Applications of Social Attribution Theory in XAI †

Guihong Zhang

School of Humanities and Social Sciences, University of Science and Technology of China, Hefei 230026, China; guihong@ustc.edu.cn
† Presented at the 5th International Conference of Philosophy of Information, IS4SI Summit 2021, Online, 12–19 September 2021.

Abstract: A major problem facing artificial intelligence ethics is how to give AI the status of the agent and gain the trust of human beings. This is accompanied by the rise of Explanatory AI (XAI). The social attribution theory (SAT) started from the perception experiment of Hyde and others, then psychologists conducted a long-term analysis of the causes and reasons of human intentional action, and Mahler built a complete social psychological attribution model from this. Because SAT is based on action analysis, it can reflect its vast application potential in XAI, including the interpretation of artificial agent action, interpretation of beliefs and other folk psychology concepts, analysis of group intelligence, and the construction of normative agent action and the corresponding ethical explanation. These explorations can provide new ideas for XAI.

Keywords: social attribution theory; intention; reason; action; explanation; XAI

1. Introduction

With the development of AI, the interpretation of algorithms is becoming a research hotspot, making XAI gradually become the focus to some scholars. In fact, observing how humans interpret each other can be a helpful starting point for artificial intelligence interpretation. Much research has been done on the interpretation of human action by philosophers, psychologists, and social scientists; they have studied the process of interpretation in detail by focusing on cognitive biases and social expectations. For decades, SAT has been analyzing how people attribute and evaluate the social action of others in the physical environment. There is much room for injecting this significant research result into XAI.

SAT is about perception. Although the cause of action can be described at a neuro-physical level or even lower, the SAT does not care about the actual cause of human action, but how other people attribute or explain the action of others. Many works by Malle and others show that intention and intentionality are the keys to his work [1]. Intention can provide a psychological guarantee for human beings to carry out their specific action, so it is a psychological state.

The SAT uses the “ordinary” term for attribution of human action. Although these concepts may not cause human action, the purpose of these concepts is to model and predict how humans do act with each other, so the SAT explains what people understand and think about action, rather than focusing on how people think. In its model, the action consists of three parts: (1) the premise of the action, including the conditions for the successful execution of the action, such as the actor’s ability or environmental constraints; (2) the action itself that humans can take; (3) the effect of the action, that is, the environmental or social changes it brings. The actions taken are usually explained by plans or intentions. In most work in the social sciences, goals equate to intentions. The goal is defined as the goal that the mean helps to reach, and the intention is defined as the short-term goal adopted to achieve the final destination. Beyond achieving positive utility goals, these intentions have no utility. Malle and Pearce divided people’s interpretation of action into two dimensions:
(a) observable and unobservable action; (b) intentional and unintended action; combining actions according to this division can produce four different types of action characteristics: observable intentional action, unobservable intentional action, observable unintended action, unobservable unintended action [2]. Since observable and intentional actions are easy to understand, and unintentional actions do not need to be explained, for interpreters, intentional unobservable actions are the key aspects of explanation.

2. SAT’s Interpretation Model

In addition to intentions, SAT research shows that other factors are also important for the attribution of action, especially beliefs, desires, and traits. Malle has done a lot of pioneering work in this field. Malle proposed a model based on psychological theory that people attribute their actions to others and themselves by assigning specific mental states that explain the action [1]. He believes that the following represent the assumptions and distinctions people make when they attribute their action to themselves and others.

(1) People distinguish between intentional and unintentional action. (2) For unintentional action, people provide legitimate reasons, such as physical, mechanical, or circumstantial. (3) For intentional action, people use three interpretation methods according to the specific circumstances of the action: (a) Reason explanations are explanations related to the mental state of the action, and the reasons for their intentions. (b) The causal history reason (CHR) interpretations are those that use factors “in the context of agency reasons”. These factors may include unconscious motives, emotions, culture, personality, and background. CHR interpretation refers to the causal factors that lead to causes, etc. (c) Enabling factors (EF) explain factors that do not explain the intention of the actor, but explain how the intentional action achieves its result.

The core of the Malle model is the intentionality of action. For action that is considered intentional, the action must be based on certain desire, and the belief that the action can be performed and the desire can be achieved. This forms the intention. If the agent is capable and aware that they are performing an action, then the action is intentional. The explanation given is to attribute intentionality to action, and to identify desires, beliefs, and values based on the assumption of subjectivity and actions based on it. Therefore, reason means intentionality, subjectivity, and rationality.

3. Field of Application

As de Graaf and Malle pointed out in their article that the folk psychology conceptual framework can well explain human actions in different situations [3], and such models are very useful in XAI, folk psychology models are very needed in XAI. At the same time, the analysis of intention by the BDI model can also be well applied in XAI, which is helpful for the research of XAI. Therefore, work on the relationship between premises, results, and competitive goals is helpful in the following aspects.

(a) Cognitive process and evaluation. The general view that people use covariance is valid. The three cognitive processes used for explanation include: (1) Choice of explanation, (2) causal connection, and (3) evaluation of explanation. Choice of explanation is the way people choose some specific reasons to explain in the explanation of action; causal connection is the process of explaining the cause of action; and evaluation of explanation is a process used to assess the quality of explanations of action. Due to various cognitive biases among different interpreters and evaluators, such biases can have a certain impact on the generation, selection, and evaluation of explanations.

(b) Explanation of action. Malle’s model is by far the most mature SAT model. Malle’s conceptual framework provides a suitable framework to characterize different aspects of action causes. Obviously, reason explanations are useful for goal-based reasoners. The fact that the agent optimizes costs is the agent’s “personality,” which is constantly given a specific plan or goal.

(c) Collective intelligence. Research on the attribution of group action is important for people engaged in collective intelligence work, including areas such as multi-agent
planning, computational social choice, or argumentation. Although compared with the attribution of individual action, this area of work seems to be seldom explored; however, O’Laughlin and Malle found that people assign intentions and beliefs to groups that act together, and research on aggregated groups shows that a lot of work attributing individual actions can be used as a solid foundation for explaining collective actions [4].

4. Ethics

The application of social attribution theory can help us further think about ethical issues in AI, including the following.

(a) Norms. Norms have been proven to occupy a special place in social attribution. Uttich and Lombrozo studied the relationship between norms and their influence on attribution to specific mental states, especially in terms of morality [5]. They provide a reasonable explanation for the side effects of the Nobel effect, which is the effect of people attributing certain mental states based on moral judgments. Samland and Waldmann further studied social attribution in the context of norms, focusing on permission rather than obligation [6]. They provided participants with scenarios where two results would occur.

(b) Ethics. For humanoid agents, morality is very important. First, the connection with morality is important for applications that cause ethical or social issues. General explanations or actions that violate the norms may give people the impression of an “immoral machine”. Therefore, such norms need to be clearly regarded as part of the interpretation and interpretability. People mostly think an explanation of what they think is abnormal or abnormal, and violation of norms is such an abnormality [7].

(c) Responsibility. Responsibility and blame are related, and they are both associated with the causality of action, which can provide an explanation for the cause of the action and can identify those who are responsible for it. Responsibility is an inevitable element in causal explanations because there is a necessary connection between cause and responsibility. Chockler and Halpern used the structural equation model to define the responsibility of the result, which is a formal model that is easier to adopt in artificial intelligence [8].

5. Conclusions

After the previous analysis, it can be found that SAT has broad application in XAI. The SAT uses ordinary terms for attribution of human action. Although these concepts do not directly cause some specific actions, they can play an explanatory role in the interpretation of actions, so that actions can be better predicted and analyzed. Malle proposed an interpretation model from this, which asserts that people attribute their action to others and themselves by assigning specific mental states to explain action [9]. This model can rationally explain cognitive processes and evaluations, action explanations, and collective intelligence. It can also further help us think about ethical issues in AI, including issues such as norms, ethics, and responsibility.

Funding: This research is supported by the Fundamental Research Funds for the Central Universities: WK2110000013.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Malle, B.F. How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction; MIT Press: Boston, MA, USA, 2004.
2. Malle, B.F.; Pearce, G.E. Attention to behavioral events during interaction: Two actor-observer gaps and three attempts to close them. J. Pers. Soc. Psychol. 2001, 81, 278–294. [CrossRef] [PubMed]
3. De Graaf, M.M.; Malle, B.F. How people explain action (and autonomous intelligent systems should too). In Proceedings of the AAAI Fall Symposium on Artificial Intelligence for Human–Robot Interaction, Arlington, VA, USA, 9–11 November 2017.
4. O’Laughlin, M.J.; Malle, B.F. How people explain actions performed by groups and individuals, J. Pers. Soc. Psychol. 2002, 82, 33. [CrossRef]
5. Uttich, K.; Lombrozo, T. Norms inform mental state ascriptions: A rational explanation for the side-effect effect. Cognition 2010, 116, 87–100. [CrossRef] [PubMed]
6. Samland, J.; Waldmann, M.R. Do social norms influence causal inferences? In Proceedings of the 36th Annual Conference of the Cognitive Science Society, Cognitive Science Society, Quebec City, QC, Canada, 23–26 July 2014; Psychology Press: London, UK, 2014; pp. 1359–1364.

7. Hilton, D.J. Mental models and causal explanation: Judgments of probable cause and explanatory relevance. *Think Reason.* 1996, 2, 273–308. [CrossRef]

8. Chockler, H.; Halpern, J.Y. Responsibility and blame: A structural-model approach. *J. Artif. Intell. Res.* 2004, 22, 93–115. [CrossRef]

9. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 2019, 26, 1–38. [CrossRef]