Contents lists available at ScienceDirect

Engineering

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

Views & Comments

如何解读机器知识

李发伸 [a,#]，**李廉** [b]，**殷建平** [c,#]，**张勇** [d]，**周庆国** [e,#]，**况琨** [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*

ARTICLE INFO

Article history:
Received 14 September 2019
Accepted 26 November 2019
Available online 17 January 2020

机器知识指的是人工智能所蕴含的知识。本文讨论了如何获取机器知识，特别是获取机器知识中的因果知识，而后者是解读机器知识的过程。通过对物理学和人工智能领域的一些研究方法进行分析，我们提出了解读机器知识的一些原则和模式，同时对一些具体方法，如解读过程自动化和局部线性化进行了讨论。

当前人类已经进入了一个由自然世界、人类世界、信息世界、智能体世界组成的四元社会。智能体已经成为我们这个世界的一种客观存在。智能体能够做出预测、进行判断、表达情感，甚至可以主动调整自己的行为以适应环境的变化[1,2]。因此，我们可以认为智能体是一个具有知识结构和功能的知识系统，即机器知识。本文就如何解读隐藏在智能体中的机器知识进行讨论。本文所指的智能体是指基于硅工艺技术和图灵算法的人工智能机器，如各种学习模型、计算模型、仿真模型等，不包括利用生物或遗传技术构建的智能体。

对于什么是知识，要做出一个令人普遍接受的定义还有待继续深入的研究。这里我们先借用一般的说法，即知识是现象变化的规律。一个智能体能够通过变换将输入转变为输出，或者根据前一个输出调整下一个输出。这种输入和输出以及输出和输出之间的变化规律就是现象的变化规律，所以它属于知识。这种知识被称为初级知识（primary knowledge）。例如，将现象中所有的变化放到一个表格中就是一种知识的表达（如穷举表达）。然而，人们所需要的知识往往不是这种初级形式的知识，而是要在更高层次上经过抽象的知识，即反映现象变化的普遍规律的知识。这种知识被称为高级知识（advanced knowledge）。我们可以根据抽象程度对高级知识继续分层。以Tycho Brahe和Johannes Kepler的工作为例，Tycho Brahe通过详尽的观测，列出了大量行星运行的轨迹数据，这些数据只反映了现象（如行星运行）之间的一种关联。直到Kepler成功地总结出三大规律，并揭示出这些现象之间的因果关系，这些知识才真正成为高级知识。另外，牛顿第二定律是一种更高层次的知

[#] These authors contributed equally to this work.

英文原文: Engineering 2020, 6(3): 218–220
引用本文: Fashen Li, Lian Li, Jianping Yin, Yong Zhang, Qingguo Zhou, Kun Kuang. How to Interpret Machine Knowledge. *Engineering*, https://doi.org/10.1016/j.eng.2019.11.013

识表达。关联关系和因果关系都属于知识，但它们却处于不同的层次。就人类获取知识的过程而言，通过观测获取现象之间的关联关系是最基本的科学活动，而要获取现象之间的因果关系，则需要我们对所观测到的数据背后的现象进行分析和归纳。由于人类总是喜欢且执着地追求现象变化背后的“为什么”，因此因果关系在人类科学体系中占有重要的位置。

在本文中，我们主要对人们是否能够以及怎样从智能体获取因果知识这一问题进行了关注。这一过程就是解读机器知识。经过训练的智能体可以完成十分复杂的工作，而且其中一些成果已经超越了人类上千年的文化积累。然而，我们仍然不清楚这些智能体究竟是如何成功的。例如，像神经网络那样的智能体，过度的拟合训练并不能使之具有更好的泛化能力。我们不清楚神经网络成功的边界在哪里。我们不知道如何设计神经网络的结构去完成预定的任务。我们不知道是否可以更换训练集使神经网络的表现更加出色。我们甚至不知道神经网络是根据什么来做预测，即它是基于数据还是基于特征。一句话，我们不知道如何信任一个智能体。

到目前为止，因果关系仍然是人类理解自然世界的根本基石，而用概率思维描述的关联关系则是推动我们理解世界因果机制的表层现象。Pearl [3]曾说过:“回顾过去，我所遇到的最大挑战是摆脱概率思维，并且我接受两种观点：第一，人们并不是从概率角度思考，而是从因果效应的角度思考；第二，因果思维很难通过概率语言来描述，它需要一种属于它自己的新语言。”第一种观点讲的是事实，即科学知识不是用概率思维的形式来表达的，而是通过因果思维进行表达的。第二种观点讲的是如何解读因果思维。Pearl认为人类现在还没有发明出解读因果思维的数学工具。遗憾的是，当前最受青睐的智能体大都是以概率方式运行的，其所表达的现象之间的关系都属于关联关系。我们能从这些关联关系中解读出蕴藏在其中的因果关系吗？这是一个很有挑战性的问题。如果人类和智能体之间无法相互交流和理解，或者如果人类不能将智能体的知识翻译为因果形式，那么人工智能的发展将会遇到极大的障碍，甚至隐藏着风险[4]。

物理学是一种典型的利用因果关系来解读自然世界的科学。自然世界也可以被看作是一个巨大的智能体，其中的现象每时每刻都发生着变化。人类在认识自然世界这些变化及其规律时，采取了因果关系的解读形式。他们希望对现象变化背后的规律给出清晰而准确的表达。这种解读形式主要是通过正则表达式和数学表达式来实现的，从而使得人们不仅可以描述已经发生的现象，而且可以预测未来可能发生的现象，其中后者尤为重要。由于对自然世界的实际运行规律无法直接获取，所以人们只能通过现象观测来“猜测”其所蕴含的规律。即使有大量的与现象相关的数据，我们也很难准确、完整地从这些数据中总结出相应的规律。因此人们在解读自然世界时采用了两条原则（或者信仰），这在牛顿的*Mathematical Principles of Natural Philosophy, Volume III: On the System of the Universe* [5]书中有明确阐述。以下是四条“哲学中的推理规则”中的前两条：

（1）最简描述原则（如Occam’s razor）：没有什么比既真实又足以解释其现象的原因更能说明自然事物的原因；

（2）功能相似原则：对于相同的自然现象，我们必须尽可能地找到相同的原因。

对于物理学来说，一些具有基础意义的定律和原理，不仅是对自然世界中现象变化规律的高度抽象和因果描述，而且也遵循了上述两条基本原理，从而形成了当前人们对于自然世界基本规律的认知和构建了人类自然科学知识的结构。例如，既然任何一次测量都不能精准验证牛顿第二定律，那么我们为什么还要接受它呢？这里面就隐藏了一种公认的原则。

让我们回到对智能体的解读上来。在绝大多数情况下，虽然我们可以知道智能体的结构，但我们无法预测智能体的行为，这就像我们不能根据大脑的神经连接结构来判断它会做怎样的思考一样。我们能观测的只是大脑输入和输出之间的关系，即数据。对于任何一个智能体，只要有足够充分的观测和大量的观测数据，理论上我们都可以通过归纳计算来获取其中的因果关系，而无需考虑智能体内部的结构和运行模式。也就是说，只要与智能体的外部性能（如功能）高度吻合，我们就可以认为该因果关系成立。这就是“功能相似原则”。这种方法在物理学中得到了充分的体现。例如，宇宙就像一个巨大的时钟，我们只能从它的外部运行来猜测内部的结构。通过连续不断的观测及其精度的提高，我们的猜测与观测到的现象越来越一致，但是我们可能永远也无法知道宇宙内部的实际结构。尽管如此，物理学仍然推动了人类的社会发展和科学进步。

虽然人类对于因果关系的探索已经有几千年的历史，但长期以来，人们对因果关系的描述还一直停留在

定性的和经验主义的阶段。直到20世纪70年代以后，C. Granger、J. Pearl和D. Rubin陆续提出基于数学表达的因果关系定义后，人类才真正开始在数据基础上对因果关系进行定量化研究。Pearl关于因果关系的描述方法比较系统化，该方法能够处理变量之间的混杂干扰、发现隐性变量的存在、解决反事实等归因问题。基于Pearl因果关系的研究在许多实际应用中取得了很好的效果，该研究可以被用于解决因果悖论问题。因此Pearl因果关系已成为人工智能理论和应用的重要方法。从原则上说，Pearl的因果关系与Fisher的实验设计具有相同的科学假设和数学基础，因此Pearl的因果关系的数学基础是牢固的。

然而，Pearl的因果关系仍然存在一些问题，以至于其在稍微复杂的问题上表现得不尽如人意。例如，Pearl的因果算法对于数据的分布和数量要求较高，这在很多实际应用中是很难满足的。再则，Pearl的因果关系对于隐藏变量比较敏感，因此，观测数据的不充分或者不准确会对计算结果产生很大影响。用于构建Pearl因果关系及其算法的结构方程模型或者因果结构图模型也还存在诸多的不确定性。

统计学家Imbens和Rubin [6]提出了另一种因果模型，该模型被称为潜在结果模型，利用该模型通过研究数据中所反映的潜在结果和现象关联，可以挖掘其内在的因果知识。Rubin的因果模型已被广泛应用于实际问题，尤其是那些需要因果知识去帮助决策的领域，如医疗诊断、公共政策制定等。但是Rubin的因果模型同样存在一些问题，例如，它对数据的假设过于强大，而且其中一些假设在实际问题中是不可测试的。

除了继续深入研究Pearl和Rubin的因果关系及其算法外，当前我们也研究了一些其他方法。虽然因果关系无法被这些方法直接计算出来，但它仍然可以从智能体的知识中揭示出一些深刻的关系。这些方法来源于对物理学和人工智能的一些研究。当物理学家在采用因果关系去理解自然世界觉得有困难时，他们也会借助机器学习方法。例如，他们使用机器学习方法去理解多体系统的Langevin方程以及Liouville方程的Boltzmann描述（BBKGY 截断）。解释算法也被用于人工智能领域，用以理解复杂数据或特征之间的内在关系。使用智能体来解读智能体是一个绝妙的想法。实际上，当前各种各样的智能体（或者学习模型）在透明程度上是分层次的。有些智能体对于人类就比较透明，如线性模型、决策树模型，而有些智能体对于人类就比较模糊，如神经网络、Monte Carlo搜索树模型。令人遗憾但却很有趣的是，越是模糊的智能体，其学习能力就越强，而且它所蕴含的知识也越丰富。如果直接解读智能体有困难，我们可以考虑通过一个较为透明的智能体来解读不太透明的智能体。这个过程是可递归的，使得解读内容越来越容易被人类理解[7]。

通过计算影响函数，我们可以分析智能体中数据或者特征的重要性，从而尽可能地分析出哪些因素（原因）会导致智能体表现出这样的行为。我们也可以通过分析数据的质量和分布来寻找更好的观测数据，这对医学诊断和物理观测都有非常重要的意义。

对于给定的输入数据，智能体会给出相应的输出结果（或者下一步动作）。通过计算各个输入数据特征的Shapley值，我们可以估计出不同特征对于该输出结果的贡献值。那些贡献大的特征有可能是智能体行为的原因[8]。

根据一般的数学原理，复杂智能体的局部行为应该类似于一个线性系统。因此，根据功能相似性，我们可以考虑在局部范围内采用线性模型（如线性回归）来替代原来的智能体[9]。线性模型对于因果关系有很好的透明性，通过对其回归系数进行适当的处理就可以得到相应的因果关系。与此同时，通过残差分析，我们还可以确定这种近似处理的精度，以及其他因素对于主要变量的敏感程度。

另一种简单的方法是，利用较为透明的模型T来学习模糊的模型V，进而通过输入数据x得到被$[x, V(x)]$，标注的数据，其中，$V(x)$表示V关于x的输出。然后，我们使用这些数据对T进行再学习。如果T和V有基本相同的行为，那么根据功能相似性原则，我们可以认为T和V有相同的因果知识。这一方法在分析智能体内部缺陷以及黑盒攻击时取得了良好的效果。

人工智能的出现为人类发现新知识开拓了更多的途径。通过解读智能体的知识，我们可以丰富自身的知识体系，从而更好地服务于人类的发展。目前有关智能体的解读研究还有待进一步深入。随着研究理论和方法的不断完善，人类与智能体之间的关系也会达到一种高度和谐，他们相互之间将实现更好的交流与沟通。这在人类进化史上将具有里程碑意义。

致谢

本文是根据2019年7月在兰州大学举行的“机器知

识和人类认知沙龙”上参会人员的发言内容整理而成。所有参会人员都对本文做出了贡献，我们对其他参与者表示感谢。他们是物理学和计算机科学的专家，具体包括黄亮（物理学，兰州大学）、安宁（计算机科学，合肥工业大学）、杨磊（物理学，中国科学院近代物理研究所）、吴枝喜（物理学，兰州大学）、刘礼（计算机科学，重庆大学）、张家琳（计算机科学，中国科学院计算技术研究所）、俞连春（物理学，兰州大学）。

Compliance with ethics guidelines

Fashen Li, Lian Li, Jianping Yin, Yong Zhang, Qingguo Zhou, and Kun Kuang declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Pan YH. Special issue on artificial intelligence 2.0. Front Inf Technol & Electron Eng 2017;18(1):1–2.
[2] Pan YH. 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; 2016. p. 1135–44.