·科学技术哲学·

迷你图灵测试与儿童教育

Mini-Turing Test and Children's Education

张端 /ZHANG Duan[1]　吴小安 /WU Xiaoan[2]　和继军 /HE Jijun[1]

（1. 首都师范大学初等教育学院，北京，100048；2. 西北工业大学马克思主义学院，陕西西安，710072）
(1. College of Elementary Education, Capital Normal University, Beijing, 100048;
2. School of Marxism, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072)

摘　要：相当多证据表明，因果学习和因果理解能力极大地增强了人类操纵物理世界的能力，是将人类与其他灵长类动物区分开来的重要因素。如何让蠢笨的机器人因果地思考、回答“为什么”的问题，甚至理解此类问题的意义，是实现人工智能的关键之一。朱迪亚·珀尔认为要实现类人智能，先从模仿孩子的智能开始，于是提出“因果推断引擎”帮助未来的人工智能进行因果推理，以通过迷你图灵测试，甚至成为明辨善恶的道德主体。本文的目标是，既根据人工智能的理论设想和建构目标为儿童教育的发展提供思路，又根据儿童教育的实践来反观人工智能的因果模型径路。

关键词：因果推断引擎　反事实　自由意志　科学教育

Abstract: There is considerable evidence to prove that causal learning and causal understanding ability have greatly enhanced our ability to manipulate the physical world and they are important factors that distinguish humans from other primates. How to get stupid robots to think causally, answer the questions raised with “why” and even understand the meaning of such questions is one of the keys to realizing artificial intelligence. Judea Pearl believes that to achieve human-like intelligence, we must start by imitating the intelligence of children, so he proposed a “causal inference engine” to help future artificial intelligence conduct causal inference, pass the mini-Turing test, and even become a moral subject who can discern good from evil. This paper attempts to provide some different thinking dimensions for the development of children's education from the basic assumptions and construction goals of artificial intelligence, and at the same time reflects on the current practice of artificial intelligence through children's education.

Key Words: Causal inference engine; Counterfactuals; Free will; Science education

中图分类号：TP18　**文献标识码：**A　DOI: 10.15994/j.1000–0763.2022.09.004

基金项目：国家社科基金后期资助项目“实际因果的结构方程径路研究”（项目编号：21FZXB040）；教育部人文社会科学项目“实际因果前沿问题研究”（项目编号：21XJC720002）；中央高校基本科研业务费专项资金资助项目“因果模型前沿问题研究”（项目编号：G2021KY05109）。

收稿日期：2022年2月24日

作者简介：张　端（1983–）男，陕西蒲城人，首都师范大学初等教育学院副教授，研究方向为科学教育及小学教师教育。Email: zhangduan@cnu.edu.cn

吴小安（1984–）男，江苏南通人，西北工业大学马克思主义学院副教授，研究方向为因果模型、条件句逻辑和死亡哲学。Email: wuxiaoan1984@126.com

和继军（1979–）男，河北邢台人，首都师范大学初等教育学院教授，研究方向为自然环境与生态保护、科学教育及小学教师教育。Email: hejiun_200018@163.com

一、导　　论

所谓人工智能，大体思路就是把“人的部分智能活动机器化，让机器具有完成某种复杂目标的能力”。而人脑中一个决策和判断的做出是数十亿神经元相互作用的结果，且神经元之间地交互无比丰富又异常迅捷。因为今天的科学还不具备洞悉这数十亿神经元如何交互的能力，所以当代人工智所做的也还只是对人的宏观认知模式的一种模拟。

人工智能的先驱之一艾伦·图灵（A. Turing）曾提出这样一个问题：怎么才能说计算机像人一样在思考？[1]他提出“图灵测试”，即向计算机提问，如果它对问题的回答，不能让人区别出与谈者是计算机还是人，那么这台机器就可以算是一个思维机器。但要模仿一个成人的智能是困难的，随着教育和阅历，她的智能水平一直在发展和推进。所以图灵主张，不妨从模仿孩子的智能开始，循序渐进，最终实现智能：

> 与其试图编写一个模拟成人思维的程序，何不尝试编写一个模拟儿童思维的程序？然后让它再接受一个适当的教育过程，最终它将拥有一个成人的大脑。儿童的大脑就像从文具店买来的一本记事本，机制很少，而且有很多空白页。[1]

图灵的时代，认为孩子的大脑“只有很少机制，是有很多空白页的记事本”，正如皮亚杰（J. Piaget）指出，学龄前儿童是没有因果思维的。[2]但是近三四十年的心理学研究已经基本推翻了这个结论，转而主张孩子的大脑中有丰富的表征和学习机制，所以他们才那么快速和有效地学习。比如，在因果方面，甚至2岁的儿童就能进行因果预测，提供因果解释，并理解反事实的因果断言。[3]，[4]但图灵主张制造出一个拥有儿童智能水平的机器人（childlike intelligence），然后才能制造出一个类人智能水平机器人（humanlike intelligence）的思路则得到了人工智能学家们的支持。

人的智能体现在很多方面，比如，视觉、自然语言，推理能力。在珀尔（J. Pearl）看来，实现拥有儿童智能水平的机器人的关键一步是：掌握因果关系。因为对因果的理解是人类认知最普遍也是最不可或缺的一环。[5]只有所造的机器人理解因果，才能把对于这个世界所知教给它们，比如，可以告诉机器人，对于一个新冠的病人，你用冰块给他物理降温，并不能缓解他呼吸困难的问题；如果现实中一个人纵火导致了火灾，那么要教会机器人，如果这个纵火犯选择不纵火，会发生什么。

具体到实践和理论上，在制造出一个理解因果的机器人，至少有两个疑难要面对。首先，机器如何从环境中获取因果知识？其次，机器如何处理因果信息？第一个问题并不是珀尔关心的，尽管它同样是一个困难的问题，涉及到如下输入的复杂组合：积极的实验、被动的观测以及程序员的输入。之所以不关注，还有如下一些理由，首先，如果表征问题不解决，就算获取了知识也不知道如何来存储它以供将来调用；其次，好的表征对如何获取知识也有其启发作用。且“人工智能认知研究的一个主要贡献，就是确立‘表征第一，获取第二（Representation first, acquisition second）’的范式”。他的主要贡献和努力在于回答第二个问题，即给出机器如何表征因果知识的方式，并说明这种表征方式如何使得机器能够快速地访问相关的信息，正确地回答因果的问题。珀尔称之为“迷你图灵测试（mini–Turing Test）”。通过迷你图灵测试是珀尔几十年的主要努力。

二、因果推断引擎蓝图

如何能让机器人在因果交谈中游刃有余，回答起因果问题从容不迫，珀尔提出给机器人装备因果推断引擎的设想，[6]装备这个引擎的机器人能够获取、存储因果知识，并以之为基础做因果推断，进行有效的因果交流。因果推断引擎的目标如下：（1）因果知识的表征及存储；（2）对人如何做因果推断的模拟；（3）实现知识的迭代与更新；（4）整个流程算法可实现。下图1就是珀尔因果推断引擎的详细蓝图，

此图同样也反映了，在今天的科学应用中，因果模型如何工作，如何与数据交互。接下来，本文将阐述其理论设想，并通过吸烟肺癌的例子来具体说明。

在珀尔看来，每一个因果推断任务都必然依赖于一些断言性的、超出数据之外的假定，这些假定源自知识$\mathbb{K}$即："过去的观察、以往的行为、接受过的教育和文化习俗等所有这些与目标问题相关的内容"。从阐释学的角度，所谓知识可以理解为"前见"，"微观分析，遗传因素、家庭背景、童年经历、学习与工作过程，特别是一些有别于他人而其独有的创伤性经历"，"宏观地看，文化社会背景、传统风俗与观念、民族心理结构与偏好、时代认知水平与伦理倾向"。[7]大体上可以把知识的来源分为三个部分：首先，进化；其次，文化传承；最后个体经历和体验，所有这些构建了人们对于这个世界的理解，并以此来解决问题和完成任务。

在图1中，之所以把"知识"放在虚线框里，是因为这些知识还在推理主体的脑袋中隐而不彰。但是当考察具体问题时，就会调用储存的知识，一开始也许是粗略且宏观的假定集$\mathbb{A}$，一些模式（pattern），如果了解的够多，也会进一步细化，细化为一个明确的方程或者逻辑表达式。这些假定是研究者出于科学的理由或者坚定的信念所要捍卫的。比如，当考察吸烟是否导致肺癌，作为这个领域的研究者，他应该对于吸烟、肺癌、焦油和基因之间的关系有自己的理解。

在珀尔看来，如何来表征这些模式和假定是很重要的，因为涉及到三个方面问题：首先，它会影响一个人正确且确切地说出这些假定。如果表达的语言不够简洁，呈现的方式不够直观，不仅不会促进反而会妨碍认为对于假定的获知；其次，它会影响从这些假定出发推导出它们所蕴涵的结论，如果推导的过程过于繁复，自然会影响到人们对其使用的意愿和信心；再次，当遇到更具说服力的证据时，可以便捷地扩展或者修改这些假定。想想人工智能的符号主义流派用一维的现代逻辑的形式语言来记录四维事物的时空信息结构的低效与复杂，以及表达四维事物的动态规律时所面临的框架问题（frame problem）。[8]图语言是一种非常直观易读的语言，对于认知、理解和表征世界的知识有其不可替代的优势，而且图理论也在近几十年取得了长足的发展。

图语言定性地编码了研究者对于所探究情形的因果假定，并且其呈现方式非常透明，通过图，任何人都可以便捷快速地读取到其中所蕴涵的因果假定，相关领域专家会就其蕴涵进

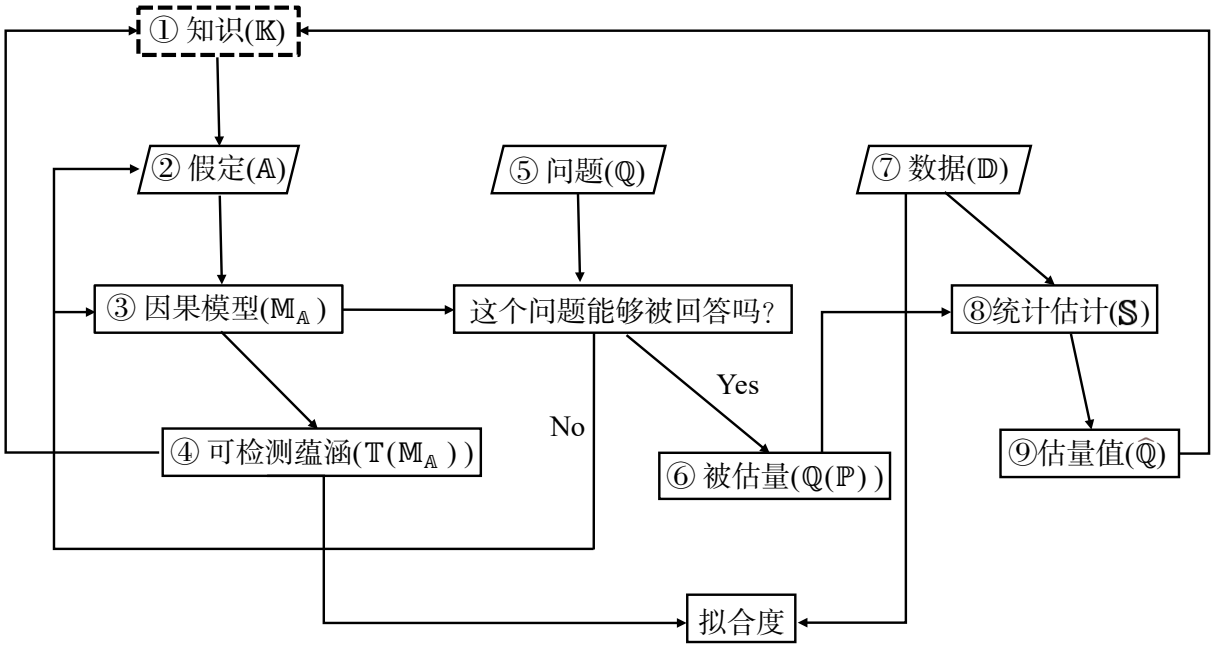

图1 未来人工智能处理因果推理的推断引擎蓝图

行探讨和辩论，以最终确证或者否证图本身的合理性。如果没有因果图来清楚地表达实质的因果假定，那么做推断时，就会有犯错误的风险。鲁宾（D. Rubin）就给了一个例子，业界传奇费舍尔（R. A. Fisher）在估量一个处理效应时，如何犯了对一个中介结果做调整的基本错误。[9]

> 研究者必须绘制提出一个因果图，其反映的是他对于某个研究课题所涉及的因果过程拓扑结构的定性判断，或者更理想的是，它所属的专业领域的研究者对于该研究课题的共识。（[10]，p.67）

当然图语言只是珀尔结构因果模型的一个理论资源，当要处理反事实问题时，图的工具就不够用了，所以他把结构方程模型引入进来。一个结构因果模型M是一个有序三元组$<U, V, F>$，其中U表示外在变量的集合，表示那些不在模型考虑范围之内的因素。根据相对论，如果要确定某个事件的律则性（nomologically）充分条件，就需要对发生在它之前的整个逆时间光锥的某个截面有一个描述。就算相对粗粒化地来刻画，其因素的数量也是不可想象的，必然得有所取舍，所以用U来概括那些暂时不予考虑的因素。V是内在变量的集合，即那些被外在变量或者其他内在变量所决定的变量。F是结构方程的集合，每一个内在变量都有一个对应的结构方程，即对于每一个内在变量$V_i \in V$，存在一个从$U_i \cup PA_i$到V_i的映射f_i：

$$v_i = f_i(pa_i, u_i),\ i=1, \cdots, n.$$

其中$U_i \subseteq U$, $PA_i \subseteq V/V_i$。一般称PA_i为V_i的父集，PA_i中的变量对于V_i有直接的因果影响，每一个PA_i的选择都不是任意的，而是“表达了在决定V_i的值之前，建模者对于自然（或者造物主）必然会探求的那些变量的理解”。

首先，每一个因果模型都有一个与之对应的因果图（表示为$G(M)$）。这个图中的节点分别对应着一个内在或者外在变量，如果某个变量在一个结构方程的右边，那么就画从这个变量到方程左边变量的一个箭头，最终就得到一个因果图。如果因果图是一个有向无环图，就称与之对应的模型为半–马尔可夫模型（semi–Markovian model），它具有如下性质：当确定了U的值为u，那么就唯一确定了这个模型中所有变量的值；或者如果确定U中变量的一个概率分布，那么就可以唯一确定定义在内在变量集V上的一个分布。如果除了因果图是有向无环图，外在变量U还联合独立，那么就称这个模型为马尔可夫模型（Markovian model）。当然，在很多时候，确定具体的结构方程是非常困难的，特别方程是非–线性的时候。但哪怕只是知道表示变量之间定性关系的因果图，很多因果问题还是可以得到回答，即在算法上可以判定：上述图模型对于所要求的因果效应的识别是否充分，哪些协变量需要测度。

其次，每一个结构方程（$v_i = f_i(pa_i, u_i)$）表征了一个稳定和自主的物理机制（a stable and autonomous physical mechanism），即可以设想在不改变其他机制的情况下来改变它。这实质上就是以一种模块化的布局方式来组织知识，使得以极小的额外信息就能够预测干预的效应。再次，上述模型的构建蕴涵着建模者很强的先见：变量之间如何彼此因果影响。比如，确定V中所有那些对于V_i有直接因果影响的变量集PA_i。这种通过因果信息来够建模型的好处在于，因果关系是人类知识建构的基石，以之为基础所做出的判断也会更有意义、更可接受、更可靠。根据上述描述，模型刻画了具体情形，当作用于这个模型中某个（某些）变量的干预发生，根据自主性，干预对于这个模型的改变是定域的（local），并不波及除了和干预变量相关的机制之外的其他机制。

在构建因果图之前，似乎要给何为因果关系一个好的定义才行。哲学家数千年来对于因果的本质为何讨论不休，但从未有过共识。相较于哲学家的劳心费力，[11]，[12]珀尔关于两个变量之间因果关系的确立简直有点形同“儿戏”：

> 如果变量Y“听从于”（listen to）变量X，并根据所“听到”的内容决定自己的值，那么变量X就是变量Y的一个原因。（[10]，p.XX）

首先，“听从”关系是一种很弱的关系，

要满足听从关系，变量之间并不必然要是因果关系，逻辑关系或者纯粹的数学关系同样也满足要求，比如，W_F表示最终体重，W_I表示初始体重，$Y=W_F-W_I$表示增加的体重，在画因果图时，还是可以画从W_F和W_I指向Y的箭头；其次，你可能会觉得这个建议似乎过于草率，但其实不用担心，因为由所构建出的因果模型或者因果图（$\mathbb{M}_{\mathbb{A}}$），根据d-分离，双子网络等等技巧，可以挖掘出这个图所蕴涵的变量或者反事实变量之间的条件独立关系，即可检测蕴涵（$\mathbb{T}(\mathbb{M}_{\mathbb{A}})$），如果所收集的数据和其蕴涵相抵牾，那么就得修正这个图。再次，鉴于数据和图模型几乎不可能会完美一致，所以需要另外一个引擎来计算“拟合度”，即模型与数据相容的程度。通过数据，可以挑战某些因果假定的合理性，以及判定，在多强意义上，一个替代的因果关系可以更好地解释观测到的数据。在上述吸烟肺癌的研究中，研究者接受的定性假定可以表征为如下的因果图及参数未知的结构方程集，见图2。

根据图理论，编码在图中那些实质因果假定是由否定性的因果断言，即图中结点之间因果联结的缺失，所表达。缺失的联结有两类：[3]第一类是缺失箭头，比如，图2中变量X和变量Y之间没有直接指向彼此的箭头。它表示我们预设如下假定，那些被忽略的直接影响X的因素和那些直接影响Y的因素不相关，而且当通过干预固定Y的父变量（即变量Z）的值之后，X对于Y没有因果影响。在建模者看来，吸烟对于肺癌没有直接因果影响，两者之间的影响要通过肺中焦油的累计为中介。所以缺失的箭头编码了排除限制（Exclusion restrictions），即把方程所排除的变量添加到方程中去（比如把变量X添加到方程$z=f_2(x, u_2)$中去）并不会改变方程的结果。用潜在结果模型的记法表达为：

$$Y_z=Y_{zx},\ Z_x=Z_{xy},\ X_y=X_{zy}=X_z=X$$

第二类是缺失双向虚线弧，比如，图2中没有联结变量X和变量Z的双向弧。它表示那些直接影响X的因素和那些直接影响Z的因素不相关。即在建模者看来，那些影响吸烟的因素和那些除X之外影响肺中焦油累积的因素不相关。具体而言，缺失的双向弧编码的是独立性限制（Independence restrictions），比如，X和Z之间（以及Z和Y之间）没有双向弧，那就意味着在$P(u)$中U_1和U_2是独立的，如果确定$y=f_3(z, u_1)$中z的值，那么y的值就完全由u_1决定；如果确定$z=f_2(x, u_2)$中x的值，那么z的值就完全由u_2决定。根据U_1和U_2是独立的，很自然地，Z_x和Y_z独立，同理有Z_x和X独立。

所以图2中所缺失的弧，而不是已有的弧，应该得到更多地关切和更仔细地证成。毕竟添加一个额外的双向弧最糟糕不过也就是参数不可识别，但是删掉一个存在的双向弧很可能就会产生错误的结论。在构建模型时，应当默认任意两个节点之间存在双向弧，只在有良好动机证成的情况下才可以认为双向弧不存在，比

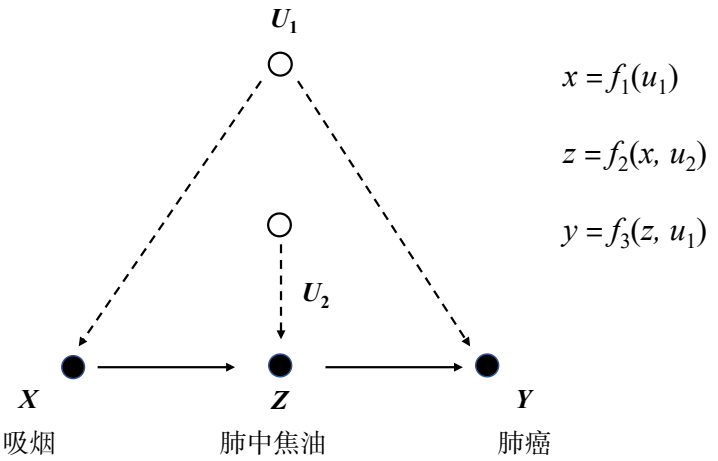

图2 说明吸烟对肺癌影响的因果图

如，不可能存在两个变量有一个共同原因但选择偏倚存在。尽管永远不可能知晓影响变量的所有因素，但个体所具有的实质知识有时可以断定一个可能的共同因素的影响并不重要。

接下来，用因果语言来表达因果问题（$\mathbb{Q}$），它可以是一个干预或者展望式（prospective）反事实问题，也可以是反省式（retrospective）反事实问题等等。如图2的例子，要求吸烟对于肺癌的因果影响，可以用因果语言形式化为：

$$P(y|do(x))$$

或者假定张三吸烟30年且得肺癌的情况之下，假定设使他不吸烟，那么他得肺癌的概率，可以用因果语言形式化为：

$$P(y_x|x', y')$$

那在给定因果模型和数据的条件下，上述问题能够被回答吗？珀尔给出在基于图的展望式反事实的转换规则，（[13]，pp.85–86）给出了计算反省式反事实的三步骤，（[13]，p.206）给出了计算自然直接效应（Natural Direct Effect, NDE）和自然间接效应（Natural Indirect Effect, NIE）的中介公式（Mediation Formula），（[13]，p.132）甚至在决定论之下的实际因果问题，也给出一个在算法上可实现的因果模型定义。[14] 珀尔的框架是否还遗漏了一些反事实和因果问题，目前还有争议，（[15]，p.1136）但毋庸置疑，大部分的因果问题和概念都可以在因果模型的框架内得到表征和回答。

如果上述问题不可回答，即回答为“否”，那么就要重新考察假定和因果模型的合理性或者进一步优化模型；如果回答是“是”，那么根据上面所给出的公式和计算规则，就可以得到一个被估量（Estimand，$\mathbb{Q}(\mathbb{P})$）：

> 被估量的工作在于告诉我们，如何把统计量转换为一个基于模型假定的表达式，而这个表达式在逻辑上等价于那个因果问题（比如：$P(y|do(x))$）。（[10]，p.XXI）

在吸烟肺癌的例子中，考察图2，得先考虑影响肺中焦油累积各种因素，再考虑在焦油累积之下对于肺癌的影响，最后再做加权计算：

$$\begin{aligned}(P(y|do(x)) &= \sum_z P(y|do(z))\, P(z|do(x)) \\ &= \sum_z P(y|do(z))\, P(z|x) \\ &= \sum_z \sum_{x'} P(y|z, x')\, P(x') P(z, x) \\ &= \sum_z P(z|x) \sum_{x'} P(y|z, x')\, P(x')\end{aligned}$$

方程右边最终所得到的式子就是“被估量”，是问题的目标量的一种计算方法，最终的被估量是以概率的形式表达，即只要通过可观测的数据（$\mathbb{D}$）就可以求出目标量的估计值。被估量可能并不唯一，根据*do*–演算，可能得到很多被估量，但有些被估量中的变量不可观测，比如，图2还可以有：

$$(P(y|do(x)) = \sum_z P(z|x) \sum_u P(y|z, u)\, P(u)$$

上式中基因（U）的数据不可观测。

如果在因果模型中，一个变量和它的父集之间的函数关系得到确切的说明，那么反事实问题的算法化也成为可能，比如，图2中，张三吸烟并得肺癌，而考虑设使他不吸烟，他不得肺癌的概率：

$$P(y_x|x', y')$$

根据反事实算法的三步骤（[13]，p.206）可以计算出这个反事实的被估量和估计值。

有了被估量，又有了数据，会得到估计值（$\hat{\mathbb{Q}}$），但是得到的估计值只可能是一个近似值，因为所收集到的数据永远是有穷的，它属于一个理论上无穷的总体，就会有选择偏倚（selection bias）的问题。但是借助于强大的机器学习技术，统计学已经发展出了多种方式来应对这种不确定性，以及对不确定的程度给出一个统计估计（$\mathbb{S}$），比如，最大似然估计、倾向得分、置信区间、显著性检验等等。（[10]，p.XXII）如果模型正确，数据充分，也就得到因果问题的一个答案，如果为真，比如，确定了吸烟导致肺癌，那么就可以把它添加到自己的科学知识库中去，这样就实现了自我知识的增长。

总之，珀尔力图实现的是人的意识能力的一个方面：因果推断能力。在他看来，不管是来自基因的遗传，还是来自经验的习得，人脑中存储着海量的因果信息，这些知识可以用一种简洁的数学语言来刻画：逻辑、图、数学方程、反事实变量等等。当遇到新的问题，就会调用储备知识，再结合获取的信息数据，来力图解决问题。这个实践本身也让我们获得新知或者认识到自己的无能为力。当然智能并不一蹴而就，是一个由弱到强的学习过程，对未知现象的探究，对遭遇问题的解决、新的实验方法的设计及对自然的拷问，在摸索砥砺中，智能本身也经历了一个不断更新迭代的过程，其“智”也越发彰显，也许终有通过图灵测试的一天。

三、反事实算法与自由意志的幻觉

在这些围绕着因果推断引擎的诸多努力中，“反事实的算法化”是珀尔工作的重要成果之一，因为反事实推断能力对于智能的实现有其重要意义：

> 反事实（counterfactuals）是道德行为和科学思考的基石。回溯自己过去的行为以及设 想其他可能情景的能力是自由意志和社会责任的基础。反事实的算法化使“思维机器”（thinking machine）有可能习得这种人类特有的能力，并掌握这种目前仍为人类所独有的思考世界的方式。（[10]，p.XVII）

所谓意志（will）可以理解为做某事的意志（Will to do something），所谓自由意志，可以理解为，你做某事的决定完全取决于你自己，比如，你决定抽烟或者不抽烟，你完全遵循自己的内心，在理性审慎地估量之后做了一个选择。当然，上述对自由意志的描述非常粗陋，如果严格分析下去，可以证明，任何选择是被决定的，并没有所谓真正的自由意志。可以区分两个描述层面：在神经元层面，一切过程都是决定论的（在神经科学家看来，人的身体和大脑是一个遵守决定论定律的自动化系统），而在认知层面，我们鲜活地感受到选择的自主性。（[10]，p.337）生而为人，似乎一直拥有着一种自由意志的感觉或者幻觉，“我是我命运的主人，我是我灵魂的舵手”，人生的很多行为选择完全出自于我们自己，如果我想伸手拿这个瓶子，就可以伸手拿它，如果我不想，我就可以不伸手拿它。

那么在哪种意义上，机器具有“自由意志”（free will）呢？珀尔指出，“如果机器有做 $X=x$ 的意图（intention），它在意识到了这个意图之后，转而选择做 $X=x'$”，那么这就说明机器具有自由意志。进化不辞劳苦地赋予我们一种自由意志幻觉的目的何在或者功能何在呢？即功能问题（function）。如果自由意志的“幻觉”有其合理的功能，那么如何赋予一个被算法决定的机器人以自由意志呢？则是模拟问题（smiulation）。得先确定功能问题的重要性，工程的实现才有其动机和动力，哪怕工程实现极端困难。

首先，珀尔指出，正是因为有了自由意志，才能有意义的谈论意图（intent），并通过设想其他可能的替代情形，来让意图遵从理性，即给予行动本身以理性的解释。比如，一个人考虑买华为（$do(A)$）、小米（$do(A')$）和苹果（$do(A^*)$）的手机，他会考察这些行动本身将会带来的不同后果的偏好，对虚荣心满足、性能需求、当下经济状况等等，即确定效用函数，在综合地考虑了这些可能及其蕴涵之后，最终决定买华为，如图3。

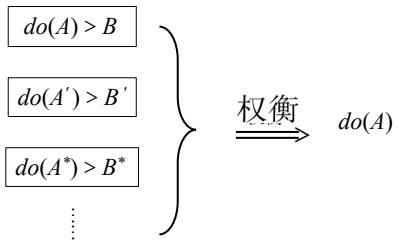

图3 各种可能行动的权衡

上述理性思考过程已经在结构因果模型的框架内被算法化。

其次，机器人一旦有了自由意志，它和人类之间沟通的不确定性会减少，因为机器一旦相信自己有自由意志，当遇到具体情况时，它就能够观测到自己的意图，并采取另一种行动。

比如，你的家庭机器人，在凌晨你熟睡的时候，打开吸尘器做卫生，你被吵醒，非常恼火，于是你大声告诉它："你不应当在我睡着的时候使用吸尘器（You should not have used the vacuum while I'm sleep）"。通过这个反事实表达，我们是要让机器人回到过去，把自己的经验重新体验一番，并修正决定其行为的程序。比如，通过这个反事实交流，机器人明白你被吵醒是你生气的原因，它内置的模式也让它知道，噪声过大会影响一个人的睡眠，而噪声的产生是因为它打开了吸尘器，根据反事实算法，它可以反事实地设想：设使没有打开吸尘器会怎么样？推断出不会产生噪音，没有噪音不会吵醒正在睡觉的人，而一个没有被吵醒的人也就不会因为自己被吵醒而生气。既然你的生气行为让它意识到了一种可以规避你再次生气的可能选择，那么下一次它就不会在你睡着的时候打开吸尘器，而是选择不打开吸尘器。其实很多时候，我们也是通过这样一种方式与孩子沟通，慢慢规训孩子的行为，导其入正途。（图4）

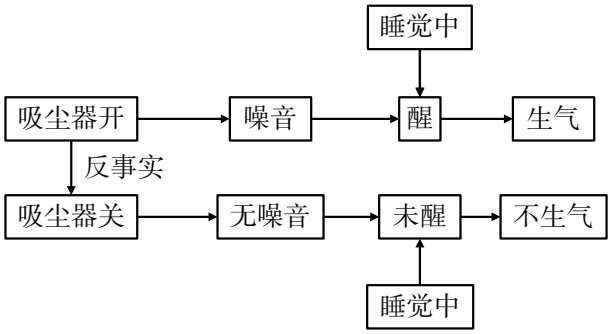

图4 机器人的反事实推断过程

所以，如果机器人有自由意志的幻觉，则其每一个行动背后都有其意图，通过指出其行动的不恰当，会让机器人思考其意图、反思其错误，并设想替代的行动可能会有怎样的结果，再根据对于结果的效用（utility）与偏好来调整行为选择，进一步优化其行动。且仅当机器人有能力做其他选择，即进行反事实思考时，才能够有效地谈论如下反事实概念：信誉（credit）、责备（blame）、责任（responsibility）和遗憾（regret）。对于一个具备因果理解和智能体能力的强人工智能的实现，珀尔是笃定的，这个笃定建立在他对于因果之梯三个层级问题的解决之上，只待在工程上来贯彻和实现：

> 推理自己的信念、意图和欲望的能力，一直是人工领域研究者们的一个重大挑战，同时也定义了"智能体"的概念。机器可以有意识吗？是什么使软件智能体区别于普通的程序？诸如此类的问题已经困扰了几代科学的精英。（[10]，p.339）

珀尔并不认为自己已经完全回答了这个问题。但是他相信，他所提出的反事实算法化，对于理解上述问题，以及使得意识和智能体成为一个可计算的现实迈出了一大步。

最后，在他看来，如果要让一个软件包赋予思维机器以智能体的能力，那么这个软件包至少有三个部分组成："一个关于世界的模型，一个关于软件自身的模型，哪怕只是对于软件蓝图的一个概述，以及一个内存，来记录心灵中的意图如何与外部世界中的事件相对应"。（[10]，p.340）通过内存所记录的数据，就能追索决策的过程，来解释和修正自己的行为。珀尔似乎认为装备因果推断引擎的机器人就具有了意识，但推理能力是否等价于主观意识依然是一个有待商榷的问题。

就算机器人作为智能体成为可能，那么它是一个道德主体（moral agent）吗？泰勒（P. Taylor）指出作为一个道德主体最为重要的五个能力（abilities）："形成关于对与错的判断能力；进行道德考量的能力，即考虑和权衡各种支持和反对可供选择行为的道德理由；根据这些理由做出决策的能力；行使必要的决心和意志力来执行这些决策的能力；以及因自己未能执行这些决策而进行解释的能力。"（[16]，p.14）即要明辨善恶对错，就要对自己的行为有所反思，对各种可能的选择有所考量，对选择的后果有所理解，且在最终的决策还要受道德规范或者社会规范的影响。既然机器具有了反事实推断的能力，上述五种能力就在算法上实现了，那么它作为一个道德主体似乎也是可以的。

四、人工智能因果推断引擎与儿童教育

朱松纯指出“搞视觉认知的，要理解物体功能就要追溯到石器时代去，搞语言的要追溯到语言起源”，[17]同样地，要实现人工智能得从娃娃抓起。通用人工智能的目标是要创造“有自主的感知、认知、决策、学习、执行和社会协作能力，符合人类情感、伦理与道德观念的通用智能体（Agents）”。[18]如果这样一个通用智能体，像人工智能的设计者所设想的那样，是从一个孩子智能体开始慢慢培育、教训而最终实现，即是通过对其进行科学教育与道德教育而最终实现，那么这与儿童教育其实是类似的，因为要培养一个“有决策能力、有道德情感的智能体”和要培养一个“有决策能力、有道德情感的孩子”在理论上并没有差别。

而今关于科学教育与人工智能的大多数讨论都只是集中在把人工智能技术应用到科学教育中。实际上，人工智能的儿童径路和儿童教育之间有其紧密联系。首先，如何教育一个孩子，使其品学兼优、守道德且懂道德，也许可以启发我们来构建一个有道德的智能体；其次，科学家在搭建通用智能体过程中的思考，对于儿童教育本身也同样有其借鉴意义。

1. 从人工智能因果推断引擎到儿童教育

在人工智能通用智能体的搭建可以看到：

首先，智育和德育其实不可分。在大部分儿童教育文献中，科学教育和道德教育都是两分的。但在人工智能学家的理解中，好的科学教育和道德教育都统一于拥有复杂逻辑论证和因果推断能力。在珀尔的人工智能构建中，反事实推理和因果推断本身除了在科学探究中发挥作用，同样也使一个能明辨善恶的道德主体成为可能的必要条件。不管是在对与错的判定中，还是在科学探究中，智能体都是根据所观察到的事实，并严格按照事件之间的因果和逻辑联系来进行推理和思考。通过对众多可能选择的结果与偏好进行理性考量，而在各个选择之间的考虑受到各种决策函数和价值函数的驱动。这其实和经济学中“理性人”假设一致，即一个人的行为和决策都是由利益和价值驱动，并追求自身利益的最大化。总之，不管是科学推理还是道德考量，都要依据事实并遵循理性。实际上，哪怕是确立与现实对照的反事实情形的构建，都建立在理性和逻辑之上，而不是被纯粹的主观臆想以及无端的情绪所决定。

其次，科学教育不能只是灌输科学事实，道德教育也不能只是一种规训教育。人的思维活动与行为决策受到两大因素（模型、表达）驱动，[18]*U*系统和*V*系统，其中*U*系统“是在长期的社会实践、规则演化、多次博弈等过程中所最终形成的一种稳定的社会规制性力量”，人在“江湖”，“固有所不得已”，社会的伦理和规范，以及客观物理定律，无时无刻不“掣肘”着每一个身处其中的“社会人”的决策；而*V*系统则是“人在进化的过程中受到环境、生物属性、激素等共同的影响进化出了价值体系”，它是很多维度的价值函数的一个集合，比如，生物进化、文化和宗教以及个体价值维度。而且随着社会的发展，装有规范和美德的袋子会越发膨胀，且并不是恒定不变的，“人群在竞争合作之中，受到外部物理环境与因果限制下，达成的暂时的准平衡态”，但随着社会的演进，旧的平衡会被打破，新的平衡又会建立。如何把上述错综复杂、甚至有的时候还互相抵牾的*U*系统和*V*系统编码到机器人的决策之中，指导其在面对具体的情形是如何加权，实际上非常不具有可操作性。

阿西莫夫（I. Asimov）给出机器人学的三大定律，以试图培养道德智能体。这三大定律和基于-规则的专家系统思路一致，都有其不可避免的问题。首先，定律本身也许并不总是铁板一块，还是应该所有权衡，这是我们行于世界的生活实践的启示；其次，完全可以构想符合上述三个定律，但还是认为机器人不应该有所行动，或者认为机器人应该行动，但是却违反了这三个定律中某一个和某两个。实际上，阿西莫夫本人在他提出三大定律的科幻小说《我，机器人》中已经指出上述三个定律会让机器人陷入“无所措其手足”的困境中。

正如瓦拉赫（W. Wallach）和艾伦（C. Allen）所说，阿西莫夫的思路是自上而下的，[19]而珀尔的思路是一种自上而下又自下而上的混

合型径路，首先，珀尔把规范编码进机器人的程序里，纳入到其理性的思考之中：

> 一旦机器被编码了自我意识，随之而来它就会有“同情心”和“公平感”，因为这两种情感的计算原则或者算法过程是一样的，只不过是把另一个智能体添加到方程里去罢了。（[10]，pp.342–343）

但道德教育本身又不完全是一种规训教育，“道德教育并不是对于做正当事情的教育，而是指道德思维的性质和作出决定的技巧和能力方面的教育”，[20]即赋予机器人以自由意志的幻觉，让其具有做因果推理的能力，使其能够在于人的交流中理解对错，修正其行为。珀尔主张，明辨是非的道德主体的实现需要从点滴的规训开始。鉴于规则的汗漫，知识的无涯，归根到底还是要赋予机器人一种思辨力和明辨力，当面对纷繁复杂的情形，它才能自我修正、自我进化、自我提升。儿童的科学教育和道德教育同样也“人同此心，心同此理”。

2. 从儿童教育到人工智能因果推断引擎

首先，如前所述，孩子在很小的时候就具有因果推断能力，且很早就知道跟人进行配合，所以，如果要设计一个机器人，那么它必须有推断别人想法和反应的能力，否则它就不能理解何为伦理与道德。机器人首先要能够想象诸多行动的结果，并理解好与坏的标准，从而可以判定这些结果的好与坏，以此来最终决定自己的行动，那么才有所谓道德与伦理。

其次，正如不可能一下子把所有知识教给孩子，同样也不可能给智能体编码关于这个世界的所有知识，我们需要它可以习得知识，实现知识的更新与迭代。根本上，儿童科学教育是为了让与经验和常识相距甚远的科学概念和原理为孩子所理解，增强他们对科学与技术的认识，以激发其学科兴趣与创新潜能。科学教育本质上不是“科学知识教育”，而是科学思想教育，比如，讲授元素周期律，关键并不在于要让学生们记忆背诵整个元素周期表，而是通过历史的梳理让学生理解发现这些元素的曲折过程，揭示其所蕴涵的思想，并通过对原子量以及其他元素性质之间的联系，让学生们理解化学元素的本质。理解本身又涉及到“理解力”的问题，当面对一个逻辑混乱，思维全然没有章法的学生，就算是再生动活泼的教育都只会事倍功半。逻辑的技巧和因果推断的思维看似简单，就像拳法的套路一下平淡无奇，但要达到“人拳合一”，即逻辑缜密、思路清晰且专注持久，需要长期的逻辑教育和数学训练。计算机在很多能力（abilities）方面远超人类，是因为它在行动中严格地贯彻算法。

再次，在儿童的科学教育中，有对于儿童科学实践的很多描述，[21]通过小孩子们所展现出来的智能水平，会发现推理能力只是智能的一个方面，还有归纳、类比、描述、联想，抽象，系统化等等，问题的提出，模型的搭建，观察与思考，尝试与创造，举一反三，这些的实现需要人工智能多个领域的突破，比如，计算机视觉问题，毕竟人的大脑皮层，70%都是在处理视觉信息。再回看因果推断引擎，会发现通用人工智能的实现还有很远的路要走，只是在因果推理方面设想出了一个解决蓝图。而且在引擎中，反事实的算法化其实和真正人的反事实推理也很不一样，毕竟人在推断反事实的时候，是不需要确定变量之间的函数关系。

最后，就算是给孩子最好的和最正大光明的教育，还是有可能最终只培养出一个不辨善恶的“坏人”。就算能够赋予机器人以自由意志的幻觉，赋予它因果和反事实推断能力，似乎并不存在一个绝对的教育过程，可以最终让它成为一个明辨善恶的道德主体。毕竟恶的诞生总让人猝不及防，比如，电影“终结者”中那个一心统治人类的天网（Skynet），设计的初衷显然并不希望它有一天来反噬其创造者。

［参 考 文 献］

[1] Turing, A. 'Computing Machinery and Intelligence'[J]. *Mind,* 1950, 59(236): 433–460.

[2] Piaget, J. *The Child's Conception of Physical Causality*[M]. New York: Harcourt, Brace, 1930.

[3] Hickling, A., Wellman, H. 'The Emergence of Children's Causal Explanations and Theories: Evidence from Everyday Conversation'[J]. *Developmental Psychology*, 2001, 37(5): 668–683.

[4] Gopnik, A., et al. 'A Theory of Causal Learning in Children: Causal Maps and Bayes Nets'[J]. *Psychological Review*, 2004, 111(1): 3–32.

[5] 梅剑华. 人工智能与因果推断——兼论奇点问题[J]. 哲学研究, 2019, (6): 86–95.

[6] Pearl, J. 'The Seven Tools of Causal Inference, with Refections on Machine Learning'[J]. *Communications of the ACM*, 2019, 62(3): 54–60.

[7] 张江. 再论强制阐释[J]. 中国社会科学, 2021, (2): 4–23; 204.

[8] 叶峰. 论语言在认知中的作用[J]. 世界哲学, 2016, (5): 72–82; 161.

[9] Rubin, D. 'Causal Inference Using Potential Outcomes: Design, Modeling, Decisions'[J]. *Journal of the American Statistical Association*, 2005, 100(469): 322–331.

[10] 珀尔、朱迪亚、麦肯齐、达纳. 为什么: 关于因果关系的新科学[M]. 江生、于华 译, 北京: 中信出版社, 2019.

[11] Papineau, D. 'The Causal Structure of Reality'[J]. A Long (double–length) Version. *Monist*. forthcoming, 2022, 1–51.

[12] Lewis, D. 'Causation'[J]. *The Journal of Philosophy*, 1973, 70(17): 556–567.

[13] Pearl, J. *Causality: Models, Reasoning, and Inference*[M]. New York: Cambridge University Press, 2009.

[14] Halpern, J., Hitchcock, C. 'Graded Causation and Defaults'[J]. *The British Journal for the Philosophy of Science*, 2015, 66(2): 413–457.

[15] Imbens, G. 'Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics'[J]. *Journal of Economic Literature* , 2020, 58(4): 1129–1179.

[16] Taylor, P. *Respect for Nature: A Theory of Environmental Ethics*[M]. Princeton: Princeton University Press, 2011.

[17] 朱松纯. 人工智能: 现状, 任务, 构架与统一[J]. 视觉求索, 2017, 1–68.

[18] 朱松纯. 三读《赤壁赋》, 并从人工智能的角度解读“心”与“理”的平衡[OL]. https://mp.weixin.qq.com/s/MuOjSBeWcmmc1t9lQC6nsA. 2022–01–06.

[19] Wallach, W., Allen, C. *Moral Machines: Teaching Robots Right From Wrong*[M]. Oxford: Oxford University Press, 2008.

[20] 霍尔、罗伯特、戴维斯、约翰. 道德教育的理论与实践[M]. 陆有铨、魏贤超 译, 杭州: 浙江教育出版社, 2003.

[21] 李雁冰. 从皮亚杰的早期科学作品看中小学科学教育[J]. 教育发展研究, 2010, 30 (20): 68–72.

[责任编辑 李斌]