FILOSOFIJA. SOCIOLOGIJA. 2022. T. 33. Nr. 1, p. 40–48

The Epistemology of Machine Learning

HUIREN BAI
Department of Philosophy, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, P. R. China
Email: bhuiren@gmail.com

This paper argues that machine learning is a knowledge-producing enterprise, since we are increasingly relying on artificial intelligence. But the knowledge discovered by machine is completely beyond human experience and human reason, becoming almost incomprehensible to humans. I argue that standard calls for interpretability that focus on the epistemic inscrutability of black-box machine learning may be misplaced. The problems of transparency and interpretability of machine learning stem from how we perceive the possibility of 'machine knowledge'. In other words, the justification for machine knowledge does not need to include transparency and interpretability. Therefore, I am going to examine some sort of machine learning epistemology and provide three possible justifications for machine knowledge, which are formal justification, model justification and practical justification.

Keywords: machine learning, epistemological foundation, transparency, interpretability, machine knowledge

INTRODUCTION

At present, the researches on the philosophy of artificial intelligence (AI) among the philosophical community mainly focus on the ethical issues and moral dilemmas, and few literatures relate to the epistemology of AI. Therefore, this paper will discuss the epistemology of artificial intelligence, especially the epistemology of machine learning. The change brought by machine learning is huge, and for human beings, machine learning is undoubtedly the best means for people to reacquaint themselves with the world. Machine learning algorithms establish certain behavioural models and use the models to analyse massive amounts of data to complete predictions for the future. The development of machine learning will help us understand the world better and more accurately.

Through machine learning, AlphaGo Zero explored a large number of walks that no human had ever attempted in a short period of time without previous human experience or guidance and without providing knowledge of any domain other than the basic rules (Li et al. 2020). The knowledge discovered by the machine was not only completely beyond human experience but also beyond human reason, becoming almost incomprehensible to humans. Both scientists and philosophers have addressed various facets of this black box dilemma of machine learning, which has also been labelled a problem with opacity, understanding, transparency and interpretability (Burrell 2016; Walmsley 2020; De Laat 2018; Carvalho et al. 2019; Krishnan 2019). So the idea of explainable artificial intelligence has been developed by

some scholars (Linardatos et al. 2021; Arrieta et al. 2020). This gives rise to the possibility of discussing 'machine epistemology' (Wheeler 2016). Machine learning based on different algorithms expresses and advocates different epistemic views, which not only argue for or support a certain epistemology, but also embody a reverse constraint: the kind of epistemology or epistemic view held determines what kind of paradigm machine learning is designed; thus, for machine learning to gain new breakthroughs, it also awaits the integration and breakthrough of epistemology.

I argue that machine learning is a knowledge-generating enterprise, since we are increasingly relying on machines. But it is widely acknowledged that the precise mechanisms by which machine learning generates predictions are quite mysterious. The output of machine learning is often beyond human understanding, they 'think' about the world differently than we do. People might think that machine learning is epistemologically inscrutable. Thus, if we presume that machine learning can be a source of knowledge, then we must deal with the transparency and interpretability problem mentioned above, and give appropriate justification for 'machine knowledge'. Therefore, in the first part, I will give a brief explanation of the epistemological foundations of machine learning; then, in the second part, I will build on my description of the transparency and interpretability of machine learning by suggesting that these two factors are obstacles to the acceptance of 'machine knowledge'; finally, the third part will show my three possible justifications for machine knowledge, which are formal justification, model justification and practical justification.

EPISTEMOLOGICAL FOUNDATION OF MACHINE LEARNING

Machine learning is a kind of technology seeking to simulate human cognitive ability (intelligence) on machines. Thus, it is closely related to philosophical epistemology, which also studies epistemic activity. Machine learning was originally born from an epistemological analogy as follows. When the rationalist epistemology reveals that the human epistemic process should be governed by strict logic and rules, the information processing in computers is controlled by precise algorithms and programs, and there is undoubtedly a high degree of analogy between the two. Moreover, the two processes are consistent in form and even in essence. Thus, it is possible to simulate the human epistemic process in machines. Besides, the main purpose of epistemology is to regulate those properties and relations necessary for the combination of knowledge. Machine learning is the incorporation of some epistemic relations (especially inference relations) into its own theory of knowledge, enabling the process of knowledge representation or knowledge discovery by machines to highlight the universal function of epistemology.

Machine learning has evoked a new way to understand the world. Two research paradigms have emerged in machine learning: one is inherited from the reductionist and rationalist traditions of philosophy; the other is based on idealised and holistic connectionism. The two research paradigms have now merged into one for data mining and knowledge discovery. The functionality of deep learning networks is established by training and cannot be explained and justified by reference to a predefined rule-based procedure (Schubbach 2021). Machine learning built on artificial neural networks is associated with the epistemology of empiricism, in which learning is performed by the brain. Learning is defined as the process of generalising general principles from continuously accumulated experience, and machine learning is to simulate this learning function of the brain. When a human provides empirical data to a machine system, models based on these continuously accumulated data are generated and then

used to perform recognition (such as image and sound recognition, collectively called pattern recognition). This is understood to parallel the learning performed by a human. The formula can be likened to the cognitive progression from experience to theory, which is the process of extracting knowledge and models from data, as well as the process of generalisation from individual to general. Then, the general models formed will provide corresponding judgments under new situations.

Machine 'experience' can be the basis for constructing an epistemology of machine learning, suggesting that the idea of a machine-oriented, non-anthropocentric epistemology is feasible. Machine 'experience', as an experience characterised by data, is characterised by big data, including a large amount, multidimensionality, rapidity, and overall completeness and automation. Therefore, the computation of data by machines can be defined as the reflection and representation of things by machines based on an algorithmic extension of the human perceptual system. It is similar to human experience but different from human experience. Machines emulate aspects of human experience insofar as they can calculate, compute, register, record, and correlate data that is contained within human experiencing, indicating that human experience data can be directly used by machines. Similarly, humans can learn machine 'experience', reflecting that the use of data in traditional science can be regarded as the learning of machine 'experience'.

Nevertheless, it is commonly recognised that the exact mechanisms through which experience is generated by machine learning are still a mystery to human, which is due to machine learning integrating myriad data with un-concrete mathematical objects which cannot be interpreted by humans. Due to our inability to understand how they work, the models are often labelled 'black boxes'. Therefore, the next section is dedicated to showing machine learning from the perspective of epistemic opacity.

TRANSPARENCY AND INTERPRETABILITY OF MACHINE LEARNING

With the far-reaching impact of machine learning in supporting various fields of human life, cutting-edge questions about the ethical guidelines, black boxes, and attribution of responsibility for machine learning urgently need answers from the philosophical community. Can humans control the pace of machine learning development? How can the models be made more credible in the process of machine learning system-assisted decision-making? Besides, how can the stability and reliability of machine learning output be ensured? All of the above questions are related to the transparency and interpretability of machine learning as well as the clarity and consistency of our epistemological notions. Transparency means that a model is considered to be transparent if by itself it is understandable. Interpretability is defined as the ability to explain or to provide the meaning in understandable terms to a human.

As black-box Machine Learning models are increasingly being employed to make important predictions in critical contexts, the demand for transparency is increasing from the various stakeholders in AI. The danger is on creating and using decisions that are not justifiable, legitimate, or that simply do not allow obtaining detailed explanations of their behaviour (Arrieta et al. 2020). The transparency problem of machine learning is reflected in epistemic opacity. Specifically, 1) at the algorithm level, complex algorithms are composed of multiple implicit layers of artificial neural networks, like a black box, resulting in the problem of uninterpretability of algorithms; 2) at the data level, opacity is closely related to data bias, and bias will penetrate into the data with the human's own access in the human collection and selection of data, resulting in the data used for training; 3) at the level of an intelligent agent,

intelligence is composed of the algorithm, data, and hardware, and its internal working mechanism is quite complex. Moreover, humans may not be able to transparently know the intelligent agent as a holistic being, though it is assumed that humans can achieve transparent knowledge of algorithms and data.

As machine learning relies more and more on deep neural networks, the problem of epistemic opacity is further exacerbated by the facts as follows. 1. The transformation between layers of multilayer neural networks involves hundreds of millions of parameters and complex nonlinearities, making it impossible for us to understand this formal transformation process transparently. As a result, the difficulty of human understanding is significantly increased. 2. The neural networks are in an automatic construction process throughout the whole process of co-action, and modelers can understand only the static structure of the model instead of thoroughly grasping the automated dynamic working of the model.

The interpretability problem of machine learning is mainly in the application domain, especially in scientific research (George, Hautier 2021; Katuwal, Chen 2016). In science, explaining causality retrospectively is the keynote of scientific research when machine learning is applied to specific scientific problems. This process requires experts in the field to understand the models and how machine learning systems can make predictions about unknown samples. For molecular biologists using machine learning to analyse protein structures, it is not enough if machine learning can only perform 'input–system–output' of data. Scientists need machine learning systems that can be interpreted in a molecular biology sense to present a more specific scientific analysis process and correct the inadequacy of training data with scientific experience. Therefore, the interpretability of machine learning systems is even more imperative for more fundamental scientific research problems, in which the cause-and-effect relationships should be understood. Some researchers argue that if it is difficult to interpret a machine learning process directly, one can consider interpreting it through a more transparent agent (Li et al. 2020).

The transparency and interpretability problem of machine learning are bringing into question the assumptions embedded in long tradition – knowledge was about finding the order hidden in the chaos and simplifying the world. The epistemic opacity and uninterpretability seem to prove us that we were wrong. Knowing the world may require giving up on understanding it. Both transparency and interpretability are strongly tied to understandability. Pearl (2018) argues that if humans and machines cannot communicate and understand each other, or if humans cannot translate the machine knowledge into a causal form, then the interpretability problem is really a problem. But I think that the problems of transparency and interpretability of machine learning stem from how we perceive the possibility of 'machine knowledge'.

JUSTIFICATIONS FOR MACHINE KNOWLEDGE

Epistemology is the systematic philosophical examination of knowledge and is concerned with the nature of knowledge and how we acquire it. Thus it might appear that 'machine knowledge' does not have sufficient justification to count as knowledge. But I argue that standard calls for interpretability that focus on the epistemic inscrutability of black-box machine learning may be misplaced. If we presume, for the sake of this paper, that machine learning can be a source of knowledge, then it makes sense to wonder what kind of justification it involves. Then I will give three possible justifications for machine knowledge, which are formal justification, model justification and practical justification.

FORMAL JUSTIFICATION

If knowledge requires the ability to give explanation to our held criterion – Plato's conception, with a path of more than two millenniums, which is the knowledge pattern difficult or even impossible to master? Amongst many philosophers, there is a consensus reached that it shall be a justified, true belief at least counting for a mental condition being a knowledge state. The knowledge which has been specifically expounded is used based on the perspective of epistemology. That is to say, knowledge is justified, true belief. It appears self-contradictory to embrace machine learning if machine learning does not count or is regarded as a source of knowledge. As mentioned above, it shall be a justified, true belief at least, which is the prerequisite to regard a mental state as knowledge. These requirements of justification, truth, and belief are adequate and crucial for viewing a mental state as knowledge. As suggested by the Gettier Problem, there is something extra required. However, since the focus of this study is justification, the discussion about Gettier problems is put aside.

In various fields, machine learning produces remarkable outcomes, which are frequently reliable if applied. It is argued that machine learning is implicitly reflective of a generic reliabilist stance on the outcomes derived from machine learning algorithms. Reliabilism provides a means for justifying knowledge. Provided that it arises out of through a more trustworthy procedure or an approach, a belief can be warranted. Based on reliabilism, it is sufficient that a belief results from a reliable process for justification. That is to say, it is not required that how reliable the process or method is can be indicated by the knower. Despite that the exact causes of the reliability are of less or no transparency, our awareness still lies in those things stemmed from a securely trustworthy procedure.

Particularly for the fields like medicine, the absence of information on what causes models to make prediction is unsettling. As human being, we explore causes and seek explanations for understanding the way and the reason that our situations are still in existence as it is or working on with some particular methods. Nevertheless, those reliabilist ways of justifying it aims at making justification without interpretation. Owing to the truth-revealing properties featured by the machine learning models, our ideas focus on being justified through the acceptance of the outputs of a model. As for the reliabilist approach, it is needless to comprehend the black box model for believing the output outcome, with the output as knowledge.

Undeniably, reliabilism is the mere-available epistemic method of justification oriented with the belief of the outputs of machine learning models considering the theory-and-epistemology-based situation applied in machine learning and in it the technology can be put ahead of scientific comprehension. Meanwhile, we have to accept that a great deal of knowledge work concerns highly complex problem solving and must be understood in contextual, social and relational terms. These aspects have no generic nor universal rules and solutions and, thus, cannot be easily replaced by machine learning (Pettersen 2019).

MODEL JUSTIFICATION

A common view in philosophy of science is that simple idealised models provide more understanding than complex or hyper-realistic models. However, an increasing number of scientists are going in the opposite direction by utilising machine learning algorithms using large data corpuses to create classifications, make predictions, and draw inferences (Sullivan 2020). We get increasingly reliant on machines to draw conclusions from the models created by us, the models which is of toughly comprehensive aspect faced with human beings; beyond that,

it is also the models that 'think' about the world in a different way to us. Apparently, machine learning has outpowered us in their capability to discriminate, identify patterns, and calculate outcomes, which explains why we apply it. Instead of fitting a comparably simplified model by reducing phenomena, our present instructions can be realised in virtue of our machines to construct models according to the demands. However, this appears to suggest what we know is determined by the output of machines, the role of which cannot be followed, explained, or understood by us.

It is of significance that the model used to generate knowledge is also accurate in reflecting how the world works. For the reason that the world reflected by the model can be understood, we believe that the model reflects the world. However, we have constructed a different model. Similar to traditional models, they allow us to make accurate prediction. Similar to traditional models, they play a role in the advancement of knowledge. Nevertheless, some of the new models are beyond comprehension. With weighted triggers representing a large number of variable values which are exerting variously mutual influence, they are existent only in the importance of endless numbers of digital triggers networked together on the basis of satisfying successive layers of networked. Therefore, general principles cannot be drawn from them.

We are left in an odd position due to the dependence upon the unknown models undertaking the source of justification oriented with our beliefs. If knowledge demands our beliefs be justified, knowledge is nearly impossible to construct a class of mental contents. The reason comes from that the justification is a required prerequisite for the models which is present within machines and models beyond the comprehension of human beings. Besides, the matter is beyond that limitation, so that we are incapable of exploring and measuring them as a lay person is doomed to fail in make measurements to the ideas held of a parade of theorists. Essentially, the truth is that the nature held by machine learning justification is not absolutely similar to human justification for its differentials and distinctions. Those machines are likely to be situated in a closer distance to the truth, when compared to humans, could ever be as for understanding how things are.

Sincerely, we can stick to the illusion that the world is going on in an analogous way to the gained knowledge, so do our models, if our machine learning models inspired our own ideas. However, the lost conception comes from the well-fitted assumption once machines began to construct the models of their own. They have exceeded the existing mental capacity volume of us. By nature, the world can be possibly forged to be much alike to how it is represented by our network of machines and sensors than how it is perceived by humans. Currently, machines are functioning automatically, we are often deprived of the illusion that the world merely takes place in an adequately simple way for us tiny creatures to get the hang of.

PRACTICAL JUSTIFICATION

A large part of human knowledge and problems are based on statistics or are provable and reasonable. According to these two dimensions, as long as it is reasonable, it can be divided through logical deduction, using tools such as symbols and rules. Thus, the machine can finally complete the reasoning. Besides, anything that can be statistical can be conducted with big data, statistical means, and various data mining analysis methods, to obtain a model with good enough performance. Consequently, the model output can meet the accuracy required to fit the realistic results, and this kind of problems can be solved through simulation.

In this two-dimensional space, human knowledge can be divided into four categories:

(1) statistical and inferential, suggesting that we know what we know, we know the phenomena and the laws, and we can find the answers to such inferential and statistical questions following the principle, regardless of the method used;

(2) non-statistical but inferential, suggesting that we know what we do not know, and we know the law but cannot observe the phenomenon, such as the position of the earth in a thousand years that is the main work of rule-based machine learning (such as decision trees), but the progress of discovering and proving the law autonomously remains in a very preliminary stage;

(3) statistical but not inferable, suggesting that we do not know what we know, and phenomena can be observed but no laws can be summarised, such as weather systems are almost impossible to obtain satisfactory results by strict inference, but we can make very effective predictions based on various features recorded before weather changes over time. This is the main work of statistical-based machine learning (such as Bayesian classifiers);

(4) non-statistical and non-inferential, suggesting that we do not know what we do not know, neither the phenomenon nor the law. This is the problem that artificial neuron network-based machine learning (such as convolutional neural networks) tries to solve, namely, generating machine knowledge.

The core difference between machine knowledge and scientific or tacit knowledge is that machine knowledge relies on data and scientific or tacit knowledge relies on information. Information is the observable representation of things or the external representation of things. The amount of information in any object is so large that an accurate description of an object requires the description of the forms of all the fundamental particles in it, the relationships between them, and the relationship of the object to its surroundings. Data is part of the information already described. For example, data about an object is usually much less than information, containing only its shape, weight, colour, and species relationships. Information is transformed into knowledge only when it is properly processed and used to perform comparisons, draw conclusions, and make connections. Moreover, knowledge can be understood as information accompanied by experience, judgment, intuition, and value, in which the human being as an epistemic subject plays a key role.

In contrast, machine knowledge can be portrayed as the relationships of data in space-time, which are expressed as certain patterns. The recognition of the pattern is knowledge, and the prediction of the pattern is the application of knowledge. In most cases, knowledge is expressed as a collection of correlations among data, and only a small part of these correlations can be perceived and understood by human beings. This is influenced by the human perception that human sensory experience is limited to three-dimensional physical space and one-dimensional time, and humans can only partially perceive external information. Therefore, the relationships between these data are beyond human understanding and belong to machine knowledge when data cannot be perceived and the relationships between them cannot be expressed by mathematical tools. With the increase in layers and numbers of artificial neural networks, machine learning can handle large-scale complex data, that is, machine knowledge. The main current manifestation of machine knowledge is similar to the full parameters of the neural network in AlphaGo Zero.

Machine learning based on artificial neural networks surpasses humans in two basic bits of intelligence: memory and recognition. Nevertheless, it is still far behind in higher intelligences such as reasoning and imagination. Due to the lack of self-awareness, machine learning

systems capable of producing machine knowledge still cannot be considered epistemic agents, and their ability to produce knowledge is an unconscious epistemic activity. Hence, a lot of epistemological controversies are required to be addressed, such as what constitutes the basic condition of epistemic agency of machines, how to share such machine knowledge more widely (Wheeler 2016).

CONCLUSIONS

After machine learning is embedded in the human cognitive process, the new features of machine cognition come into the view of epistemological research, and there is a need for a new epistemology that can accommodate both similar and different cognitive mechanisms of human and machine in the information processing process. The ability of machines to recognise has led to the creation of 'machine epistemology'. Besides, although there are problems of transparency and interpretability in machine learning, we must acknowledge the existence of some kind of machine knowledge. In other words, the justification for machine knowledge does not need to include transparency and interpretability. When machines have developed certain cognitive functions that are superior to those of humans, how to make them compatible with the human cognitive process, so that machine epistemology and human cognition can form an organic interface, creating a new system of human-machine cognition in harmony, will play a very important role in the future development of AI. The epistemology of human-machine integration is a new epistemological topic in the age of intelligence and big data.

ACKNOWLEDGEMENTS

The work was supported by the National Social Science Fund of China 'Research on Distributive Justice of Scientific Knowledge' (17CZX022).

Received 13 June 2021
Accepted 13 December 2021

References

1. Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; Chatila, R.; Herrera, F. 2020. 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI', *Information Fusion* 58: 82–115.
2. Burrell, J. 2016. 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms', *Big Data & Society* 3(1): 1–12.
3. Carvalho, D. V.; Eduardo, M. P.; Jaime, S. C. 2019. 'Machine Learning Interpretability: A Survey on Methods and Metrics', *Electronics* 8(8): 1–32.
4. De Laat, P. B. 2018. 'Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?', *Philosophy & Technology* 31(4): 525–541.
5. George, J.; Hautier, G. 2020. 'Chemist Versus Machine: Traditional Knowledge Versus Machine Learning Techniques', *Trends in Chemistry* 3(2): 86–95.
6. Katuwal, G. J.; Chen, R. 2016. 'Machine Learning Model Interpretability for Precision Medicine', *arXiv preprint* arXiv:1610.09045.
7. Krishnan, M. 2019. 'Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning', *Philosophy & Technology* 33(3): 487–502.
8. Li, F.; Li, L.; Yin, J.; Zhang, Y.; Zhou, Q.; Kuang, K. 2020. 'How to Interpret Machine Knowledge', *Engineering* 6(3): 218–220.
9. Linardatos, P.; Vasilis, P.; Kotsiantis, S. 2021. 'Explainable AI: A Review of Machine Learning Interpretability Methods', *Entropy* 23(1): 1–45.
10. Pearl, J.; Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

11. Pettersen, L. 2019. 'Why Artificial Intelligence will not Outsmart Complex Knowledge Work', *Work, Employment and Society* 33(6): 1058–1067.
12. Schubbach, A. 2021. 'Judging Machines: Philosophical Aspects of Deep Learning', *Synthese* 198(2): 1807–1827.
13. Sullivan, E. 2020. 'Understanding From Machine Learning Models', *The British Journal for the Philosophy of Science*. Available at: https://doi.org/10.1093/bjps/axz035
14. Walmsley, J. 2020. 'Artificial Intelligence and the Value of Transparency', *AI & SOCIETY* 36: 585–595. Available at: https://doi.org/10.1007/s00146-020-01066-z
15. Wheeler, G. 2016. 'Machine Epistemology and Big Data', in *The Routledge Companion to Philosophy of Social Science*, eds. L. McIntyre and A. Rosenberg. Taylor & Francis, 321–329.

HUIREN BAI

Mašininio mokymosi epistemologija

Santrauka

Straipsnyje argumentuojama, kad mašininis mokymasis yra žinias kurianti veikla, nes mes vis dažniau remiamės dirbtiniu intelektu. Tačiau mašinų atrastas žinojimas yra už žmogiškosios patirties ir proto, jis yra beveik nesuprantamas žmogui. Vadinasi, įprasti kvietimai suprantamumui, orientuoti į „juodosios dėžės" pobūdžio mašininio mokymosi episteminį neperprantamumą, gali būti netinkami. Mašininio mokymosi aiškumo ir suprantamumo problemos kyla dėl „mašininių žinių" suvokimo galimybių. Kitaip tariant, mašininių žinių pagrindimas nereikalauja skaidrumo ir suprantamumo. Straipsnyje nagrinėjama tam tikra mašininio mokymosi epistemologija, pateikti trys galimi mašininių žinių pagrindimai: formalusis, modelio ir praktinis.

Raktažodžiai: mašininis mokymasis, epistemologinis pagrindimas, skaidrumas, suprantamumas, mašininės žinios