Explanation Perspectives from the Cognitive Sciences—A Survey

Ramya Srinivasan[*] and **Ajay Chander**
Fujitsu Laboratories of America
{ramya, ajay}@fujitsu.com

Abstract

With growing adoption of AI across fields such as healthcare, finance, and the justice system, explaining an AI decision has become more important than ever before. Development of human-centric explainable AI (XAI) systems necessitates an understanding of the requirements of the human-in-the-loop seeking the explanation. This includes the cognitive behavioral purpose that the explanation serves for its recipients, and the structure that the explanation uses to reach those ends. An understanding of the psychological foundations of explanations is thus vital for the development of effective human-centric XAI systems. Towards this end, we survey papers from the cognitive science literature that address the following broad questions: (1) what is an explanation, (2) what are explanations for, and 3) what are the characteristics of good and bad explanations. We organize the insights gained therein by means of highlighting the advantages and shortcomings of various explanation structures and theories, discuss their applicability across different domains, and analyze their utility to various types of humans-in-the-loop. We summarize the key takeaways for human-centric design of XAI systems, and recommend strategies to bridge the existing gap between XAI research and practical needs. We hope this work will spark the development of novel human-centric XAI systems.

1 Introduction

Recently, there has been a surge of AI-based applications. Amidst this, regulatory bodies, policy makers, and consumers have expressed concern about AI being a black-box technology. As a result, there has been a growing interest in making AI technologies explainable. These include regulatory reforms [GDPR, 2017], government endeavours [Gunning, 2017], industry initiatives [PWC, 2018; Kyndi, 2018] and academic papers such as [Koh and Liang, 2017; Park *et al.*, 2018; Floyd and Aha, 2017; Vigano and Magazzeni, 2018]. Yet, most of these efforts largely cater to the needs of AI scientists or domain experts, and seldom address the explanation needs of other types of stakeholders such as business executives or consumers.

As an instance, consider the following explanation of an AI based loan decision: "Person A was denied a loan because their credit score was similar to person B". Does such an analogical explanation help all kinds of humans-in-the-loop? Consider another explanation of the following form: "If $w1 < 0.02$ and $w2 > 0.4$, then predict = loan denied." How generalizable are such rule-based explanations across different datasets? While such explanations might help an AI engineer in debugging the system, it offers little to no value to the loan applicant [Chander *et al.*, 2018].

Thus, there are several design aspects that need to be considered in building XAI systems. First, explanations should serve a variety of stakeholders such as customers, business executives, software engineers, regulatory bodies, policy makers etc., and not just domain experts or AI scientists [Ras *et al.*, 2018]. Second, an explanation may need to serve different purposes based on the requirements of the human-in-the-loop. Some such purposes of explanations could include fostering trust in the users, educating the users about the model's decision, assisting users in taking appropriate actions, etc. [Chander *et al.*, 2018]. Third, the effectiveness of XAI systems have to be justified not just quantitatively by means of algorithmic validation, but also qualitatively by estimating factors such as user satisfaction, trust assessment, and other social factors [Miller, 2018]. Thus, robust evaluation strategies have to be designed [Narayanan *et al.*, 2018; Doshi and Kim, 2017].

An understanding of explanation's context is necessary for crafting effective human-centric explanations. Context includes information such as who is providing the explanation, who is seeking the explanation, what is the purpose of the explanation, what is the use-case, and so on. More fundamentally, there is a need to understand what constitutes a good explanation for a particular type of user and use-case, i.e., what sort of explanation structures (analogical, causal, rule-based, etc.) are best suited for different types of users and use-cases.

Motivated by the aforementioned considerations, we analyze psychological underpinnings of explanations by means of a cognitive science literature survey. A closely related work to our survey is that by [Miller, 2018]. Our survey differs from [Miller, 2018] in at least three aspects. First, unlike

[*] Contact Author

[Miller, 2018] wherein the emphasis is more on the social aspects of explanation, our survey focuses more on the cognitive behavioral purpose an explanation serves for individuals, and the structure that the explanation uses to reach those ends. Second, we illustrate characteristics of good and bad explanations, discuss suitability of various explanation structures based on the application domain, and analyze their utility for various kinds of humans-in-the-loop. Last but not the least, in order to enhance both understanding and adoption of the theories propounded in cognitive sciences, we provide a *succinct* survey with practical insights and recommendations for XAI designers.

We survey papers from the cognitive science literature that address the following broad questions: What is an explanation, What are explanations for and What are the characteristics of good and bad explanations. Section 2 discusses various theories of explanations highlighting their applicability and shortcomings. Section 3 describes the functions of explanations. The characteristics of good and bad explanations are discussed in Section 4. Section 5 outlines insights gained, and our takeaways for the XAI community.

2 What is an Explanation?

Theories of explanation date back at least as far as the times of Plato and Aristotle, who divided explanations into at least four basic modes [Lombrozo, 2006]. There are several contemporary theories of explanations and a vast body of research concerning the nature of explanations [Joseph, 1988; Friedman, 1974; Achinstein, 1983; Harman, 1965; Ruben, 1990]. In this section, our goal is to provide a succinct organization of various theories, mention where they may be best suited, and state their shortcomings.

2.1 Explanations as Deductive Proofs

One of the earliest contemporary works concerning the definition of explanations was proposed by [Hempel and Oppenheim, 1948]. Here, explanations are believed to be like proofs in logic. A set of basic laws are stated as axioms and the deductive sequences involved in the proofs constitute the explanation. For example, an explanation related to the buoyancy of objects might assume certain laws of density and properties of the objects under consideration. These laws are then leveraged to explain why the object might float or sink.

Where such Methods are Best Suited

Deductive proof-like explanations are useful in applications where there are well defined laws governing the phenomena of interest [Hempel and Oppenheim, 1948]. Thus, deductive rule based explanations such as "If $w1 < 0.02$ and $w2 > 0.4$, then predict = loan denied" can be understood only by AI engineers who know what the weights $w1$ and $w2$ correspond to. More broadly, such types of deductive explanations are primarily suitable for domain experts in scientific fields such as logic, physics, mathematics, etc. [Rosemary, 1999].

Shortcomings

These models do not generalize easily. As one considers explanations across many disciplines, even superficial similarities to deductive chains start to disappear [Salmon, 1989]. For laypeople, the deductive model seems even less feasible. For example, people frequently prefer one explanation to another without explicitly being able to say why. They often seem to draw on implicit explanatory understandings that are not easy to put in explicit terms [Kozhevnikov and Hegarty, 2001]. Moreover, these methods fall short in explaining the mechanism of the phenomena and hence are not easily comprehensible by lay people— Many researchers believe that merely stating laws such as "if X, then Y", are just means of citing "effects," and that they are not really explanations as the mechanism governing the process is still opaque [Cummins, 2000]. Current XAI systems fall short in explaining the mechanism of failures and instead rely on rule based explanations which are often opaque to lay people.

2.2 Explanations as Causal Patterns

Several theories state that explanations often refer to causal relations. It is believed that there are at least four distinct ways in which these relations can be perceived. These are as stated below:

- *Common Cause:* In common-cause explanations, a single cause is seen as having a branching set of consequences [Sober, 1984]. In such scenarios, tracing the path to the root cause provides the required information in explaining a phenomena. Common-cause explanations are frequently found in diagnoses of problems, such as in medical disease, equipment malfunction, or software bugs [Keil, 2006].

- *Common Effect:* Common-effect explanations involve cases where several causes converge to create an event. These sorts of explanations are common in history, environmental science, and economics, wherein a major event might be attributed to the confluence of several factors.

- *Linear Chains:* Explanations as simple linear chains are a special limited case of common cause and common effect explanations; namely, there is one unique serial chain from a single initial cause through a series of steps to a single effect [Keil, 2006]. That said, simple linear explanations may be quite rare in real life. Even if things start with a single chain of effects and causes, at some point those effects start to have multiple effects of their own and the structure starts to branch.

- *Causal Homeostasis:* Causal homeostatic explanations seek to account for interlocking sets of causes and effects resulting in a set of properties [Boyd, 1999]. A causal homeostatic explanation does not seek to explain how a cause progresses over time to create some effects, but rather how an interlocking set of causes and effects results in a set of properties enduring together as a stable set over time that then exists as a natural kind. For example, this theory can explain why feathers, hollow bones, nest building, flight, and a high metabolic rate might all reinforce the presence of each other in birds [Boyd, 1999].

Where such Methods are Best Suited

Explanation based on causal patterns has been very popular in domains such as epidemiology, economics, marketing, environmental sciences, law, policy making, and in medicine, domains in which the structure of the physical world can be well-established. Causal patterns unveil the underlying mechanism behind phenomena and thereby provide a more intuitive explanation than deductive proofs. As a result, these methods are applicable to a variety of stakeholders, including lay people.

Shortcomings

Many philosophers believe that some causal patterns are highly domain specific. For example, causal homeostasis is mostly applicable to living entities [Keil, 2006; kyoung Ahn, 1998]. Another concern regarding the use of causal models is that the underlying causal structure often stems from certain assumptions which are in turn based on human subjectivity. Thus, for the same problem, the structure could vary from person to person based on how they perceive the underlying mechanisms.

2.3 Explanations as Mental Models

This approach incorporates aspects of both traditional AI and neuroscience and makes use of the idea of a mental model [Holland *et al.*, 1986]. Mental models are internal representations that occur as a result of the activation of some part of a network of condition-action type rules. These rules are clustered in such a way that when a certain number of conditions becomes active, some action results. For example, a mental model of a squirrel can be described as an activation of rule *If (small, hops, chirps) then (squirrel)* [Mayes, 2019]. By definition, these models are hierarchical. Essentially, when the expectations activated at a certain level of the default hierarchy fail, the system searches lower levels of the hierarchy to find out why. Thus, in this view, explanation is mostly a neurological process and explanatory understanding is understood by reference to activation patterns within a human [Mayes, 2019].

Where such Methods are Best Suited

It is widely believed that explaining is a purely cognitive activity, and an explanation is a certain kind of mental representation that results from or aids in this activity [Mayes, 2019]. Thus, such methods of explanation are well suited for situations in which the purpose of explanation is to aid in communication. For example, a teacher could educate new concepts (e.g. a new animal or plant) to children based on some mental models she has. Other examples of mental models of explanation could include counseling-based use-cases (e.g. a counselor trying to communicate a strategy to a user seeking help based on the mental models of treatments that the counselor has).

Shortcomings

Mental models can range from logical patterns [Johnson-Laird, 1983] to image-like representations of the workings of a system [Gentner and Stevens, 1983] and other spatial representations. Thus, it has been argued that explanations cannot simply be a read-out of mental blueprints, but must include the interpretations of such blueprints. It has been widely perceived that explanations based on neural activation are largely black-box models offering little insights about the internal functionalities of a system.

2.4 Explanations as Stances

Some papers define explanations as stances or modes. People may adopt a stance or mode of construct to frame an explanation. A single phenomenon often can be characterized by different stances and thereby yield quite different insights and explanations. These stances could include:

- *Mechanical Stances:* The idea here is to use physical objects and interactions to explain a phenomenon or effect. For example, the action of a driver taking a sharp turn may be explained by means of presence of adjacent vehicles, the mechanics of wheel movement, etc. Thus, explanations are constructed leveraging the presence and functions of physical objects and the interactions these objects might have with the rest of the environment.

- *Illustrative Stances:* The idea here is to provide explanations by means of different kinds of illustrations. Illustrations could include examples or demonstrations, comparisons and analogies, contradictions or counterfactual reasoning, or by highlighting salient and distinguishing features pertaining to a phenomenon of interest. By far, this is one of the most popular means of explanations. Explanations by means of highlighting salient regions, counterfactual reasoning [Goyal *et al.*, 2019], explanations by means of contrasting examples [Kanehira and Harada, 2019] are already in use within the XAI community. Analogical stances are also gaining popularity with the XAI community. Typically, when explanations are being provided to people with reasonable knowledge within an area, close analogies are used; but when explanations are to be provided to a larger section of lay people, analogies between more distant domains may become more common [Gentner, 1983]

- *Intentional Stances:* These are stances concerning general beliefs and desires among people. Some argue that these are similar to explanations based on mental models. A hypothetical example of an intentional stance could be "there is resentment concerning the proposal as there is a frown on people's face". Here, the notion of resentment is being attributed to frown and frown is being used as an intentional stance to explain resentment [Dennett, 1987]. However, these stances are highly subjective and not predictable easily.

Where such Methods are Best Suited

Explanations via stances are suited for scenarios where there are well established relations and properties. Such relations and properties could include feasible object-object, person-object, and person-person interactions, and spatio-temporal relations. Such methods are also suited in applications with illustrative instances that aid in understanding. These methods are very popular in the XAI community.

Shortcomings

It has been argued that stances are not in themselves theories, but are rather modes or structures of explanations. Another concern about using stances for explanations is that they are vague and sometimes not predictable (especially those concerning intentional stances). XAI systems which use analogical stances as explanations must note that distant analogical stances can cause confusion to the person seeking the explanation.

In the next section, we discuss the functions of explanations.

3 What are Explanations for?

Understanding the purpose of an explanation is crucial in crafting effective explanations. The authors in [Miller, 2018] state that explanations are social in that they are meant to transfer knowledge, presented as part of a conversation or interaction and are thus presented relative to the explainer's beliefs about the user's (i.e., explainee's) beliefs. In a similar vein, the authors in [Chander and Srinivasan, 2018] state that an explanation is a filter of facts in a given context. The context includes various factors such as the human-in-the-loop providing the explanation (the explainer) and the person seeking it (the explainee), the communication medium, the reason for seeking an explanation (purpose of an explanation), among other factors.

Consider for example the scenario of explaining an AI-based loan denial to a loan applicant. While an explanation such as "person A was denied a loan as his profile was similar to person B" might be reasonable to an AI engineer to understand and verify the model, it offers little to no value to a loan applicant who wants to understand how to secure a loan or what aspect of their application contributed to the denial. Thus, there are different functions of an explanation based on the human-in-the-loop in need of the explanation.

The authors in [Chander and Srinivasan, 2018] argue that explanations should serve a cognitive-behavioral purpose. Some of these purposes are:

- **Trust:** Some users may primarily expect explanations that speak to how their personal values (e.g., privacy, safety, etc.) are met by AI systems. In this case, the cognitive value of the explanation is to *engender trust* in the user. For example, a user of an online banking platform might want an explanation concerning the safety of a transaction.
- **Troubleshooting and/or Design:** Some other users may largely expect explanations to describe functional aspects of the AI models such as accuracy, speed and robustness. Here, the cognitive value of the explanation is in aiding *troubleshooting and/or design*. For example, an AI engineer might want to know why there were a lot of false positives in some classification task.
- **Education:** Some users may expect explanations to help them understand the AI's recommendation and aid them in analysis. In this case, the cognitive value of the explanation is in *educating* the user. For example, a sales representative might want to know why the AI model recommended a certain product to a customer.
- **Action:** Some users may expect explanations to help them take an appropriate *action* based on the AI model's decision. For example, a loan applicant who was denied a loan might want to know what they need to do to secure a future loan.

In addition to the above, there may be other purposes an explanation might serve, as noted in the following works:

- **Justification:** Sometimes explanations may be to *justify or rationalize* an action. Explanations are attempts to represent our actions to others as sensible, well intentioned, pragmatic, or appropriate. For example, we explain that we punished a child for his own good, or that we didn't bother voting in an election because our vote would not have mattered anyway [Keil, 2006].
- **Aesthetics:** Explanations could simply be for the purposes of *aesthetic pleasure*. For example, one could explain the intricacies of a poem with the sole goal of increasing appreciation in another reader [Keil, 2006].
- **Communication:** Explanations are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs and facilitate *social and emotional communication* [Miller, 2018].

A good summary of other functions of explanations can be found in [Lombrozo, 2006]. Independent of the choice of theory behind an explanation or the function of an explanation, a good explanation is characterized by certain features. In the next section, we enlist characteristics of good and bad explanations.

4 What are the Characteristics of Bad Explanations and Good Explanations?

Understanding the characteristics of bad explanations is useful in evaluating explanations. There are at least four primary features of a bad explanation as listed below:

4.1 Characteristics of Bad Explanations

- **Circularity**: Consider an explanation of the form—"This diet pill is effective because it helps people lose weight". This explanation is actually not providing any substantiation; instead, it is circling back to the problem statement. Repetition of facts and assertion of obvious factors cannot be considered as good explanations. Moreover in the case of complex and lengthy circularity, people are often confused and cannot evaluate the effectiveness of an explanation [Rips, 2002].
- **Lack of Relevance:** Relevance is essential for a good explanation. Within the XAI community, people have identified the importance of relevance in crafting good explanations [**?**]. Yet, levels of abstraction, analogies, and surprising connections to other domains can complicate the assessment of relevance and make an explanation bad [Keil, 2006]. Furthermore, an explanation will fail either if it provides too much detail or if it presupposes too much and skips over essential details. This is commonly known as the egocentric bias [Ross *et al.*,

Examples	Reason
The hotel is affordable because it is close to the zoo	lack of relevance
The diet pill is effective because it helps reduce weight	circularity
This insurance is the best option because it has the highest premium	lack of coherence

Table 1: Examples of Bad Explanations

1976]. This is essentially the tendency to rely too heavily on one's own perspective than reality. It appears to be the result of the psychological need to satisfy one's ego, hence the name. Such a bias will make an explanation highly subjective and thereby unreliable.

- **Lack of Coherence:** Explanations can also be seen as bad if they fail to cohere, or "hang together". The different elements of an explanation must work in conjunction to achieve a consistent justification. An example of an incoherent explanation could be—"This insurance is the best available option because it has the high premium". Here, the fact that the insurance has a high premium is contradicting the statement that it is the best. Inconsistent and contradictory explanations are often disregarded by people [Keil, 2006] as they contradict halfway through the explanation.
- **Lengthy:** An explanation that is long often causes confusion and makes the explanation less comprehensible [Narayanan *et al.*, 2018]. Longer explanations demand more processing time for the humans involved in the loop. This, in turn, contributes to lower satisfaction for the explainee.

Table 1 provides some illustrations of bad explanations along with the reason for the explanations to be deemed as bad.

4.2 Characteristics of Good Explanations

There has been substantial research concerning the characteristics of good explanations [Lipton, 2004; Rosemary, 1999]. These are valuable in the design and evaluation of XAI systems. Below, we list some primary features that are characteristics of good explanations.

- **Simplicity:** Good explanations should be simple. People prefer simpler explanations because they are easier to comprehend.
- **Robustness:** Explanations should generalize to new contexts, exhibiting a high degree of "goodness-of-fit" [Johnson *et al.*, 2017].
- **Address the gap between knowledge and understanding:** A good explanation should help reduce the gap between knowledge and understanding [Lipton, 2004]. Knowing the reason behind a particular phenomenon is necessary but not sufficient for understanding why it is the reason. Thus, a good explanation should not only state the reason but also explain why something is the reason.
- **Self-evidencing:** Good explanations are those where what is explained provides an essential part of our reason for believing that the explanation itself is correct [Lipton, 2004]. Good explanations should be testable or verifiable whenever feasible, which is an aspect of explanation's self-evidence.

5 Insights

In this section, we summarize what we learned from the survey and what we recommend to the XAI community.

5.1 Lessons from Explanation Theories

- There is no one theory that is universally applicable. The choice of a theory is determined by the context surrounding the explanation, and the purpose or the function of an explanation.
- Certain types of explanation theories are more suitable for certain domains. For example, explanations by deductive reasoning may be more suitable for domains such as physics where there are well-defined laws explaining several phenomena.
- Explanation by means of illustrative stances is one of the popular methods of explanations. Such illustrative stances are applicable to a wide set of users and domains. This is because these explanative stances offer a variety of ways to construct the explanations and thereby aid different types of explainees. For example, to explain to a consumer why they should buy a product, it may be helpful to explain via analogies to other people who purchased the product. A business executive may better understand why they should invest money in a certain project through counterfactual reasoning about what would happen if they did not make the investment.

5.2 Lessons Considering Human-in-the-Loop

- The choice of an explanation mechanism should be determined based on the human-in-the-loop seeking the explanation—For example, a lay person may prefer an explanation based on causal patterns as opposed to explanations based on deductive proofs. Thus, the choice of an explanation structure/mechanism should be determined based on the explainee.
- Appeal of an explanation depends on the type of explanation structure, the explainee, and also the cognitive task requiring the explanation. This result is intuitive—some tasks are inherently harder to explain, and also it is hard to explain to some types of users who have limited comprehension skills.
- Explanations should serve a behavioral purpose such as educating the user, engendering trust in the user, helping with system design and debugging, and so on.
- Some characteristics of good explanations include: shortness/terseness, simplicity, and coherence. Furthermore, good explanations should be compelling to the human-in-the-loop.
- Good explanations should be verifiable whenever possible.

5.3 What we Recommend

In order to bridge the gap between research and practice and to accelerate the practical adoption of XAI across a variety of domains, we suggest:

- **Adoption of a user-centered approach in generating explanations**.
  As stated earlier, most XAI systems are designed to provide explanations to AI engineers or data scientists, and offer little to no value to other types of users. Thus, understanding the type of user seeking the explanation is essential in the design of efficient XAI systems.

- **Incorporation of domain-specific knowledge in the AI models in order to enhance explainability**.
  In some scientific applications, domain knowledge can aid in providing explanations. By incorporating this knowledge, it is possible to reduce inconsistencies in model explanations [Karpatne *et al.*, 2017] and reduce the probability of a model generating unrealistic explanation. Note, by domain knowledge, we do not mean knowledge specific to AI, but knowledge specific to a scientific discipline where an explanation is sought. For example, in providing an explanation to a physicist regarding AI-based estimation of lake temperature, some principles regarding density and flow of water from physics may be used [Karpatne *et al.*, 2017] to avoid fact inconsistencies and model overfitting. It is to be noted that incorporation of such domain-specific knowledge is beneficial in providing explanations mostly to domain experts.

- **Understanding of the explanation context: who is seeking explanation, what is the purpose of the explanation, etc.**
  Explanations are meant to bridge the knowledge gap between the person providing the explanation (explainer), and the person seeking it (explainee). Furthermore, explanations should serve a purpose such as educating the user, help them with appropriate actions, etc. Thus, it is important to design explanations keeping in mind the human-in-the-loop seeking it, and the purpose an explanation is trying to serve. Thus explanations need to be personalized to satisfy the explainee.

- **Creation of new datasets along the lines of aforementioned points**
  Studies have shown that existing datasets often contain features that are seldom understood by lay people [Srinivasan *et al.*, 2018]. It therefore becomes necessary to build new datasets to understand the nature of user friendly explanations and to train new AI models using these datasets. In creating such datasets, it is very much possible that one may encounter unfamiliar or uncommon explanations for a particular use case. It is important to include these uncommon responses as long as they are valid in order to capture diversity and to address the needs of various kinds of users.

- **Design of new algorithms to cater to the limitations of datasets (e.g. a small dataset)**
  Datasets that are representative of user friendly explanations may not be as big as other datasets typically used in training machine learning models. In this context, it may become necessary to design robust algorithms that do not require large training data.

- **Evaluation of XAI systems based on how the explanations satisfy the purpose of the human-in-the-loop**.
  The evaluation of an explanation must include metrics that determine how the explanation satisfies the explainee's purpose. As stated earlier, purposes could include educating the explainee, help them in taking appropriate actions based on the model's decision and so on. Thus, incorporating user-experience studies is necessary in the evaluation of XAI systems, in addition to algorithmic validation. A survey of some such techniques may be found in [Lage *et al.*, 2019].

6 Conclusions

Studies from the cognitive sciences offer enriching perspectives for the design of XAI systems. These studies emphasize the role of the human-in-the-loop seeking explanations and study the structure and functions of good explanations from the viewpoint of the human-in-the-loop seeking the explanation. In this paper, we provided a survey of some of these studies addressing aspects of what is an explanation, what are explanations for, and what are the characteristics of good and bad explanations. We highlighted the applicability of various explanation structures and their shortcomings. We outlined the insights gained from the survey and laid out some guidelines for the design and development of future XAI systems.

References

[Achinstein, 1983] Peter Achinstein. The nature of explanations. *Oxford University Press*, 1983.

[Boyd, 1999] Richard Boyd. Homeostasis, species, and higher taxa. *Species: New Interdisciplinary Studies-MIT Press*, 1999.

[Chander and Srinivasan, 2018] Ajay Chander and Ramya Srinivasan. Cdmake workshop on make explainable ai. *Evaluating Explanations by Cognitive Value*, 2018.

[Chander *et al.*, 2018] Ajay Chander, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino. Working with beliefs: Ai transparency in the enterprise. *Explainable Smart Systems Workshop, Intelligent user Interfaces*, 2018.

[Cummins, 2000] Robert Cummins. How does it work?" versus "what are the laws?" two conceptions of psychological explanation. *Explanation and Cognition*, 2000.

[Dennett, 1987] Daniel Dennett. The intentional stance. *MIT Press*, 1987.

[Doshi and Kim, 2017] Finale Doshi and Been Kim. Towards a rigorous science of interpretable machine learning. *ArXiv*, 2017.

[Floyd and Aha, 2017] Michael Floyd and David Aha. Using explanations to provide transparency during trust-guided behavior adaptation. *AI Communication*, 2017.

[Friedman, 1974] Micheal Friedman. Explanations and scientific understanding. *Journal of philosophy*, 1974.

[GDPR, 2017] GDPR. General data protection regulation. *EUGDPR*, 2017.

[Gentner and Stevens, 1983] Dedre Gentner and Albert Stevens. Mental models. *Hillsdale NJ*, 1983.

[Gentner, 1983] Dedre Gentner. Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, 1983.

[Goyal *et al.*, 2019] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and S. Lee. Counterfactual visual explanations. *ICML*, 2019.

[Gunning, 2017] Dave Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA)*, 2017.

[Harman, 1965] Gilbert Harman. The inference to the best explanation. *Philosophical Review*, 1965.

[Hempel and Oppenheim, 1948] Carl G. Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175, April 1948.

[Holland *et al.*, 1986] John Holland, Holyoak Keith, Nisbett Richard, and Thagard Paul. Induction: Processes of inference, learning, and discovery. *Cambridge University Press: MIT*, 1986.

[Johnson *et al.*, 2017] Samuel Johnson, J. J. Valenti, and Frank C. Keil. Opponent uses of simplicity and complexity in causal explanation. *CogSci*, 2017.

[Johnson-Laird, 1983] Philip Johnson-Laird. Mental models. *Harvard University Press*, 1983.

[Joseph, 1988] Pitt Joseph. Theories of explanations. *Oxford University Press*, 1988.

[Kanehira and Harada, 2019] Atsushi Kanehira and Tatsuya Harada. Learning to explain with complemental examples. *CVPR*, 2019.

[Karpatne *et al.*, 2017] Anuj Karpatne, Gowtham Atluri, James Faghmous, Michael Steinbach, Arindam Banerjee, Aroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[Keil, 2006] Frank Keil. Explanations and understanding. *Annual Review of Psychology*, 2006.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *International Conference on Machine Learning*, 2017.

[Kozhevnikov and Hegarty, 2001] Maria Kozhevnikov and Mary Hegarty. Impetus beliefs as default heuristics: dissociation between explicit and implicit knowledge about motion. *Psychon Bull Rev.*, 2001.

[Kyndi, 2018] Kyndi. How explainability is driving the future of artificial intelligence. *White Paper*, 2018.

[kyoung Ahn, 1998] Woo kyoung Ahn. Why are different features central for natural kinds and artifacts? the role of causal status in determining feature centrality. *Cognition*, 1998.

[Lage *et al.*, 2019] Isaac Lage, Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Greshman, and Finale Doshi. An evaluation of the human-interpretability of explanation. *ArXiv*, 2019.

[Lipton, 2004] Peter Lipton. What good is an explanation. *Explanations: Styles of Explanation in Science: Oxford University Press*, 2004.

[Lombrozo, 2006] Tania Lombrozo. The structure and functions of explanation. *Trends in Cognitive Science*, 2006.

[Mayes, 2019] Randolph Mayes. Theories of explanation. *Internet Encyclopedia of Philosophy*, 2019.

[Miller, 2018] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *ArXiv*, 2018.

[Narayanan *et al.*, 2018] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Greshman, and Finale Doshi and. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *ArXiv*, 2018.

[Park *et al.*, 2018] Dong Park, Lisa Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. *CoRRabs*, 2018.

[PWC, 2018] PWC. Explainable ai driving business value through greater understanding. *White Paper, Intelligent Digital*, 2018.

[Ras *et al.*, 2018] Gabrielle Ras, Marcel van Gerven, and Pim Haselager. Explanation methods in deep learning: Users, values, concerns and challenges. *ArXiv*, 2018.

[Rips, 2002] Lance Rips. Circular reasoning. *Cogn. Science*, 2002.

[Rosemary, 1999] Roberts Rosemary. What makes an explanation a good explanation? : adult learners' criteria for acceptance of a good explanation. *Masters Thesis: University of Newfoundland*, 1999.

[Ross *et al.*, 1976] Lee Ross, David Greene, and Pamela House. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Psychology*, 1976.

[Ruben, 1990] David-Hillel Ruben. Explaining explanation. *Routledge Publications*, 1990.

[Salmon, 1989] Salmon. Four decades of scientific explanation. *Minneapolis University Press*, pages 3–219, 1989.

[Sober, 1984] Elliott Sober. Common cause explanation. *Philos Sci.*, 1984.

[Srinivasan *et al.*, 2018] Ramya Srinivasan, Ajay Chander, and Pouya Pezeshkpour. Generating user friendly explanations for loan denials using gans. *NeurIPS Workshop*, 2018.

[Vigano and Magazzeni, 2018] Luca Vigano and Daniele Magazzeni. Explainable security. *IJCAI XAI workshop*, 2018.