Towards a User-Centered, Goal-Driven, and Context-Involved Explaining System

Xinyi Yang
Department of Automation
Tsinghua Universtiy
xy-yang21@mails.tsinghua.edu.cn

Abstract

In this essay, we focus on characteristics of selected and social in human explanations, thus arguing that, a user-centered, context-involved and goal-driven system is what we need to build a good explaining system. Two characteristics of human explanations, selected, *i.e*., select the most relevant as explanations, and social, *i.e*., how to select is determined by the context of the targeted users, are first introduced based on psychological research. Then, three forms of human explanations, deductive proofs, casual patterns, and mental models, are analysed about their advantages and limitations from the view of selected and social explanations. The consequent challenge is that all these theories are usually without context surrounding and not universally applicable. Finally, two ways of designing priors for types of users and applying bidirectional human-robot communication models are proposed to offer possible avenues.

1 Introduction

In recent years, while Artificial Intelligence (AI) has achieved remarkable levels of performance in more and more complex computational tasks [17], the sophistication of these methods has also increased to such an extent that it is difficult for humans to understand and then interfere with the processes they work [2]. Machine Learning (ML) models, always called "black-box", are typical as they create and use decisions that are not justifiable, or in other words, detailed explanations of decision-making processes are not obtained [5]. Obviously, we do not dare to apply such methods into some key areas where far more information than a simple binary prediction is required [2], *e.g*. medicine, finance, security and so on. Thus it is essential to build explainable systems for current and future AI methods, *i.e*., Explainable Artificial Intelligence (XAI).

Since the purpose of XAI is for humans to understand AI, finding out the psychological foundations of explanations in humans is vital for developing effective explaining systems. Studies in the fields of cognitive science and social psychology present us a variety of ideas and also challenges to realize better explaining. This essay focuses on characteristics of selected and social in human explanations, thus arguing that there is no one form of explaining that is universally applicable, a user-centered, context-involved and goal-driven system is what we need. The following content is organized as Fig. 1.

2 What Kind of Explanations do We Need

2.1 Selected: Applicable to the Current Scenario

It is stated in Chander and Srinivasan [4] that an explanation is a filter of facts in a given context. People do not typically provide everything that makes sense in causal chains in an explanation, as it is too large to comprehend [7]; Instead, they select what they believe are the most relevant [11]. Several criteria play important roles in explanation selection: Lipton [9] argues that necessary causes

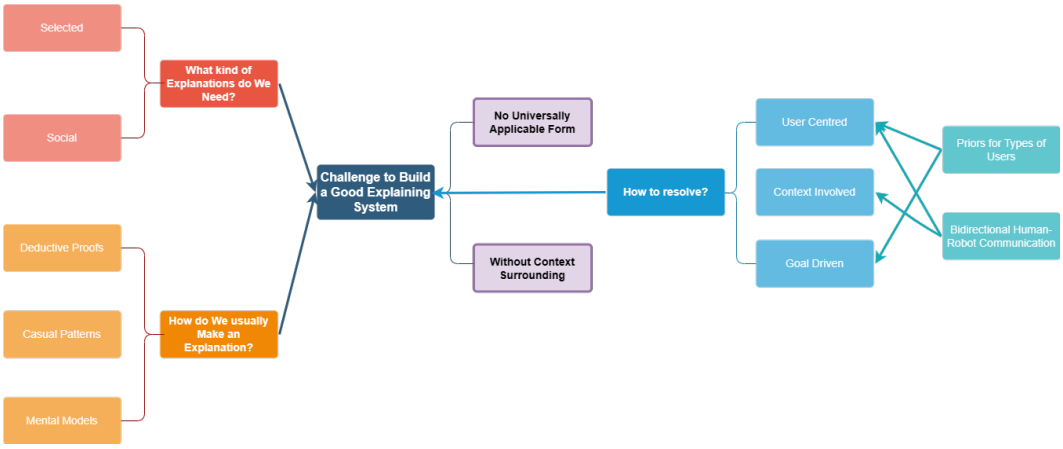

Figure 1: In Sec. 2, two characteristics of human explanations, selected, *i.e.*, select the most relevant as explanations, and social, *i.e.*, how to select is determined by the context of the targeted users. Then, three forms of human explanations are analyzed in Sec. 3, from the view of selected and social explanations, and it can be seen that all these theories are usually without context surrounding and not universally applicable. How to resolve such a consequent challenge? We argue that a user-centered, context-involved and goal-driven system is needed. Two possible ways are proposed in Sec. 4.

are preferred when explaining, for example, when either in sunny or snowy days Tom will go out if Mary invites him, the explanation of Mary's invitation is preferred to the weather as Mary's invitation is necessary for the observed event but neither sunny weather nor snowy weather individually is; Also, Woodward [18] argues that sufficiency is another strong criteria, for example, if it is found that Tom goes out every day he doesn't work, even though Mary doesn't invite him, the explanation of a vacation would be preferred most as it is both sufficient and simple. However, on what basis do we judge the necessity, sufficiency or other criteria of an explanation? This question needs to be answered with another characteristic: social.

2.2 Social: Targeted to the Context of the User With a Purpose

According to Miller [11], explanations are social as they are only presented as a part of conversation or interaction with a certain purpose. The content and format of an explanation must take into account its function for a specific purpose and the context of the targeted user. In the first place, it is intuitive that people don't explain things for no reason, there must be something driving them, *e.g.*, transferring knowledge, gaining trust, or whatever. Evidence has also been presented in an experiment of Slugoski et al. [14]. In this experiment, participants were shown information about a student called George, who had seriously injured another boy in a school fight. The information consisted of two types: (1) a situational profile of George's family background; and (2) the circumstances of the fight. Then these participants were asked to explain why George had assaulted the other boy to another one played by a researcher, who was set to have either of two types of information or neither. Results showed that participants tailored the content of their explanations to suit their expectations of knowledge of the enquier, and tended to select single causes as they believed the enquier was ignorant. The different explanations provided to the enquier indicated that different contrasts they believed they were resolving for their partner.

3 How do We usually Make an Explanation

3.1 Contemporary Theories

Authors in Srinivasan and Chander [16] have provided a succinct organization of contemporary theories of explanations and research concerning the nature of explanations, which will be introduced briefly below (In our opinions, explanations as stances occur sometimes in other forms of explanations, so they will not be included). And based on their analysis of suitable scenarios and limitations of these theories, a discussion form the view of selected and social explanations will also be made.

3.1.1 Deductive Proofs

According to Hempel and Oppenheim [6], explanations are like the deductive sequences of proofs in logic with a set of basic laws. Such form of explanations is primarily suited in scenarios where there are already defined laws that control the phenomena accurately, *e.g*., logic, mathematics, physics, *etc*. [12]. However, it doesn't generalize upon most occasions as the defined laws in one situation always are not applicable in another situation [13]. It also doesn't take into consideration the goal of explaining and the context of the user as the deductive sequences don't change when the context changes.

3.1.2 Causal Patterns

Explanations often refer to causal relations [15, 3, 8]. Casual attributions are widely used in domains such as epidemiology, economics, marketing, environmental sciences, law, policy making, and medicine, where there is a well structured physical model [16]. However, some of them are highly domain specific [1], thus sometimes don't generalize well. Moreover, the structure in a specific domain could vary from person to person as everyone perceives the world differently, so they are based on human context but how the structure exactly looks like is usually hard to grasp.

3.1.3 Mental Models

Mental models are internal representations of beliefs, goals and intentions under a certain situation. According to Mayes [10], explaning is a purely cognitive activity and thus explanation is a kind of mental representation that is involved in an event. Furthermore, when mental models include inferences about the explainee's mental states, they are well context-tailored. But actually, mental models are tightly tailored to the context of the explainer self, whose beliefs are not always the ground truth.

3.2 Consequent Challenges

It can be found that explanations as deductive proofs and mental models don't pay much attention to the context of the specific user; explanations as casual patterns need a more comprehensive context that always doesn't exist to build a personalized structure. There is no explicit mechanism in these forms about how to change according to different social purposes. Furthermore, as stated in Srinivasan and Chander [16], "There is no one theory that is universally applicable". Therefore, the major challenge to be resolved is how to explain with the purpose of explaining and the context surrounding explanations.

4 How to Build a Good Explaining System

In order to bridge the gap between psychological research and practice in XAI about how to build a good explaining system that is user-centered, context-involved and goal-driven, two possible ways will be proposed below.

4.1 Priors for Types of Users

Details needed to be explained are different in models targeted at different tasks and different types of users, *e.g*., natural scientists, data engineers, AI scientists and common users; Forms of explanations that are easy to understand are different for different types of users. Although it cannot meet individual needs, designing prior knowledge about what kinds of explanations different types of users prefer is useful for understanding what content of the explanation a specific type of users is seeking and in what forms they will grasp the meaning more effectively and more comfortably.

4.2 Bidirectional Human-Robot Communications

Understanding the explanation context of current moments means value alignment between the explainer and the explainee. Humans always use communication as an efficient tool to establish a common understanding of context in the process of interaction. Successful communication can serve as effective explanations as there exists a bidirectional value alignment. The XAI system proposed

in Yuan et al. [19] integrates a cooperative communication model for inferring human values from human feedback, which indicates that real-time human-robot mutual understanding is achievable based on bidirectional communication.

References

[1] Woo-kyoung Ahn. Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition*, 69(2):135–178, 1998. 3

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 1

[3] Richard Boyd et al. Homeostasis, species, and higher taxa. *Species: New interdisciplinary essays*, 141:185, 1999. 3

[4] Ajay Chander and Ramya Srinivasan. Evaluating explanations by cognitive value. In *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*, pages 314–328. Springer, 2018. 1

[5] David Gunning. Explainable artificial intelligence (xai): technical report defense advanced research projects agency darpa-baa-16-53. *DARPA, Arlington, USA*, 2016. 1

[6] Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175, 1948. 3

[7] Denis Hilton. Social attribution and explanation. 2017. 1

[8] Frank C Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006. 3

[9] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990. 1

[10] G Randolph Mayes. Theories of explanation. 2001. 3

[11] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019. 1, 2

[12] Rosemary Roberts. *What makes an explanation a good explanation?: adult learners' criteria for acceptance of a good explanation*. PhD thesis, Memorial University of Newfoundland, 1999. 3

[13] Wesley C Salmon. *Four decades of scientific explanation*. University of Pittsburgh press, 2006. 3

[14] Ben R Slugoski, Mansur Lalljee, Roger Lamb, and Gerald P Ginsburg. Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology*, 23(3):219–238, 1993. 2

[15] Elliott Sober. Common cause explanation. *Philosophy of Science*, 51(2):212–241, 1984. 3

[16] Ramya Srinivasan and Ajay Chander. Explanation perspectives from the cognitive sciencesa survey. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4812–4818, 2021. 2, 3

[17] Darrell M West. *The future of work: Robots, AI, and automation*. Brookings Institution Press, 2018. 1

[18] James Woodward. Sensitive and insensitive causation. *The Philosophical Review*, 115(1):1–50, 2006. 2

[19] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. In situ bidirectional human-robot value alignment. *Science robotics*, 7(68):eabm4183, 2022. 4