# Explainable, Interpretable, Trustworthy, Responsible, Ethical, Fair, Verifiable AI... What's next?

Rosa Meo[0000−0002−0434−4850], Roberto Nai[0000−0003−4031−5376], and Emilio Sulis3[0000−0003−1746−3733]

University of Torino, Italy
{rosa.meo,roberto.nai,emilio.sulis}@unito.it
https://www.cs.unito.it/do/home.pl

**Abstract.** Artificial Intelligence plays an increasingly important role in many knowledge fields: computer science, technology, and other sciences such as health care, one of its most compelling applications. Artificial Intelligence has impacted arts, linguistics, law, sociology, society, and everyday lives. We are demanding many properties from the products of Artificial Intelligence: users of their application fields need trust and ask for fairness, accountability, and privacy. We overview the desired properties and recall the technology that enables Artificial Intelligence to satisfy them.

**Keywords:** Explainable AI · Trustworthy AI · Fairness · Ethics · Accountability

## 1 Introduction

Artificial Intelligence (AI) refers to computational systems whose actions and decisions resemble human intelligence, including functions typically associated with intelligence, such as learning, problem-solving, planning, and acting rationally, as defined by Russell and Norvig [18]. We interpret the term AI broadly to include closely related areas such as machine learning (ML). Systems that heavily use AI, have had a significant impact in domains that include healthcare, transportation, finance, social networking, e-commerce, and education. These "intelligent" systems have almost pervaded all the areas of our modern society. This growing societal impact has brought a set of risks and concerns, including the mistakes that AI systems can make. As a response, researchers are trying to design and deploy a new generation of systems that are trustworthy, i.e, meritable of trust from human beings and more robust to errors in software, resilient to cyber-attacks, and secure, in presence of incomplete scenarios.

The ingredients for a trustworthy Artificial Intelligence (AI) are manifolds. This is related to the deployment of an AI product: sometimes the output of an AI system is used to support the decision-making, and in this case, end-users will need to trust the outcomes of the artificial model. Other times the system is

used to inform the user about the inner structure of the instances coming from the application domain. In these cases, the end-user needs to be convinced that the system has grasped a meaningful organization of the application domain examples.
Systems whose outcomes cannot be well-interpreted are difficult to trust, especially in sectors, such as healthcare or self-driving cars, in which the impact of an erroneous decision has moral and fairness implications [15]. This need for models that are trustworthy, fair, robust with respect to missing data, high-performing in the real-world applications led to the revival of eXplainable Artificial Intelligence (XAI) [13]. This field focuses on the understanding and interpretation of AI systems' behavior. The popularity of the search term "Explainable Artificial Intelligence" in the last five years, as measured by Google Trends, is illustrated in Figure 1. The noticeable spike in recent years reflects also the increased research output of the same period.

XAI is not a monolithic concept: it reflects several related notions. The *explainability* and *interpretability* terms are often usually used interchangeably [14,5]. However, while they are very closely related, some works identify differences among related concepts [16]. We will distinguish them in the following. XAI has numerous applications: model validation, model debugging, and knowledge discovery [11]. The obtained explanations should show whether a machine learning model is grounded upon the possible biases in the training data or show when the learned models ignore important parts of the input data and instead rely on irrelevant ones. They could show that the flaws of the models could be caused by flaws in the training data.

As the demand for more explainable machine learning models with interpretable predictions rises, so does the need for methods that can help to achieve these goals. XAI is centered on the challenge of demystifying the black boxes but also implies *Responsible AI* as it can help to produce transparent models. Responsible AI takes into account societal values and moral and ethical considerations. Responsible AI has three main concepts: *Accountability*, *Responsability*, *Transparency*; these are called the A.R.T. of AI [9]. Finally, XAI is a part of a new generation of AI technologies called the *third wave AI* [21]. One of the objectives of this ambitious "wave" is to precisely generate models than can explain themselves.

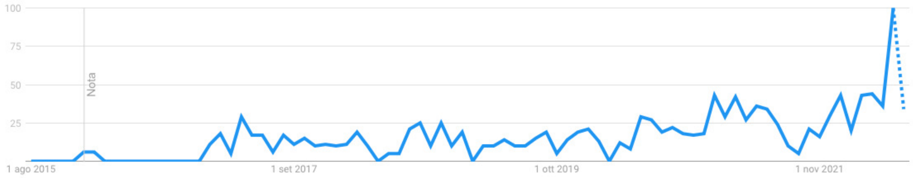

**Fig. 1.** Google Trends popularity index of the term "Explainable Artificial Intelligence" over the last five years (2017–2022).

## 2 Explainability

Explainability is more related to the techniques thought to convince the end-user about the validity of the model outcomes. The most common methods are providing post hoc explanations or recalling from the domain similar instances to the given one in input [11]. These post hoc explanations are local, and specific to single instances and can be model-agnostic or specific to the single method. The model agnostic ones treat the model to be explained as a black-box and assume the predictions of the global model can be approximated as the application of many interpretable white-box models, valid locally, in a small neighborhood of each input. Then, they sample the feature space in the neighborhood of each instance to prepare a training set that is passed to train a white-box model, such as a sparse linear model (Lasso), or if the local behavior is non-linear using if-then rules. Another approach is to determine the importance of each feature on the model by measuring the impact of features' perturbations on the output score. The results may be interpreted as counterfactual explanations, that describe a causal relationship between the input X and the output Y. They have the form: "If input X had not occurred, output Y would not have occurred".

Explanation approaches, designed for a specific type of model, leverage on the characteristics of the model to explain them. For instance, for Deep Neural Networks (DNN) we need to treat their structure as a white box and describe their components. There are three methods: back-propagation methods (top-down) compute the gradient of specific outputs with respect to the input and back-propagate it to derive the contribution of each feature. This method can be efficiently implemented in software libraries (PyTorch or TensorFlow) as a modified gradient function but can give noisy explanatory results. Perturbation methods work bottom-up (with mask perturbations in an optimization framework) and learn a perturbation mask that preserves the contribution of each feature and can be trained by an additional DNN. The intermediate methods either transform the representations at the higher layers of the DNN into a synthetic image together with an encoding of the target object in a mask, or they adopt a prediction's decomposition through the additive contribution of the hidden vectors in the DNN corresponding to each input (e.g, a word in the textual input to a Recurrent Neural Network). Therefore, each component of the decomposition quantifies the contribution of each input to the DNN output.

## 3 Interpretability

One of the most popular definitions of interpretability is the one of Doshi-Velez and Kim, who, in their work [10], define it as "the ability to explain or to present in understandable terms to a human". Interpretability is more focused on the task of exploration of the model properties with the goal of providing transparency to humans. For instance, clarifying the meaning of the components of a black-box model, like a deep neural network or a Support Vector Machine with the goal of understanding the model. The most common technique is to put aside an obscure model a “white-box” model, trained on the same instances. The latter model incorporates interpretability directly into its structure. This is the case of logical models (decision tree or rule-based model), linear models (that accompany features with coefficients whose magnitude informs their impact on the model outcome), attention model (for natural language, referred to as the words in the context).

One of the more interesting goals of learning an interpretation of a black box model is to understand the representations of the input (images) captured by the Deep Neural Network (DNN) model like a Convolutional one (CNN). Here we refer to the CNN internal network nodes because we know they encode artifacts learned from the input images. One of the most effective methods is finding the inputs that best activate neurons at a specific layer [11]. The optimization should be regularized using natural image priors produced by a generative model (GAN). Instead of directly optimizing the image, these methods optimize the latent space codes of the GAN to find an image that activates a given neuron. The visualization results provide several interesting observations. The neurons from the first layer to the last layer learn representations at several levels of abstraction, from general to task-specific. The second interesting learned issue is that a neuron is multifaceted, i.e., could respond to different images, semantically related to the same concept (i.e. faces). CNN learns distributed code for objects and learns objects by the representation of their parts that can be shared across different categories [11].

Based on the above, interpretability is mostly connected with the intuition behind the outputs of a model [1] and the idea that the more interpretable a machine learning system is, the easier it is to identify cause-and-effect relationships within the system inputs and outputs. Doshi-Velez and Kim [10] proposed the following classification of evaluation methods for interpretability: application-grounded, human-grounded, and functionally-grounded; Figure 2 shows the taxonomy proposed. Application-grounded evaluation concerns itself with how the results of the interpretation process affect the human, domain expert, and end-user in terms of a specific and well-defined task or application. Human-grounded evaluation is similar to application-grounded evaluation; however, there are two main differences: first, the tester, in this case, does not have to be a domain expert, but can be any human end-user, and secondly, the end goal is not to evaluate a produced interpretation with respect to its fitness for a specific application, but rather to test the quality of the produced interpretation in a more general setting and measure how well the general notions are captured. Functionally grounded evaluation does not require any experiments that involve humans but instead uses formal, well-defined mathematical definitions of interpretability to evaluate the quality of an interpretability method. This type of evaluation usually follows the other two types of evaluation: once a class of models has already passed some interpretability criteria via human-grounded or application-grounded experiments, then mathematical definitions can be used to further rank the quality of the interpretability models.

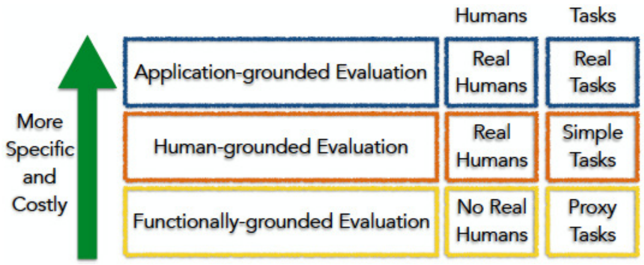

**Fig. 2.** Taxonomy of evaluation approaches for interpretability [10].

## 4 The problem of bias

AI explanations might reveal that decisions are influenced by factors that do not align with explicit organizational policies. Amazon canceled a plan to use AI to identify the best job candidates for technology positions upon discovering the models were biased against women because the training data consisted predominantly of males, reflecting historic hiring practices [7]. The example above explains that biases in AI mean biases in predictions. The ethical consequences of algorithmic decision-making by AI systems are a great concern. The emergence of biases in AI-led decision-making has seriously affected the adoption of AI. In order to build an unbiased system, a strong sense of justice needs to be in place to help decision makers act fairly without having any prejudice and favoritism [4].

The survey presented in [20] discusses the different sources of bias. Figure 3 shows a taxonomy of the sources of bias according to the authors.

As it is well-known, the knowledge discovery process in AI stems from a pipeline composed of different steps: data source cleaning, integration, feature selection or feature construction, model training, selection, validation, and finally, outcome presentation. All these steps might be the source of some bias. Some might be due to the users/analysts insufficient knowledge/preparation that comes out under the multiple forms of employing a sampling bias, showing a capture bias, a device bias, a measurement bias, or a negative set bias (insufficient examples for the negative class), or a confirmation bias (that leads to ignoring some relevant issues in the domain). All these examples of bias could lead to unsuitable choices in the data preparation.

Other biases could come from the presence of an ill-posed domain problem: the framing effect bias is sometimes due to the need to formulate the problem so that the experimental measured results could reflect some business objective. Another source of mistakes in the AI model is the confounding bias, that exists when an omitted feature is not included in the training data: this makes it impossible to measure the correlation between causes and effects. Another example is the inclusion of a proxy feature that is the source of indirect discrimination (e.g., zip code could be correlated to the ethnic condition), a sensitive feature that should be omitted to avoid discrimination based on ethnic conditions).

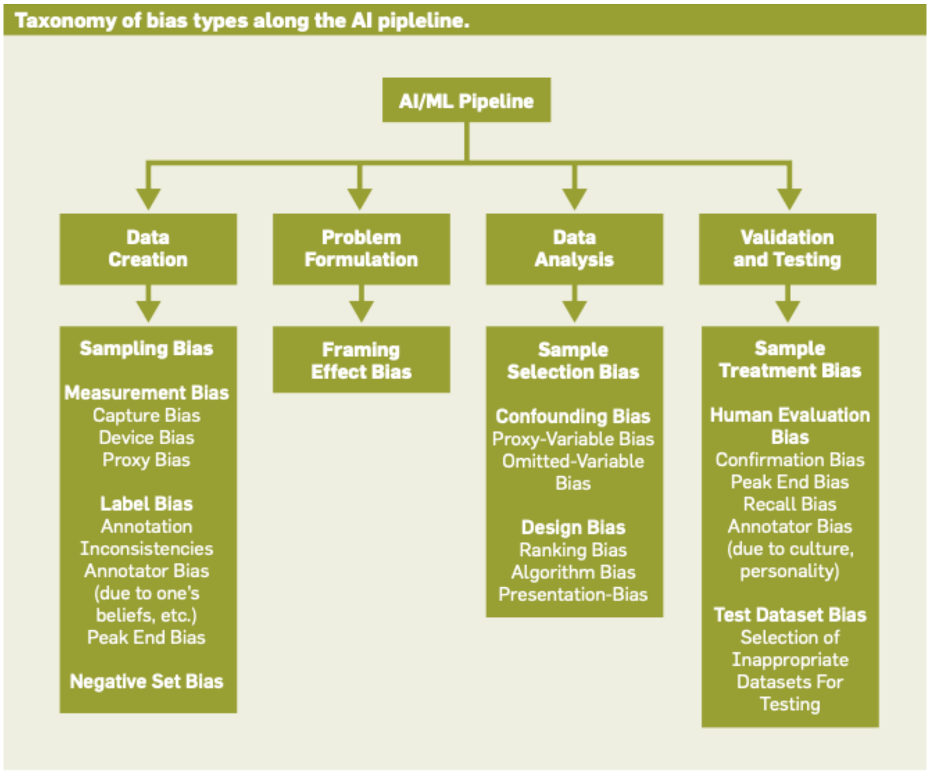

**Fig. 3.** Taxonomy of biases [20].

Other biases are algorithms biases: influence on the model outcome by how they explore the hypothesis and evaluate constraints, or how they present their results in a ranking to the users waiting for feedback. The users could not be impartial or could not be ready in their evaluation due to a recall bias. Finally, the deployment of the AI model might in turn influence the studied scenario and alter it.

Some effects of the biases could have serious effects on people's discrimination and on give serious doubts about the fair application of some AI models. These are discussed in Section 5.

## 5 Fairness

Because machine learning systems are increasingly adopted in real-life applications [1], any inequities or discrimination that are promoted by those systems have the potential to directly affect human lives [1]. Machine Learning Fairness is a sub-domain of machine learning interpretability that focuses solely on the social and ethical impact of machine learning algorithms by evaluating them in terms of impartiality and discrimination [6]. Traditionally, the fairness of a machine learning system has been evaluated by checking the model predictions and errors across certain demographic segments, for example, groups of a specific ethnicity or gender. In terms of dealing with a lack of fairness, a number of techniques have been developed both to remove bias from training data and from model predictions and to train models that learn to make fair predictions in the first place.

One of the proposed methods to control the bias in data is by maintaining diversity in examples' collection. It could allow the models to achieve statistical parity among the represented categories (or groups, among which some minorities exist and should be protected from discrimination). There are several measures for accounting fairness of treatment of an AI model to groups of people.

*Statistical parity* accounts for the parity of the rates of the favorable outcomes produced by the AI model when applied to exemplars coming from the unprivileged group and the privileged group.

*Equal Opportunity* accounts for the true positive rates between the unprivileged group and the privileged group.

*Average Odds* takes into consideration the odds between the false positive rates and true positive rates in the two groups.

*Disparate impact* compares the rates of the favorable outcome in the two groups by considering their ratio.

Unfortunately, not all of them could be applicable at the same time, because their satisfaction could depend on the distribution of the categories in the population and on the possible existence of some correlation between the sensitive attribute and the target.

Some software tools and standards of behavior in the data analysis and AI model development already exist [2] and are promoted by the big software vendors and by the European Community [17] and should be adopted by the software developers and the business development teams to verify the existence of some disparities in treatment by the AI model.

## 6 Verification

According to [19] the next generation of AI systems shall be verified with techniques similar to the formal methods used in computer science to test integrated circuits, debug software architectures, and Cyber-Physical Systems (CPS). It is composed of specifications, systems design that adheres to these specifications, verification by algorithmic search, specifications testing, simulation, and model checking. In Figure 4 we show the tasks involved in the verification of a complex AI-based CPS, where a modular approach is essential for scalability but not yet easily reached. Finally, correct-by-construction design methods hold promise for achieving verified AI, but they are in their infancy and are still premature. Figure 4 summarizes the five challenge areas for verified AI. For each area, the current promising approaches are organized into three principles, depicted as nodes. Edges between nodes show the dependency among the principles, with the color denoting a common thread. The authors of [19] developed open-source tools, VerifAI [8] and Scenic [12] which implement the techniques based on the principles described.

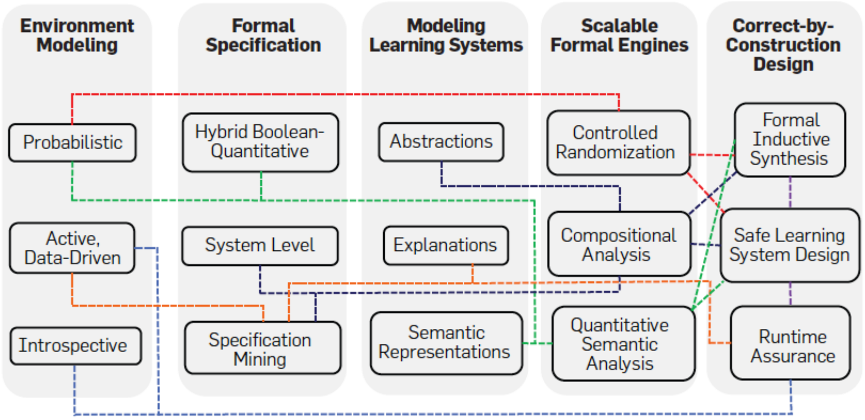

**Fig. 4.** Summary of the five challenge areas for verified AI, the corresponding principles proposed to address them, and their connections and dependencies [19].

For example, runtime assurance (fifth challenge) relies on introspective and data-driven environment modeling to extract monitorable assumptions and environment models (from the first challenge). Similarly, to perform system-level analysis (second challenge), we require compositional reasoning and abstraction.

Some AI components may require specifications to be mined, while others are generated correct-by-construction via formal inductive synthesis (from the fifth challenge).

## 7 Accountability

The terms accountability, responsibility, and liability are closely related but carry different meanings. According to the OECD group of experts on AI [17], "accountability" implies ethical, moral, or in terms of management practices, codes of conduct. It guides the individuals' or organizations' actions and allows them to explain the reasons for which the actions were taken. From the viewpoint of moral principles, accountable systems are related to the concepts that guide "moral machines" [3] AI systems that are proposed and designed in a large-scale crowd-sourced experiment conducted by MIT researchers in 2018. The aim of the experiment is to collect and study the ethical and moral principles that should guide autonomous driving cars to take their decisions, in front of moral dilemmas. "Liability" generally refers to adverse legal implications arising from a person's or an organization's actions. "Responsibility" can also have ethical or moral expectations and refers to a causal link between an actor and an outcome.

Given these meanings, the term "accountability" best captures the essence of the moral principles behind the decisions of autonomous systems. In this context, "accountability" refers to the expectation that organizations or individuals will ensure the proper functioning of the AI systems that they design, develop, operate or deploy, throughout their lifecycle. For proving this, through their actions and the decision-making process they should provide documentation on the key decisions throughout the AI system lifecycle or they should conduct or allow auditing. From these viewpoints, accountability is related to systems that can be verified, as described in Section 6.

## 8 Conclusions

We provided a summary of the overview of the rapidly evolving field of explainable and interpretable AI. While many application areas of the AI systems need trust and fairness and demand responsible principles to guide the automated decisions, other applications like Cyber-Physical systems and autonomous driving need also the principles of the formal methods for obtaining verifiable systems, to guarantee software security also against cyber-attacks.

## References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
2. AI, I.R.T.: AI Fairness 360. IBM (2022), `https://aif360.mybluemix.net`

3. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I.: The Moral Machine experiment. **563**(7729), 59–64 (Oct 2018). `https://doi.org/10.1038/s41586-018-0637-6`
4. Bennetot, A., Donadello, I., Qadi, A.E., Dragoni, M., Frossard, T., Wagner, B., Saranti, A., Tulli, S., Trocan, M., Chatila, R., et al.: A practical tutorial on explainable AI techniques. arXiv preprint arXiv:2111.14260 (2021)
5. Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U.: Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv preprint arXiv:1704.07911 (2017)
6. Chouldechova, A., Roth, A.: The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810 (2018)
7. Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of Data and Analytics, pp. 296–299. Auerbach Publications (2018)
8. Derossi, T., Fremont, D.J., Ghosh, S., Kim, E., Ravanbakhsh, H., Vazquez-Chanlatte, M., Seshia, S.A.: Verifai: A toolkit for the formal design and analysis of artificial intelligence-based systems. In: Dillig, I., Tasiran, S. (eds.) Computer Aided Verification. pp. 432–442. Springer International Publishing, Cham (2019)
9. Dignum, V.: Responsible artificial intelligence: designing AI for human values (2017)
10. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
11. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Commun. ACM **63**(1), 68–77 (dec 2019). `https://doi.org/10.1145/3359786`, `https://doi.org/10.1145/3359786`
12. Fremont, D.J., Dreossi, T., Ghosh, S., Yue, X., Sangiovanni-Vincentelli, A.L., Seshia, S.A.: Scenic: A language for scenario specification and scene generation. In: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation. p. 63–78. PLDI 2019, Association for Computing Machinery, New York, NY, USA (2019). `https://doi.org/10.1145/3314221.3314633`, `https://doi.org/10.1145/3314221.3314633`
13. Gunning, D., Aha, D.: Darpa's explainable artificial intelligence (xai) program. AI magazine **40**(2), 44–58 (2019)
14. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning. pp. 1885–1894. PMLR (2017)
15. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)
16. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
17. OECD Policy Observatory: From principles to practice: tools for implementing trustworthy AI. OECD (2022), `https://oecd.ai/en/tools`
18. Russell, S.J., Norvig, P.: Artificial Intelligence: a modern approach. Pearson, 3 edn. (2009)
19. Seshia, S.A., Sadigh, D., Sastry, S.S.: Toward verified artificial intelligence. Commun. ACM **65**(7), 46–55 (jun 2022). `https://doi.org/10.1145/3503914`, `https://doi.org/10.1145/3503914`
20. Srinivasan, R., Chander, A.: Biases in AI systems. Commun. ACM **64**(8), 44–49 (jul 2021). `https://doi.org/10.1145/3464903`, `https://doi.org/10.1145/3464903`
21. Xu, W.: Toward human-centered AI: a perspective from human-computer interaction. interactions **26**(4), 42–46 (2019)