Engineering 6 (2020) 218–220

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Views & Comments

How to Interpret Machine Knowledge

Fashen Li [a,#], Lian Li [b,#], Jianping Yin [c,#], Yong Zhang [d,#], Qingguo Zhou [e,#], Kun Kuang [f,#]

[a] Department of Physics, Lanzhou University, Lanzhou 430000, China
[b] Department of Computer Science, Hefei University of Technology, Hefei 230009, China
[c] Department of Computer Science, Dongguan University of Technology, Dongguan 523808, China
[d] Department of Physics, Xiamen University, Xiamen 361005, China
[e] Department of Computer Science, Lanzhou University, Lanzhou 430000, China
[f] College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China

Machine knowledge refers to the knowledge contained in artificial intelligence. This article discusses how to acquire machine knowledge, with a particular focus on the acquisition of causal knowledge. The latter is the process of interpreting machine knowledge. Through the analysis of certain research methods in the fields of physics and artificial intelligence, we propose principles and models for interpreting machine knowledge, and discuss specific methods including the automation of the interpretation process and local linearization.

Human beings have now entered the four-dimensional society that comprises the natural world, human world, information world, and intelligent-agent world. The intelligent agent[†] has become an objective existence of our world. An intelligent agent can make predictions, make judgments, express emotions, and even actively adjust its behaviors to adapt to changes in the environment [1,2]. Hence, we can think of an intelligent agent as a knowledge system with a knowledge structure and function, known as machine knowledge.

To establish a generally accepted definition of knowledge, it is still necessary to continuously study it in depth. In this article, we first set forth the general definition that knowledge is the law of phenomena change. An intelligent agent can change the output from the input, or adjust the next output based on the previous output. This kind of input and output—as well as the law of change between output and output—is the law of change of the phenomenon, so it belongs to knowledge. This kind of knowledge is called *primary knowledge*. For example, placing all the changes in the phenomena into a table is an expression of knowledge (i.e., exhaustive expression). However, the knowledge that people need is often not this primary form of knowledge, but rather one that is abstracted at a higher level—that is, the general and universal law that reflects the change of phenomena. This kind of knowledge is called *advanced knowledge*. Advanced knowledge can continue to be layered according to the degree of abstraction. Taking the work of Tycho Brahe and Johannes Kepler as an example, through detailed observations, Tycho listed a large amount of trajectory data of planetary operations, which only reflected the associations of phenomena (i.e., planetary operations). Once Kepler successfully summed up the three laws and revealed the causal relationship of those phenomena, high-level knowledge of planetary operations was developed. Moreover, Newton's second law is a yet higher-level expression of knowledge. Both association and causal relationships are knowledge, but they are at different levels. In the process of humans acquiring knowledge, it is the most basic scientific activity to determine the association between phenomena through observation. To determine causality, it is necessary to analyze and summarize the phenomena behind the observed data. Causality plays an important role in the human science system, since humans always want to know—and persistently pursue—the "why" behind a phenomenon change.

In this paper, we focus on the question of whether people can obtain causal knowledge from intelligent agents, and how it may be done. This process involves the interpretation of machine knowledge. Through training, intelligent agents can complete very complicated work, and some of their achievements have exceeded humanity's cultural accumulation over thousands of years. However, we still do not know how these agents are so successful. For example, for an intelligent agent such as neural network, excessive fitting training data does not make neural network more generalizable. We do not know where the boundaries of its success are. We do not know how to design the structure of a neural network to accomplish an intended task. We do not know whether it is possible to change the training set to make the neural network perform better. We do not even know what the neural network is based on for precise prediction—that is, whether it is based on data or on features. In a word, we do not understand the knowledge of an intelligent agent; hence, how can we trust it?

Thus far, causality remains the fundamental cornerstone of human understanding of the natural world, and the association described by probabilistic thinking is the surface phenomenon that drives us to understand causal mechanisms in the world. As Pearl [3] said,

[#] These authors contributed equally to this work.

[†] The intelligent agent in this paper refers to artificial intelligent machines based on silicon technology and Turing algorithm, such as various learning models, computing models, and simulation models, excluding agents constructed using biological or genetic technologies.

https://doi.org/10.1016/j.eng.2019.11.013

In retrospect, my greatest challenge was to break away from probabilistic thinking and accept, first, that people are not probability thinkers but cause-effect thinkers and, second, that causal thinking cannot be captured in the language of probability; it requires a formal language of its own.

The first point is the fact that scientific knowledge is not expressed in the form of probabilistic thinking, but is expressed as causal thinking. The second point involves how to carry out causal thinking. Pearl believes that humans have not yet invented mathematical tools that portray causal thinking. Unfortunately, most currently favored agents are run in a probabilistic manner, and the relationships between the expressed phenomena are all associations. Can we interpret the causation contained in these associations? It is still a very challenging problem. If humans and agents cannot communicate and understand each other, or if humans cannot translate the knowledge of agents into a causal form, then the development of artificial intelligence will encounter great obstacles and may even bring hide danger [4].

Physics is a typical science that interprets the natural world with causality. The natural world can also be a huge intelligent agent, with phenomena changing every moment. To recognize the changes in the natural world and their laws, humans always adopt a description form of causality. They hope to give clear and accurate expressions of the laws behind the phenomenon transformation. This is mainly done by adopting regular expressions and mathematical expressions, which not only make it possible for humans to describe what has happened, but also make it possible for them to predict what is likely to happen, where the latter is especially important. However, the actual operating laws of the natural world cannot be directly obtained: Humans can only "guess" the laws that are inherent in natural phenomena through observations. It is very difficult to accurately and completely summarize the corresponding law even with a large amount of data on the phenomena. Therefore, humans use two principles (or beliefs) to interpret the natural world, which are clearly stated in Newton's *Mathematical Principles of Natural Philosophy, Volume III: On the System of the Universe* [5]. These are the first two of the four "rules of reasoning in philosophy":

(1) The simplest description principle (i.e., Occam's razor): We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.

(2) The functional similarity principle: Therefore, to the same natural effects we must, as far as possible, assign the same causes.

For physics, some fundamental laws and principles not only are a high degree of abstraction and causal characterization of the laws of phenomena of the natural world, but also follow the two basic principles mentioned above, thus forming the current basics and cognitions for the natural world and building the structure of human natural science knowledge. For example, measurements cannot be used to accurately verify Newton's second law, so why do we still accept it? We accept it because there is a well-established principle hidden inside.

Let us return to the interpretation of the intelligent agent. In most cases, we can know the structure of the agent, but we cannot predict its behaviors, just as we cannot judge what our brain will do based on its neural connection structure. We can only observe the associations between its input and output—that is, the data. For any agent, if there are sufficient observations and a large amount of observational data, it is theoretically possible to obtain the causal relationship through inductive calculation, without considering the internal structure and operation mode of the agent. It is said that the causal relationship can be established if it is highly consistent with the external performance (i.e., function) of the agent. This is guaranteed by the functional similarity principle. This method is fully embodied in physics. For example, the universe can be said to be like a huge clock, where we can only guess its internal structure from the outside. With continuous observations that improve in accuracy, our guess will become increasingly consistent with the observed phenomenon. Yet we may never know the actual structure inside the cosmic "clock." Despite this incomplete knowledge, physics promotes human social development and scientific progress.

Humans have been exploring causal relationships for thousands of years, but the description of causality has remained at a qualitative and empirical stage for a long time—until the 1970s, when C. Granger, J. Pearl, and D. Rubin proposed a definition of causality based on mathematical expressions. At that point, humans began to establish quantitative research on causality. Pearl's description and method of causality are systematic and algorithmic, and can therefore deal with confounding interference among variables, find the existence of implicit variables, and solve the problems of attribution such as counterfactuals. The research based on Pearl's causality achieved excellent performance in many real applications and can be applied to address the causal paradox problem. Therefore, Pearl's causality has become an important method in the theory and application of artificial intelligence. In principle, Pearl's causality has the same scientific assumptions and mathematical foundations as Fisher's experimental design; hence, Pearl's causality has a solid mathematical foundation.

However, Pearl's causality still has certain issues that make it less than satisfactory for slightly more complicated problems. For example, Pearl's causal algorithm requires a high degree of data distribution and quantity—requirements that cannot be met in many real applications. Furthermore, Pearl's causality is very sensitive to hidden variables; hence, insufficient or inaccurate observational data will greatly affect the calculation results. There are still many uncertainties in constructing the causal structural equation model or causal structure diagram model required by Pearl's causality and its algorithms.

Subsequently, Imbens and Rubin [6] proposed another causal model, named the potential outcome model, to explore the underlying causal knowledge by studying the potential outcomes and phenomena associations reflected in the data. Rubin's causal model has been widely used in practical problems, especially those that require causal knowledge to assist in decision-making, such as medical diagnosis and public policymaking. However, Rubin's causal model also has some problems; for example, its assumptions on data are too strong, and some of those assumptions are not testable in practical problems.

Although the causal methods of Pearl and Rubin are still being studied, other methods have also been developed. Even though the causal relationship cannot be directly calculated, it can still reveal profound relationships from the knowledge of the intelligent agent. Those methods come from research in physics and artificial intelligence. Physicists also apply machine learning methods in their research when it is difficult to draw a causal relationship to understand the natural world. For example, machine learning methods have been used to understand the Langevin equation for multibody systems and the Boltzmann description of Liouville's equation (the Bogoliubov–Born–Green–Kirkwood–Yvon (BBKGY) truncation). Interpretation algorithms are also used in artificial intelligence to understand the intrinsic relationships among complex data or features. Using an intelligent agent to interpret an intelligent agent is a wonderful idea. In fact, the current variety of intelligent agents (or learning models) are hierarchical in transparency. That is, some agents are more transparent to humans, such as linear models and decision tree models, while other agents are more obscure to humans, such as neural networks and Monte Carlo search tree models. It is regrettable (but very interesting) that the more obscure an agent is, the stronger its learning ability is, and the more knowledge it contains. If it is difficult to interpret an agent directly, one can consider interpreting it through a more transparent agent. This process can be recursive, making the content of the interpretation more and more easily understood by humans [7].

By calculating the influence function, the importance of the data or features in an intelligent agent can be analyzed, making it possible to analyze which factors (i.e., causes) cause the agent to have such a performance. It is also possible to analyze the quality and distribution of the data to find better observational data, which is very meaningful in both medical diagnosis and physical observation.

For a given input data, the intelligent agent will give the corresponding output (or the next action). By calculating the Shapley value of each input data feature, it is possible to estimate the contribution of different features to the output. Features with large contributions are likely to be causes for the behavior of the agent [8].

For complex agents, according to the universal mathematical principle, the local behavior of the agent should be similar to a linear system. Hence, according to functional similarity, it is possible to consider replacing the original agent with a linear model (e.g., linear regression) in a local range [9]. The linear model has good transparency for causality, and its causal relationship can be obtained by the appropriate processing of its regression coefficients. Simultaneously, through the analysis of residuals, the accuracy of this approximation can be determined, as well as the sensitivity of other factors to the main variables.

Another straightforward approach is to use a more transparent model T to learn the obscure model V, in order to obtain the data labeled by $(x, V(x))$ by inputting the data x, where $V(x)$ represents the output of V with respect to x. Then T is relearned based on those data. If T and V have basically the same behaviors, then, according to the functional similarity principle, T and V can be considered to have the same causal knowledge. This method has achieved good results in analyzing the internal defects of an agent and in black-box attacks.

The emergence of artificial intelligence has opened more ways for humans to discover new knowledge. By interpreting the knowledge of intelligent agents, we can enrich our own knowledge systems and better serve human development. At present, the interpretation of the intelligent agent still requires further study. As the theory and methods continue to improve, humans and agents will achieve a higher level of harmony in their relationship and will achieve better communication and cooperation with each other. This will be a milestone in the history of human evolution.

Acknowledgements

This article is based on the talks of participants at the Salon of Machine Knowledge and Human Cognition held by Lanzhou University in July 2019. All participants contributed to this article, we are greateful to other participants. They are experts in physics and computer science, including Liang Huang (Physics, Lanzhou University), Ning An (Computer Science, Hefei University of Technology), Lei Yang (Physics, Institute of Modern Physics, Chinese Academy of Sciences), Zhixi Wu (Physics, Lanzhou University), Li Liu (Computer Science, Chongqing University), Jialin Zhang (Computer Science, Institute of Computing Technology, Chinese Academy of Sciences), and Lianchun Yu (Physics, Lanzhou University).

References

[1] Pan Y. Special issue on artificial intelligence 2.0. Front Inf Technol Electron Eng 2017;18(1):1–2.

[2] Pan Y. 2018 special issue on artificial intelligence 2.0: theories and applications. Front Inf Technol Electron Eng 2018;19(1):1–2.

[3] Judea Pearl on his inspiration and the breakthrough moments of his research [Internet]. Cambridge: Cambridge University Press; 2012 [cited 2020 Jan 03]. Available from: http://www.cambridgeblog.org/2012/07/qa-with-judea-pearl-part-one/.

[4] Pearl J, Mackenzie D. The book of why: the new science of cause and effect. New York: Basic Books; 2018.

[5] Newton I. Mathematical principles of natural philosophy. 2nd ed. London: A. Strahan; 1802.

[6] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. New York: Cambridge University Press; 2015.

[7] Lucci S, Kopec D. Artificial intelligence in the 21st century. Sterling: Stylus Publishing, LLC; 2015.

[8] Molnar C. Interpretable machine learning: a guide for making black box models explainable [Internet]. 2019. Available from: https://christophm.github.io/interpretable-ml-book/.

[9] Ribeiro MT, Singh S, Guestrin C. "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; 2016 Aug 13–17; San Francisco, CA, USA. New York: Association for Computing Machinery; 2016. p. 1135–44.