Defining Explanation in an AI Context

Tejaswani Verma
Mercedes-Benz Innovation Lab, Berlin, Germany
Saarbrücken Graduate School of Computer Science, Saarland University
tejaswani.verma@daimler.com

Christoph Lingenfelder
Mercedes-Benz Innovation Lab, Berlin, Germany
christoph.lingenfelder@daimler.com

Dietrich Klakow
Saarland University, Saarbrücken, Germany
dietrich.klakow@lsv.uni-saarland.de

Abstract

With the increase in the use of AI systems, a need for explanation systems arises. Building an explanation system requires a definition of *explanation*. However, the natural language term *explanation* is difficult to define formally as it includes multiple perspectives from different domains such as psychology, philosophy, and cognitive sciences. We study multiple perspectives and aspects of explainability of recommendations or predictions made by AI systems, and provide a generic definition of *explanation*. The proposed definition is ambitious and challenging to apply. With the intention to bridge the gap between theory and application, we also propose a possible architecture of an automated explanation system based on our definition of *explanation*.

1 Introduction

Definitions of explanation have been proposed by many researchers from the perspective of various fields of science, however the scope of our research is focused on computer science and cognitive science perspectives, as we are looking for automatically generated explanations for people affected by recommendations or predictions made by AI systems. Explanation is often described as a phenomenon which enables transfer of specific information. Philosophically, acceptance of a statement as an explanation is subject to acceptance by the recipient (Gilpin et al., 2018). However, according to computer scientists explanations must be *complete*, i.e. encapsulate all factors of complex internal workings, and *accurate*, i.e. hold high fidelity to the AI model in question (Gilpin et al., 2018).

We can find many definitions in different domains, often conflicting. Some choose to focus only on selected aspects of explanation. For generating explanations automatically, a strong understanding of how people define, generate, select, evaluate, and present explanations is essential. In order to design and implement intelligent agents that are truly capable of providing appropriate explanations to people, analyzing how humans explain decisions and behavior to each other is a good starting point.

In the past, researchers have defined models and conducted social experiments, many of which uncovered novel and essential aspects of human explanations (Harman, 1965; Hanson, 1958; Hesslow, 1988). Specifically social experiments help us understand the impact of unquantifiable attributes of human behavior which include belief, desire, social norms, intention to understand, emotions, personality traits, etc. These attributes are aptly defined by folk psychology, or commonsense psychology (Heider and Simmel, 1944; Malle, 2006). Folk psychology does not describe our thought process or behavior but our expected thought process or behavior, which makes it an important aspect of explanation.

As mentioned earlier, most of the past research is focused on selective aspects of explanation. This is because building a generic explanation system is a tedious task. With the intention to take a step towards realization of this task, we assume the AI system to be a *black-box* i.e. all internal structure and/or processes are hidden and only input-output pairs are perceivable. No restriction is imposed on inputs and outputs.

Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 314–322

In this paper we will review cognitive and social models, experiments and definitions which may assist us in writing a more generic definition of explanation. In the background section we will discuss the need for explanation in detail, which will enable us to define necessary conditions and possible quality measures. We will also highlight important models, theories and conflicting definitions. In later sections, we argue which aspects of explanation are most important and propose a generic definition of explanation. Lastly, we introduce a possible architecture of a generic automated explanation system based on our proposed definition.

2 Background

In this section, we will discuss why explanation is required, the theoretical advances made towards building explanation models and insights about conflicts in the definitions proposed so far.

2.1 Need for Explanation

Many AI systems are complex and hard to understand. Even though AI systems possess potential to address plethora of problems, their applications may be limited due to our inability to comprehend their results and reasoning process. A definition of explanation would be beneficial for multiple groups, the most prominent of which is receivers of a prediction or recommendation made by any AI system. Other groups include researchers, AI developers, users of the system, etc. We refer to all these groups together as *recipients* and a *recipient* can be any individual from this group.
We reviewed the significance of explanation covered by several authors. Here, we reiterate the reasons majorly presented in Doshi-Velez and Kim (2017), Samek et al. (2017), and Lipton (2018). The reasons presented below depict the importance of explanation:

- Enhancement of scientific understanding: Most AI systems are quite powerful, they can process data faster and can find patterns which may go unperceived by humans. Understanding the reasons which contribute to the prediction of a *black-box* offers a great opportunity for learning and extending our scientific understanding. (Doshi-Velez and Kim, 2017; Samek et al., 2017)

- Verification of the system: Explanations can enable us to examine if the system is working as intended and abides by the user's and the recipient's objectives. The workings of the overall system and its objectives are important as the user of an AI system and the recipient of the prediction might have different objectives. (Doshi-Velez and Kim, 2017; Samek et al., 2017)

- Understanding multi-objective trade-offs: In case of multiple objectives an explanation provides information to the recipient regarding which features outweigh others, their magnitude and their collaborative impact on the output. (Doshi-Velez and Kim, 2017)

- Improvement of system:An explanation can also help developers detect fallacies within the system, which allows further improvement of the system. (Samek et al., 2017)

- Inculcating Trust in user: AI system predictions can be confusing at times, explanations will give users an insight of why a certain prediction was made. This makes it easier for users to accept the prediction. It also provides a rough estimate to the user regarding how often the system is right and also what makes it right. (Lipton, 2018)

- Inferring Causal relations in data: Inferring causal relations in data is not trivial, researchers wish to discover such relations which can help them in generating hypotheses about the natural world. Explanations can play a crucial part in deepening our knowledge of the universe. (Doshi-Velez and Kim, 2017; Lipton, 2018)

- Enabling Transferability: Humans possess great abilities such as generalizing and transferring learned skills. Artificial neural networks were inspired by the ability of the human brain to execute multiple tasks given the correct data and processing power, however their capabilities are still limited due to their inability to learn and transfer skills like humans. Explanation might bring us one step closer to understand, how to enable the ability to transfer learned skills. (Lipton, 2018)

- Fair and Ethical Decision-Making: Automated decision making is embedded in our

daily lives from social media platforms, to stock market, to process of approving loans and much more. It has become essential to ensure fair and ethical decision making. This requires transparency in the decision making process which can be enabled by explanations even though the system may not be originally transparent. (Doshi-Velez and Kim, 2017)

- Ensuring safety to use AI models: Although prediction and recommendation systems might not seem dangerous, in certain sensitive scenarios the results can be catastrophic. There is a risk factor involved with every automated decision making system however with some AI models, which might be used for medical diagnosis or self-driving cars, the risk factor is much higher. In order to ensure safety while using these systems we need to understand the system's actions or predictions, measure the risk factors and take appropriate steps. (Doshi-Velez and Kim, 2017)

- Compliance to Legislation: In 2016, European Union GDPR enabled the "Right to Explanation" act[1], making it a legal necessity for all automated systems to provide explanations.

2.2 Explanation vs Interpretation

Explanation definitions can be controversial. The difference between *explainability* and *interpretability* is frequently debated in the computer science community. In the process of distinguishing between explainability and interpretability, the essence of human understandability is often lost. Technically, an *explanation* is considered to be complete and often complex, whereas an *interpretation* defines internal workings of a system in an abstract human understandable way (Gilpin et al., 2018). Intuitively, the term *explanation* used by humans is much closer to the technical term *interpretation*. This is a subject of debate in the research community, however considering the objective explanations, many researchers use these terms interchangeably. From this point onwards we will use the term explanation for the technical term interpretation.

[1]"Right to Explanation" in EU Legislation (Last accessed on: 2020-10-06): https://www.privacy-regulation.eu/en/r71.html

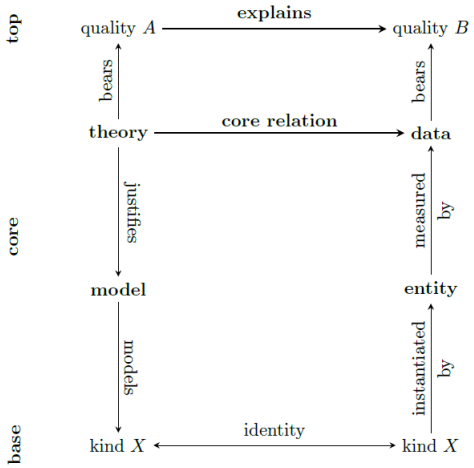

Figure 1: A general structure of a theory-data explanation proposed by Overton.(Overton, 2012)

2.3 Models for generating explanation

Many tried to model explanations and explanation systems for different types of questions (why, why not, how, etc.), Aristotle's four causes model, also known as the Modes of explanation model (Hankinson, 2001), offers an analytical solution to the why questions by classifying them into four different elements:

- Material- The substance or material of which something is made. For example, rubber is a material cause for a car tire.

- Formal- The form or properties of something that make it what it is. For example, being round is a formal cause of a car tire. These are sometimes referred to as categorical explanations.

- Efficient- The main mechanism which cause something to change. For example, a tire manufacturer is an efficient cause for a car tire. These are sometimes referred to as mechanistic explanations.

- Final- The end or goal of something. Moving a vehicle is an efficient cause of a car tire. These are sometimes referred to as functional or teleological explanations.

It is important to note that although all the elements are individually necessary for an explanation, they are not individually sufficient.

Overton defines the structure of explanations with five categories of properties or objects that

are explained in science: theories, models, kinds, entities and data (Overton, 2012, 2013). To explain this concept further, let's look at the example of a billard ball hitting another ball at rest in a billard game. According to the example, the components of the model in Figure 1 will be as follows:

- Theory: Newton's third law to motion
- Model: F_A = -F_B .
- Kind X: If A exerts force upon B, B must exert force of equal and opposite magnitude.
- Entities A and B: Ball A and Ball B respectively.
- Data/Observation: A moving ball A hits ball B (initially at rest). The momentum of ball B increases and the momentum of ball A decreases due to equal and opposite forces exerted by each ball upon the other.

Extending Overton's Work, Malle argued social explanation has three layers: base, core and top (Malle, 2006).

- Base: Encapsulates underlying assumptions about human behaviour and explanation.
- Core: Psychological processes used to construct explanation.
- Top: Responsible for linguistic realization of the explanation.

Malle also proposed a theory of explanation (Malle, 2006, 2011), which breaks down the psychological processes used to offer explanations into two distinct groups information processes (for devising and assembling explanations), and impression management processes (for governing the social interactions of explanations). These two dimensions are divided into two further dimensions, which refer to the tools for constructing and giving explanations, and the explainer's perspective or knowledge about the explanation. This particular model is quite insightful due to its proximity with computer science domain. As presented in Figure 2, there are four dimensions:

- Information requirements - What information is required to provide an appropriate explanation?
- Information access - What information is accessible to the explainer to convey as an explanation?
- Pragmatic goals - What is the objective of the explanation?
- Functional capacities - What are the functional capacities or limitations of the given explanatory tools?

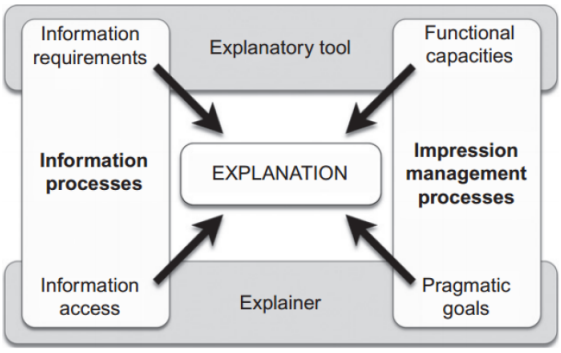

Figure 2: Malle's process model for explanation (Malle, 2011)

Some researchers portray explanations as representations of underlying causes which led to a system's output and which reflects the system's decision-making process. Other researchers assert that explanations are much more, for example when humans explain something they prioritize information, give examples and counter examples, select important details, etc. One of the most important factor of human explanations is the social component, explanations are almost always tailored for a specific audience or recipient. In Miller (2019), the author shares four important findings for explainable AI:

- Explanations are contrastive: In many scenarios, contrastive explanations are chosen by people due to their ease of comprehension.
- Explanations are selected: Humans usually don't analyze all causes of an event. They generally select a few main causes which are deemed "sufficient" as an explanation. The selected causes can be simple or complex, even global or local.
- Explanations are social: Explanation can be taken as transfer of knowledge, which includes many social aspects such as the receiver's cognitive and comprehension ability,

the explainer's and receiver's beliefs, knowledge and presumptions, etc. Explainer's presumption Recipient's prior knowledge comprehension capacities

- Probabilities probably don't matter: "The most likely explanation is not always the best explanation for a person, and importantly, using statistical generalizations to explain why events occur is unsatisfying, unless accompanied by an underlying causal explanation for the generalization itself." (Miller, 2019)

3 Aspects of Explanation

As mentioned earlier, explanation has multiple aspects. Philosophically acceptance of a statement as an explanation is subject to acceptance by the recipient. Although this is intuitive, this definition of explanation is vague and incomplete. We argue that important aspects of explanation include *the posed question*, *the type of explanation*, *the explainer* and *the recipient*. We discuss each of them in this section.

3.1 The Posed Question

The posed question defines the objective of the explanation and provides direction for selecting an abstract, comparatively simple, reasoning of a complex state. Much emphasis is put into researching explanation models with the assumption that the posed question is either a "why" or "why not" question. Even though the assumption is intuitive and holds true for most cases, in some cases the posed question could also be of the following type: "how", "what", "what if", etc.

Aristotle's Modes of explanation model mentioned in section 2.3, only caters to the why-questions. Miller (Miller, 2019) proposed a simple model for explanatory questions based on Pearl and Mackenzie's Ladder of Causation (Pearl and Mackenzie, 2018). This model places explanatory questions into three classes:

- What-questions, such as "What event happened?" – Requires associative reasoning to determine which unobserved events occurred based on the occurrence of observed events.
- How-questions, such as "How did that event happen?" – Requires interventionist reasoning to determine necessary and sufficient causes of the given event. This may also require associative reasoning.
- Why-questions, such as "Why did that event happen?" – Requires counterfactual reasoning to undo events and introduce hypothetical events. This also requires associative and interventionist reasoning.

3.2 The Type of Explanation

There is a wide spectrum from which an explanation can be produced. It is important to understand that there are numerous accurate possible explanation for each scenario. Mills argued that causal connection and explanation selection are essentially arbitrary and the scientifically/philosophically it is "wrong" to select one explanation over another (Mill, 1973).

Based on Lipton (2018), Mittelstadt et al. (2019) and Pedreschi et al. (2019), we deduce that an explanation can be of many types. It can be example-based or statistical, generalized or specialized, technical or in layman terms. We argue that the choice between the latter two may be deduced by the information available about the posed question and the recipient. However, Adhikari (2018) shows via experiments that in most cases example-based explanations are preferred.

3.3 The Explainer

As mentioned in the introduction, the impact of unquantifiable attributes of human behavior, which includes belief, desire, social norms, intention to understand, emotions, personality traits, etc., can affect explanation. Even though the explainer's intention to explain and her beliefs about the recipient(s) may be important, we argue that there are other aspects which can be just as important (if not more). These aspects include the knowledge of the explainer herself, her ability to explain and the availability of information about the recipient's knowledge. This argument becomes stronger when we remove the pre-assumption of the explainer being a human. If the explainer is a machine then the importance of social aspects reduces.

3.4 The Recipient

The recipient is arguably the most important aspect for any explanation as many other aspects are manipulated by the recipient. Similar to the explainer,

even though the social aspects such as intention to understand and her belief system are important, we argue that recipient's knowledge and ability to comprehend is also important. Although for building an explanation system, we would not relax the pre-assumption of human recipient(s).

4 Proposed Definition

In this section we present a generic definition, break down its components which can be interpreted differently depending on perspectives and discuss them from the AI perspective. Here, if an explanation is requested for a certain prediction/recommendation made via a particular black-box, we will represent the combination of input to the black-box, the black-box itself and the prediction/recommendation as a *scenario*. We argue that explanation requires pre-existing information about three major aspects: *the posed question*, *the recipient's knowledge*, and *the given scenario*. With respect to the aforementioned three aspects, we propose the following definition of explanation:

An explanation is a representation of fair and accurate assessments made by an explainer to transfer relevant knowledge (about a given scenario) from the explainer to a recipient.

In AI perspective, each component of the definition can be translated as follows:

- Assessments: It can be plain observations or analysis of observations. In AI, these observations tend to be about internal workings of a system and about factors which are important for a given input-output pair.
- Fair and accurate assessments: Here, *fair and accurate* depicts unbiased observations which hold high fidelity w.r.t the given AI model.
- Representation: It must be an appropriate depiction of assessments made by the explainer and acceptable by the recipient. It can be in the form of text, image, dialogue, etc. or a combination thereof.
- Explainer: It is an entity, human or machine, that mediates information. In AI, the explainer is a machine which reduces the impact of social aspects.
- Recipient: It is an entity to which clarification of a subject matter is presented. This clarification may be in the form of a statement or a response to a posed question.
- Transfer of Relevant Knowledge: The concepts of explainability (complete and complex) and interpretability (abstract and comprehensible) of an explanation can be generalized by using "transfer of relevant knowledge" as a function of information gained by the recipient. This transfer of relevant knowledge can be active or passive.

Explanation as a statement represents a passive form of knowledge transfer based on only the given scenario. On the other hand, explanation as a response to a question posed by a recipient represents an active form of knowledge transfer based on the posed question, the recipient's knowledge, and the given scenario.

4.1 Examples

Here we consider three different scenarios in which the same underlying logic is conveyed in different ways to the recipient according to their knowledge and experience, when a particular question is asked. Please note that these examples only provide naïve insights into possible explanations. In reality the recipient's knowledge might be a complicated model which cannot be allotted one of the basic groups such as novice, intermediate or expert. The question "How do we model a recipient's knowledge?" is one of many difficult questions which need to be answered in order to build such complex explanation systems.

In Table 1, Scenario (1) and (2) offers contrast in generated explanations for the same type of recipient's knowledge. In Scenario (1), the explanation for a novice in the medical domain is kept very simple but the expert explanation has a lot of medical details, which might be burdensome for most patients. On the other hand, in Scenario (2) the banking domain novice explanation gives reasons of a loan rejection in a detailed yet simple manner, while the explanation for an expert is short. This uncovers the possibility of having different types of explanation depending on not only the scenario but also the domain of the scenario. Since it is not feasible to create different models for every scenario or domain of scenario, we need a generalized explanation system which can adapt and learn.

Scenario	Question	Logic Found	Recipient's Knowledge	Explanation
(1) A medical treatment X was recommended for a patient with symptoms Y.	Why was treatment X recommended to me?	Treatment X worked for other patients with symptoms Y	Novice	This treatment has worked before.
			Intermediate	Similar symptoms were healed with treatment X before.
			Expert	This treatment has a high probability of success as the similarity between the current patients and a past patient, who responded to treatment X, is above 70% due to the following underlying facts ....
(2) A bank loan request was rejected.	Why was my loan request rejected?	Risk factors were found too high due to already existing loans.	Novice	Since there is a lot of money already loaned to the applicant, the return of the requested amount to the bank becomes improbable.
			Intermediate	The amount loaned already is high which makes this a risky request for the bank to approve
			Expert	Due to high risk implied by multiple prior loans.
(3) A certain movie was recommended.	Why was this movie recommended to me?	98% match with other movies in user profile	Novice	You watched similar movies.
			Intermediate	There is 98% match with similar movies that you have watched before.
			Expert	You will probably like this movie as there is a 98% match with similar movies that you have watched before, due to the following reasons...

Table 1: Here we present three examples of a scenario with three possible types recipients each i.e. novice, intermediate and expert.

4.2 Proposed Architecture for Definition

The need for a generalized explanation system which can adapt and learn has been discussed in the previous sections. Application of the presented definition from the perspective of computer science is complicated and ambitious, it has multiple factors to be considered. In order to take another step towards building explanation systems, we take inspiration from classic *Natural Language Generation* (NLG) Model to propose an architecture for generation of user understandable explanations for desired scenarios (Dale and Reiter, 1997; Reiter and Dale, 2000).The functionality of each component of the architecture is provided below.

In Figure 3, the classical NLG pipeline of Content Determination – Content Refining – Content Lexicalization – Text Generation consists of several components. The initial input of the pipeline is black box's input-output pair which is fed to the first component (Content Determination). Following which each component takes as input the output of the previous component, together producing the following sequence of outputs: Logical Content – Structured Content – Lexicalized Content – Generated Text/Explanation. This entire setup can be placed in an environment responsible for learning models about the individual recipients' preferences. For each query, multiple explanations will be generated and ranked by the learning environment. Based on recipient's choice, the corresponding model will be refined for future queries.

Each component of this architecture brings forward a different challenge. The most intriguing ones are from the section of content determination and the learning environment. Feature Importance Extraction has been researched vividly in recent years with black box solutions such as LIME (Ribeiro et al., 2016a), LEAFAGE (Adhikari,

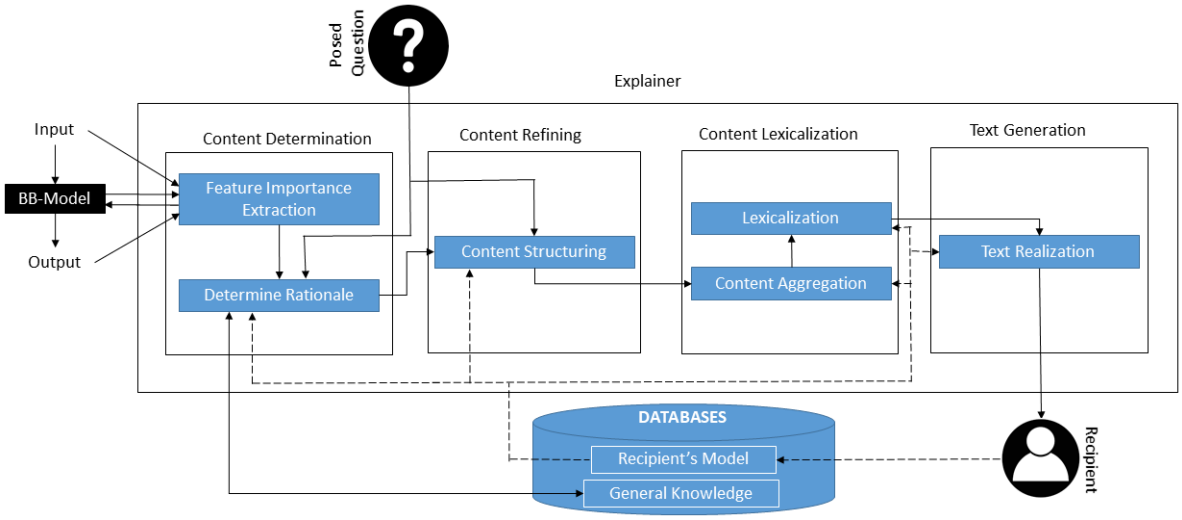

Figure 3: A possible architecture of a generic explanation system based on the proposed definition. This is a slightly modified setup of the classical NLG pipeline. This setup will be placed in an iterative learning environment for refining individual recipients' models.

2018), XEMP [2], etc. However, very little attention has been given to bridging the gap between black box inputs and human interpretable features in a generic way.

The suggested architecture enables us to partly mimic Overton's structure of explanations and Malle's structure of social explanations (section 2.3) via the NLG pipeline and the learning environment. In reference to Overton's structure of explanations, the NLG pipeline receives an *input-output pair* as *data/observation* and extracts the *model* as *Logical Content*. Similarly, the learning environment will capture the recipient's underlying behavior as described in Malle's structure of social explanation.

Also, by generating multiple explanations we acknowledge Miller's four important findings for explainable AI (section 2.3) and Mill's theory of selection of an explanation (section 3.2). We argue that there exists multiple possible explanations for each query. Each explanation generated by following the suggested procedure and supported by factual data cannot be "wrong". However, the system must be iteratively trained based on the recipients' acceptance of explanations and gain of information. This makes the learning environment a crucial part of the system.

[2]XEMP White Paper (Last accessed on: 2020-10-06): www.datarobot.com/resource/xemp-prediction-explanations/

5 Conclusion

In this paper, we review multiple perspectives and aspects of explanation. Based on this, we summarize the most important aspects of explanation and propose a generic definition. The definition and examples of explanation in different scenarios guide us towards building an explanation system. We propose an architecture for one such generic explanation system based on the proposed definition of explanation. It is capable of producing multiple explanations depending on the given recipient, the posed question and a given scenario. We hope that this paper paves the way to build generic explanation systems.

6 Future Work

Each of the components of the proposed architecture is challenging in a different way. The challenges include the process of creating a recipient model, refining the recipient's model based on their preferences, making the explanation system generic, etc. The learning environment will also bring forth a cold-start problem. These are some of the many ambitious topics which need to be researched, analyzed and applied. We intend to start by creating a prototype of a content determination module from the proposed architecture of a generic explanation system.

References

Ajaya Adhikari. 2018. Example and feature importance-based explanations for black-box machine learning models.

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13.

Robert Dale and Ehud Reiter. 1997. Tutorial on building applied natural language generation systems. In *ANLP-97*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Robert James Hankinson. 2001. *Cause and explanation in ancient Greek thought*. Oxford University Press.

Norwood Russell Hanson. 1958. *Patterns of discovery: An inquiry into the conceptual foundations of science*, volume 251. CUP Archive.

Gilbert H Harman. 1965. The inference to the best explanation. *The philosophical review*, 74(1):88–95.

Fritz Heider and Marianne Simmel. 1944. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259.

Germund Hesslow. 1988. The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, pages 11–32.

Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65.

John R Josephson and Susan G Josephson. 1996. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8.

David Lewis. 1986. Causal explanation, in his philosophical papers, vol. 2.

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266.

Zachary C Lipton. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57.

Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press.

Bertram F Malle. 2011. Time to give up the dogmas of attribution: An alternative theory of behavior explanation. In *Advances in experimental social psychology*, volume 44, pages 297–352. Elsevier.

John Stuart Mill. 1973. A system of logic, books i-iii, volume vii of collected works.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288.

James A Overton. 2012. *Explanation in Science*. Electronic Thesis and Dissertation Repository.

James A Overton. 2013. "explain" in scientific discourse. *Synthese*, 190(8):1383–1405.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.

Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. 2019. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9780–9784.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016a. " 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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016b. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.