# Mathematical decisions and non-causal elements of explainable AI

**Atoosa Kasirzadeh**
Australian National University & University of Toronto
atoosa.kasirzadeh@anu.edu.au

## Abstract

Recent conceptual discussion on the nature of the explainability of Artificial Intelligence (AI) has largely been limited to data-driven investigations. This paper identifies some shortcomings of this approach to help strengthen the debate on this subject. Building on recent philosophical work on the nature of explanations, I demonstrate the significance of two non-data driven, non-causal explanatory elements: (1) mathematical structures that are the grounds for capturing the decision-making situation; (2) statistical and optimality facts in terms of which the algorithm is designed and implemented. I argue that these elements feature directly in important aspects of AI explainability. I then propose a hierarchical framework that acknowledges the existence of various types of explanation, each of which reveals an aspect of explanation, and answers to a different kind of why-question. The usefulness of this framework will be illustrated by bringing it to bear on some salient normative concerns about the use of AI decision-making systems in society.

## 1 Introduction

A truly vast array of deployments have been found for machine learning algorithms in decision making. Both governments and private actors are using these algorithms in high-stake decision contexts, including health care, criminal justice, and the labour market ([1], [3], [6], [25]). Making these algorithms able to explain their recommendations is one way to increase societal acceptance of algorithmic decision outcomes, to establish trust in the results of the decisions, to validate the decisions, and to have fruitful conversation among different stakeholders on the justification of using these algorithms for decision making. But what is explainability and interpretability? Computer scientists have suggested that AI explainability and interpretability are not monolithic concepts, and are used in different ways ([8], [14]). Therefore, bringing in conceptual clarity to the debate is of primary significance. One promising way for such clarification is to seek inspirations from social sciences and philosophy on what an explanation is, and how an explanation is related to interpretation and understanding. So far, this inspiration-seeking approach has merely focused on elucidating what the data-driven and causal aspects of an explanation are. For instance, in the most extensive survey in this area, Miller [15, p.20] entirely dismisses the non-causal aspects of explanations:

> But what constitutes an explanation? [...] accounts of explanation both philosophical and psychology [sic] stress the importance of causality in explanation — that is, an explanation refers to causes [...]. There are, however, definitions of non-causal explanation [...]. These definitions [are] out of scope in this paper, and they present a different set of challenges to explainable AI.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

This paper broadens the discourse on the explainability and interpretability of AI by incorporating some insights about non-data driven and non-causal aspects of explanation from the recent philosophical literature.[1] Investigations on the nature of non-causal explanations, as developed by philosophers, have grown extensively in the last two decades ([2], [4], [7], [13], [19], [20]). Drawing on this literature, I provide a more enriched and philosophically-informed conceptual framework for the explainability and interpretability of AI decision making which will facilitate descriptive and normative conversation among several stakeholders affected by AI decision-making. I take an explanation to be a response to a why-question [5], and to be empirically or mathematically verifiable. I acknowledge the significance of causal and data-driven information for some types of explanations, but I argue that a thorough investigation on AI decision making and explainability requires identification of two implicit and crucial non-data driven, non-causal elements. Accordingly, I propose a hierarchy for AI explanations in which a variety of causal and non-causal information find their place in AI explanations. The explanatory hierarchy builds on Pearl's hierarchy of causation [17] by adding a variety of mathematical explanations to his hierarchy. My proposed hierarchy is composed of five levels of explanation, and is sorted in ascending order of locality from the most structurally global to the most data-driven, local explanation.

Often in the literature, the two concepts of machine-learning explainability and interpretability go hand in hand, and are used interchangeably. Here is an example: "In the context of ML systems, I define interpretability as the ability to explain or to present in understandable terms to a human" ([8]). To capture the many differences and similarities pertaining to explainability and interpretability, I distinguish between the two issues. I discuss that background assumptions (such as social, political, and institutional norms) influence human explanatory judgements. This suggests that explanations of a phenomenon are diverse, and are interpreted in relation to different precedent assumptions. Accordingly, I define two separate, yet closely related, schemas: an explanatory and an interpretative. The two schemas forms a conceptual framework for a focused evaluation of the implicit and explicit context of the applicability of the decision-making algorithm.

## 2 Mathematical explanations: structure, statistics, and optimization

Let us consider a deep supervised learning algorithm that sifts through several job applications to recommend a hire for company X. Nora, a competent candidate, applies for the job. Her application gets rejected by the algorithmic decision. Nora wants to know why she is rejected. She wants the explanation to have factual foundations and to be empirically valid. So far, the work on the explainability of AI has been mainly focused on building methods that generate data-driven explanations for an algorithmic decision outcome ([9], [12], [21], [23], [26]). In relation to Nora's case, these attempts can be mainly summarized to answer either of the following explanatory questions: (i) What is a causal explanation for Nora's rejection? (ii) What is an associational explanation for Nora's rejection? Figure 1 illustrates how (i) and (ii) are situated in respect to Nora's explanatory question.

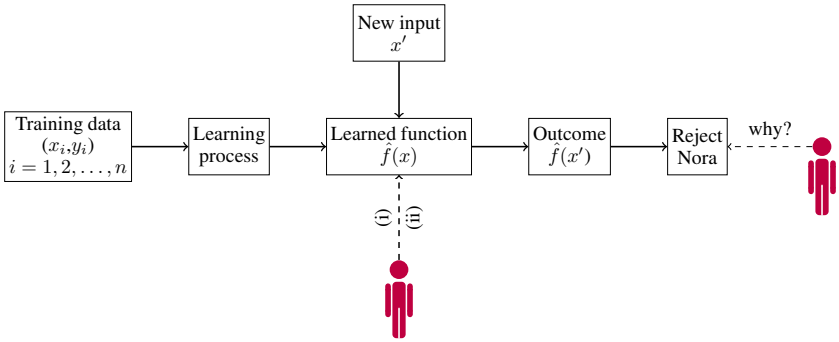


Figure 1: Generating explanations for algorithmic decision outcomes via AI

[1] For simplicity, here I only focus on explainability in the context of deep supervised learning. This assumption is also justified on the ground that several critical decision-making problems have been considered to be instances of classification or regression, two tasks performed by supervised learning algorithms.

I argue that this explanatory landscape is insufficient and inadequate. I begin by defending the importance of structural explanations.

The producer of an explanation for a given decision stands in a relation with the decision-making situation. This situation is represented and formulated in a particular way, and different concepts are used to specify the elements of the decision-making situation. The generated explanation is the outcome of investigation and reflection on the reasons-why for an outcome, relevant to the representation of the decision-making situation. The deep learning architecture for representation of the decision-making problem enforces an implicit assumption: there is an isomorphism between the input layer of the mathematical representation and the source of the input layer data coming from the real world that make up the decision-making situation. To represent mathematically is to have clear and distinct ideas about the stuff out there and also the relation between them: what features are important and relevant, how the decision-making configurations should be translated and reduced into numerical values on the neural network structure, such as the link between nodes and what the activation function should be. Mathematical structure and the form of representation as required by the algorithm gives a response to the explanatory question (iii): why does the algorithm merely observe Nora's attributes mathematically relevant to decision making? The answer is that to use these algorithms, the precondition is to consider some relevant features of Nora, measure them, and make them as the only input features to the algorithms. If the mathematical representation of the decision-making situation would be very different, we might expect a different decision outcome. Accordingly, a part of explanations for the decision outcomes must be framed in terms of the mathematical structures that are constitutive of the decision-making characterization, and hence of decision results. Having established the explanatory relevance of mathematical structures to algorithmic decision-making problems, I now turn to the explanatory role of statistics and optimization in warranting AI decision making.

Deep supervised learning algorithms use some methods of statistical analysis in order to predict the probability of the occurrence of an outcome. These probabilistic inferences are permitted due to statistical facts such as the law of large numbers and the central limit theorem. Therefore, the decision outcomes of several machine learning algorithms are partially dependent on statistical facts that warrant such learning from data. Such statistical facts partially govern and influence the design of the decision procedure, and therefore are an explanatory element for why a decision outcome is achieved. Moreover, there are several ways to extract causal information from statistical facts ([18]). The discovery process finds causal relations by analyzing statistical properties of purely observational data. These methods use the notion of conditional independence relationships in the data in order to find causal relations. Causal relations are discovered based on several statistical properties. In addition, most supervised learning algorithms are based on optimizing a particular objective function such as error minimization. Indeed, mathematical optimization is used in a variety of ways during the training of an artificial neural network. To train an artificial neural network is to resolve an optimization problem, such as Stochastic Gradient Descent. The optimization function as a part of the algorithmic design influences the decision outcome. Therefore, statistical facts and optimization methods and assumptions also warrant causal relations. Our discussion suggests the significance of the following two explanatory questions as instances of structural and optimality explanations: (iv) Why does statistical thinking and a mathematical optimization function decide Nora's hire? (v) Why does a deep structure and this training data set influence Nora's condition? The plurality of the relevant explanatory why-questions is illustrated in Figure 2.

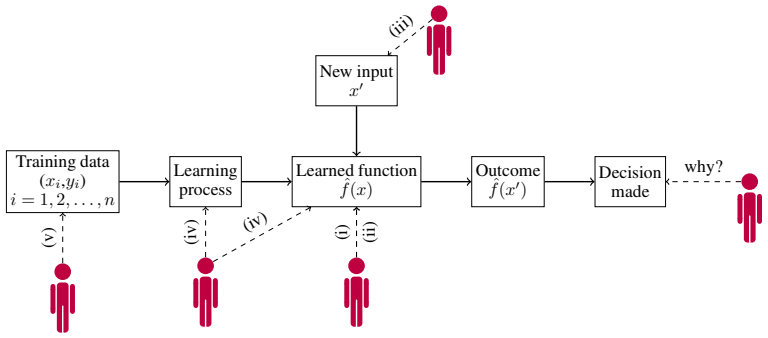


Figure 2: The plurality of why-questions about an algorithmic decision

## 3 The many faces of explanations

The question "what is an explanation?" has taken central stage in contemporary philosophy of science. Roughly, these discussions aim to answer one of the following three questions. (1) Are explanations reducible to causal explanations, or are there genuine cases of non-causal explanations? (2) Whether, and if so how, can I give necessary and sufficient conditions for explanations? (3) How do non-causal explanations work, if they exist at all? [11] and [10] offer an account of explanation that emphasizes its argumentative nature. The argumentative nature of explanations generated by AI systems has been emphasized in the literature by [15] and [16], among others. Since Hempel, some philosophers such as [22] and [24] have argued that a missing element of a account of explanation is an emphasis on providing causal information hold that explanations are providing information about causal relations that are constitutive of the world. More recently, the discussion has tilted towards unveiling the non-causal elements engaged in the production of an explanation such as the explanatory roles of mathematics. A simple example reveals the insights behind the cluster of these discussions. To explain why a mother cannot divide 23 strawberries between her three children, one can appeal to mathematical facts about numbers that 23 cannot be divided 3 evenly. Of particular interest are cases of optimality explanations in which reference to an optimality notion, such as equilibrium, is responsible for an explanation of some empirical phenomena such as natural selection ([19]). This extends to explanation of socially significant phenomena as well. I have now sufficient ingredients to propose the hierarchical explanatory framework.

On the top level, there is structural explanation which can unveil why a particular decision output is generated in virtue of a specific structural mapping of the decision-making situation on to a mathematical representation. On a lower level, we have statistical and optimality explanations that emphasize the statistical and optimality laws or facts engaged in forming the decision-making procedure. These two kinds of explanations acknowledge the importance of non-causal elements warranting the AI decision making. They open space for asking questions concerning justifications for reductive representation of significant decision-making attributes. It is correct that these explanations might not be understandable by a lay-person. However, this reason should not make us dismissive of the significance of these explanations. The proposed explanation hierarchy is powerful enough to be used to answer several normative questions about the context of the applicability of the algorithm. For this reason, we do not want to set minimal requirements on explanations to be merely stories given to us by AI algorithms. We want the explanations to be rooted in empirical and mathematical facts, in order to critically evaluate and discuss them. On the descending order of locality, three other levels of explanation are added to the hierarchy (inspired by [17]]. These levels correspond to data-driven information, and convey causal information, rather than an emphasis on the mathematical and statistical constitution of the decision-making procedure. They are associational explanations (model-agnostic or model-dependent), causal explanations (model-agnostic or model-dependent), and example-based explanations. Each of these explanations might be textual or visual. Figure 3 illustrates this hierarchy.

My proposed explanatory hierarchy enforces a platform for conveying and communicating the explicit and implicit mathematical assumptions and social and moral norms that designers of the algorithms grapple with, and the objective functions in terms of which they design the algorithm. The functional value of the explanations, partly, depends on the audience who consume them: an explanation must result in an appropriate level of understanding or some grade of cognitive achievement for the receivers of explanations. In other words, explanations are required to be interpreted and judged against different vantage points, about whether they are good or bad, satisfactory or unsatisfactory, effective or ineffective, acceptable or unacceptable.

To discuss the understandability (or interpretability) of AI-based decisions, I use a contextual conception of understanding that I borrow from the philosophy of science literature. Philosophers have discussed explanation and understanding in the context of scientific inquiry. I carry over their views to the domain of AI explainability and interpretability. In particular, I benefit from a contextual theory of understanding as proposed by De Regt and Dieks (2005). This view suggests that Scientist S (in context C) understands phenomenon P based on theory T. In non-scientific contexts, an audience of an explanation replaces a scientist, and some normative background assumptions replace the scientific theory. I get the following theory of understanding: Audience X in context C understands why decision D is made based on the correctness of an explanation as well as their normative background assumptions N. Therefore, interpretability of AI explanations depends on the appropriate

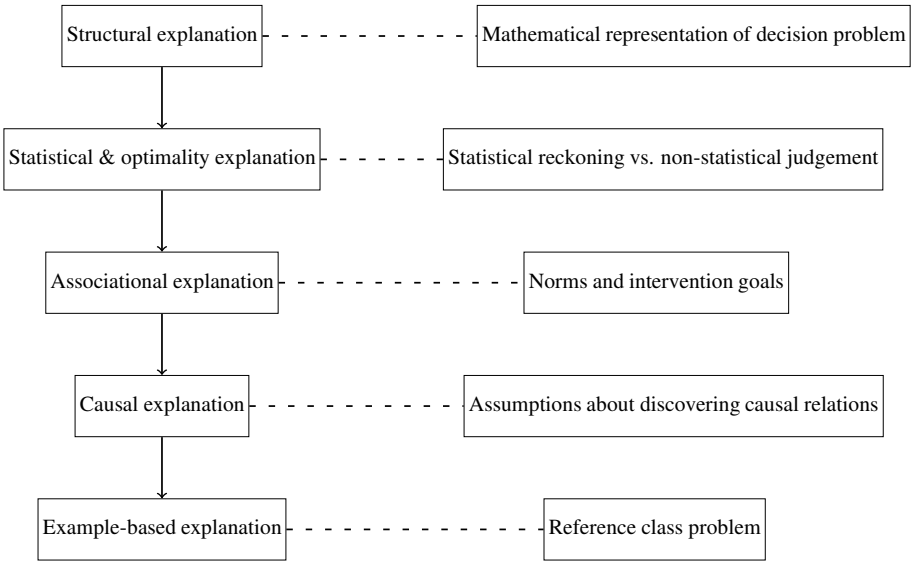


Figure 3: A conceptual framework for AI explainability/interpretability

skills of audience X and qualities of N. I do not have space left to explore and discuss the usefulness of this framework on answering socially significant normative questions. However, I hope that the hierarchical framework as illustrated in Figure 3 gives the intuition that the assumptions and norms listed on the right-hand side of the framework correspond to moral, social, or political assumptions and norms that case light on several aspects of the applicability of an algorithm. In a longer version of this paper, I discuss this point in more details.

## References

[1] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. propublica 2016.

[2] Batterman, R. W. and Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, 81(3):349–376.

[3] Bodó, B., Helberger, N., Irion, K., Zuiderveen Borgesius, F., Moller, J., van de Velde, B., Bol, N., van Es, B., and de Vreese, C. (2017). Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents. *Yale JL & Tech.*, 19:133.

[4] Bokulich, A. (2011). How scientific models can explain. *Synthese*, 180(1):33–45.

[5] Bromberger, S. (1966). Why-questions. In Colodny, R. G., editor, *Mind and cosmos: Essays in contemporary science and philosophy*, pages 86–111.

[6] Cabitza, F., Rasoini, R., and Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518.

[7] Chirimuuta, M. (2017). Explanation in computational neuroscience: Causal and non-causal. *The British Journal for the Philosophy of Science*, 69(3):849–880.

[8] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[9] Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2.

[10] Hempel, C. G. (1965). Aspects of scientific explanation; and other essays in the philosophy of science.

[11] Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175.

[12] Lakkaraju, H. and Rudin, C. (2017). Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics*, pages 166–175.

[13] Lange, M. (2016). *Because without cause: Non-causal explanations in science and mathematics*. Oxford University Press.

[14] Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

[15] Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

[16] Mittelstadt, B., Russell, C., and Wachter, S. (2019). Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288. ACM.

[17] Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

[18] Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

[19] Potochnik, A. (2007). Optimality modeling and explanatory generality. *Philosophy of Science*, 74(5):680–691.

[20] Reutlinger, A. and Saatsi, J. (2018). *Explanation beyond causation: philosophical perspectives on non-causal explanations*. Oxford University Press.

[21] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

[22] Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

[23] Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

[24] Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press.

[25] Veale, M. and Edwards, L. (2018). Clarity, surprises, and further questions in the article 29 working party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, 34(2):398–404.

[26] Zhao, Q. and Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, (just-accepted):1–19.