harm caused by the AI systems.

3.3 The Principle of Fairness

The principle of fairness makes sure that the decisions made by the AI systems should not be biased and discriminatory to any kind of users. As these AI systems have been used in wide range of applications, where a discriminatory behavior of the system will make the system unfair, hence decreasing the trust on the system. This principle ensures that the AI systems should treat all the users equally without favoring any particular section of the society. AI systems should be obligated to hold moral and ethnic values. These systems are supposed to ease the decision-making process but if not designed and implemented properly can lead to bias and unfairness. So, it is very important to make these systems fair and unbiased. Before looking at the solutions let us look at the reasons for unfairness and bias.

Different types of Bias: AI systems can suffer from different types of bias. [30][31] discussed different reasons for unfairness. One main reason of unfairness of AI system is if the data on which it is trained on is biased and crooked. That is if the data is not able to represent the clear picture of the reality. For example, [32] ImageNet dataset, which is widely used by computer vision community, does not have a fair distribution of the geodiversity of people hence causing bias in the system using it. Other reason for the algorithm to behave biased is if some underlying stereotypes are present in the data. For example, AI system makes a prediction that men are more suitable for engineering jobs than women because over all the years men have higher percentage in engineering jobs than women and this stereotype makes the training data biased. And other reason can be if the bias is introduced by the algorithm itself. This can happen when the algorithm is trying to maximize its accuracy over the training data. [33] There can be other reasons also from where bias can be introduced into the system like if people collecting the data, designing the system or interacting with the system are biased against particular section of the society. So, measures should be taken to make AI systems fair, ethical and inclusive.

To make AI systems fair and unbiased, several methods and techniques have been proposed. [34] proposed a test generation technique to detect all different combinations of input attributes that can discriminate any individual based on gender, race, ethnicity etc. [35] proposed third party rating system to detect bias using sets of biased and unbiased data. [36] proposed a subsampling technique that ensures that the sub samples used for training are both fair and diverse. [37] Facebook proposed data traceability technique to detect bias using radioactive data labeling. [38] proposed vetting of algorithms by multidisciplinary team of experts to make algorithms unbiased. [39] designed an open-source toolkit to use fairness metrics and algorithms in an industrial setting. All these proposed solutions help to ensure fairness in AI systems.

3.4 The principle of Explicability

The principle of explicability deals with providing explainability and interpretability to the opaque AI systems. With the increase in the complexity of the AI models, they have become black boxes which are difficult to understand and interpret. [40] This principle of explicability ensures that the working of these AI systems can be openly communicated with different stakeholders who are directly or indirectly affected by the decisions made by these AI systems. This principle makes the process of decision making transparent, hence increasing the trust of the users on the system. [41] It enable the users to correctly understand the reasons that lead to a particular decision. Explainability of the system will also help the policy makers to better understand the system to make appropriate laws, will help developers of the system to detect the reason for errors and making the system more accurate.

Different approaches have been proposed to make AI systems explainable and interpretable. Some approaches known as integrated approaches deal with integrating explanation mechanism into the AI development lifecycle, while other approaches are post-hoc approaches that treat AI systems as black boxes and build an interpretable model over it. Some explainability approaches are global approaches that deals with explaining the working of the whole model while other approaches are local approaches that deal with providing explanation of a particular decision made by the system. All these approaches have one thing in common that is making the AI system transparent and understandable by different type of users.

Lot of work have been done in the field of explainability and interpretability. Some of these approaches are discussed here. [42][43] proposed post-hoc method of explanation by building a proxy model on the top of machine learning model and providing explanations by highlighting the important input attributes that lead to a particular decision. [44] also proposed a post hoc method which generates diverse counterfactual explanations to provide explainability to the model. [45][46] approaches consider the internal working of the system which takes into account the internal representation of the AI system to provide explanations. [47] proposed an action-based influence model to provide explanations which is based on how humans do reasoning and provide explanations to real world problems. Other popular technique of providing explanations is through visualization, which highlight the area of the image that lead to model prediction [48][49].

All these principles of trustworthy AI if followed properly ensures the reliability of the AI system hence increasing the trust on the system.

4 Technical Challenges and Future Directions

Lot of research has been done and still going on to make AI systems trustworthy. There are some technical challenges that can hinder the development of trustworthy AI. One of the main challenges is the lack of clear requirements and standards for making AI trustworthy. The definition of the principles and the properties are still vague and there can be a conflict between the principles in various application domains [11]. For example, following the principle of explicability can disobey the principle of prevention of harm, as more interpretable and transparent the model is, more prone it is to outside attacks. Hence there should be a tradeoff between these principles based on the application requirements and there is a need for strict laws that can govern the working of AI systems. Another challenge is that one solution that worked for one problem may not works for another problem. For example, explanation provided to developers of the system may not make sense to the users with non-technical background. Hence more context specific solutions are needed. There is also a need to involve multi-disciplinary team of experts to develop AI systems. In a nutshell, this area of trustworthy AI is new, lot of research is still needed to make AI systems reliable and trustworthy.

5 Conclusion

With the increase in the adoption of artificial intelligence in various application domains, it becomes particularly important to make these systems reliable and trustworthy. Different types of approaches have been developed to make these systems accurate, robust, fair, explainable and safe. In this review we have summarized these approaches and provided some future directions.

Acknowledgments. This work was partially supported by the National Science Foundation under Grant No. 1547411 and by the U.S. Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) (Award Number 2017-67003-26057) via an interagency partnership between USDA-NIFA and the National Science Foundation (NSF) on the research program Innovations at the Nexus of Food, Energy and Water Systems.

References

1. International Data Corporation IDC. (2019) “Worldwide Spending on Artificial Intelligence Systems Will Be Nearly $98 Billion in 2023, According to New IDC Spending Guide” Available: https://www.idc.com/getdoc.jsp?containerId=prUS45481219
2. Angwin, Julia, et al. "Machine bias. ProPublica." See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).
3. Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women." San Fransico, CA: Reuters. Retrieved on October 9 (2018): 2018.
4. Mike Thomas. “Six dangerous risks of Artificial Intelligence” Builtin. January 14,2019
5. Sam Levin and Julia Carrie Wong “Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian” TheGuardian, 19 Mar 2018
6. Schlesinger, Ari, Kenton P. O'Hara, and Alex S. Taylor. "Let's talk about race: Identity, chatbots, and AI." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018.
7. Rossi, Francesca. "Building trust in artificial intelligence." Journal of international affairs 72.1 (2018): 127-134.
8. Goodman, Bryce, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a “right to explanation”." AI magazine 38.3 (2017): 50-57.
9. Naveen Joshi. “How we can build Trustworthy AI” Forbes, Jul 30, 2019
10. Jobin, Anna, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." *Nature Machine Intelligence* 1.9 (2019): 389-399.
11. Smuha, Nathalie A. "The eu approach to ethics guidelines for trustworthy artificial intelligence." *CRi-Computer Law Review International* (2019).
12. Paul R. Daugherty and H. James Wilson. 2018. Human + Machine: Reimagining Work in the Age of AI. Harvard Business Review Press, Boston, MA, USA.
13. European Commission. "White paper on artificial intelligence–a European approach to excellence and trust." (2020).
14. Veeramachaneni, Kalyan, et al. "AI^ 2: training a big data machine to defend." *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC).*

15. Ruan, Y., Zhang, P., Alfantoukh, L., Durresi, A.: Measurement theory-based trust management framework for online social communities. ACM Trans. Internet Technol. 17 (2), 24 (2017). Article 16
16. Uslu, Suleyman, et al. "Control Theoretical Modeling of Trust-Based Decision Making in Food-Energy-Water Management." Conference on Complex, Intelligent, and Software Intensive Systems. Springer, Cham, 2020.
17. Uslu, Suleyman, et al. "Trust-based decision making for food-energy-water actors." International Conference on Advanced Information Networking and Applications. Springer, Cham, 2020
18. Uslu, Suleyman, et al. "Trust-based game-theoretical decision making for food-energy-water management." International Conference on Broadband and Wireless Computing, Communication and Applications. Springer, Cham, 2019.
19. Kaur, Davinder, Suleyman Uslu, and Arjan Durresi. "Trust-based security mechanism for detecting clusters of fake users in social networks." Workshops of the International Conference on Advanced Information Networking and Applications. Springer, Cham, 2019
20. Kaur, Davinder, et al. "Trust-based human-machine collaboration mechanism for predicting crimes." International Conference on Advanced Information Networking and Applications. Springer, Cham, 2020.
21. Xu, Weilin, David Evans, and Yanjun Qi. “Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks.” Proceedings 2018 Network and Distributed System Security Symposium (2018): n. pag. Crossref. Web.
22. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
23. Raji, Inioluwa Deborah, et al. "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020
24. Katell, Michael, et al. "Toward situated interventions for algorithmic equity: lessons from the field." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.
25. Wieringa, Maranke. "What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.
26. Mehri, Vida Ahmadi, Dragos Ilie, and Kurt Tutschku. "Privacy and DRM Requirements for Collaborative Development of AI Applications." Proceedings of the 13th International Conference on Availability, Reliability and Security. 2018.
27. He, Yingzhe, et al. "Towards privacy and security of deep learning systems: a survey." *arXiv preprint arXiv:1911.12562* (2019).
28. Hintze, Mike. "Science and Privacy: Data Protection Laws and Their Impact on Research." Wash. JL Tech. & Arts 14 (2018): 103.
29. Y. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," *2015 IEEE Symposium on Security and Privacy*, San Jose, CA, 2015, pp. 463-480, doi: 10.1109/SP.2015.35.
30. Ragot, Martin, Nicolas Martin, and Salomé Cojean. "AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence?." *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020
31. Annie Brown. “Biased algorithms learn from biased data : 3 kinds of biases found in AI datasets” Forbes. Feb 7, 2020
32. Stock, Pierre, and Moustapha Cisse. "Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases." Proceedings of the European Conference on Computer Vision (ECCV). 2018
33. Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).

34. Agarwal, Aniya, et al. "Automated test generation to detect individual discrimination in AI models." arXiv preprint arXiv:1809.03260 (2018)
35. Srivastava, Biplav, and Francesca Rossi. "Towards composable bias rating of AI services." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018
36. Celis, L. Elisa, et al. "How to be fair and diverse?." *arXiv preprint arXiv:1610.07183* (2016).
37. Sablayrolles, Alexandre, et al. "Radioactive data: tracing through training." *arXiv preprint arXiv:2002.00937* (2020).
38. Lepri, Bruno, et al. "Fair, transparent, and accountable algorithmic decision-making processes." *Philosophy & Technology* 31.4 (2018): 611-627
39. Bellamy, Rachel KE, et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." *IBM Journal of Research and Development* 63.4/5 (2019): 4-1.
40. Mueller, Shane T., et al. "Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI." *arXiv preprint arXiv:1902.01876* (2019).
41. Wang, Danding, et al. "Designing theory-driven user-centric explainable AI." *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019.
42. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016
43. Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.
44. Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020
45. Zhang, Quan-shi, and Song-Chun Zhu. "Visual interpretability for deep learning: a survey." Frontiers of Information Technology & Electronic Engineering 19.1 (2018): 27-39.
46. Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. 2018.
47. Madumal, Prashan, et al. "Explainable reinforcement learning through a causal lens." arXiv preprint arXiv:1905.10958 (2019
48. Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.
49. Mahendran, Aravindh, and Andrea Vedaldi. "Understanding deep image representations by inverting them." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015