大数据时代社会科学方法论探讨*

张庆熊

摘　要：大数据要求人们改变对因果关系的追问，转而追求相关关系；要求人们改变对精确性的苛求，转而追求混杂性；行业专家及其专业知识的重要性都会因为统计学家和数据分析的出现而变低。这些说法虽然刻画了大数据时代的新特点和动向，但论证不够深入全面，可能会引起误解。其中，第一个论点涉及因果关系与相关关系，第二个论点涉及决定论和概率论，第三个论点涉及统计分析和意义理解。从哲学理路上考察上述论点，我们会发现，尽管大数据时代开辟了一条模糊地利用数据的途径，但如果没有通过理想型的理论对大数据中的相关关系的意义的理解，我们就不知道如何去应用这些相关关系。如果我们不考虑社会理论的价值观念和人生指导意义，而沉湎于预测和操控，就会存在被彻底物化的危险。数据库再大，也是依据已经积累的过去的资料来预测将来，而将来是开放的，所以决策和预测总是存在风险，因此，机器的决策永远不能够取代人的决策。

关键词：大数据；因果关系；相关关系；统计分析；意义理解

中图分类号：C0；TP311.13　文献标识码：A　文章编号：0257－5833(2018)09－0069－09

DOI:10.13644/j.cnki.cn31-1112.2018.09.008

作者简介：张庆熊，复旦大学哲学学院教授　（上海　200433）

一、引　言

要讨论大数据时代的社会科学方法论的新特点，让我们从一个例子出发，比如说调研一个二三线城市的房地产开发是否存在过剩的问题。以往社会研究主要采取抽样调查结合统计分析的方法。为了取得数据，我们制定抽样调查的表格，发给那里的房地产开发商、房产中介商和进行房屋交易的居民等，我们让他们填写表格，并结合实地采访。调查的内容包括房屋的价格、建造和出售的时间、空置的比例等。调研的成功取决于样本发放和收回的数量、样本设计、发放对象和统计分析的合理性。这样的调查方法具有局限性。毕竟，样本再多也不等于全体；此外，填表者和受访者由于利益关系或其他的原因可能不愿意说真话。我们希望通过表格设计、样本发放和分析的合理性来弥补这些缺陷，但这不免带有原始数据的不可靠性和主观估算方面的瓶颈。

收稿日期：2018－05－28

＊　本文系教育部哲学社会科学研究重大课题攻关项目“当代国外社会科学方法论新形态及中国化研究”（项目编号：17JZD041）的阶段性成果。

在大数据时代，这样的调研方法即便不算过时的话，至少也显得少慢差费。实际上，如今许多数据不用问卷调查就可以在具有数据记录的相关网站获得，而且这样的数据是客观真实、全面和动态的。这是因为在互联网时代，大量数据被自动记录下来，即便缺少某一方面的数据，也可以通过相关数据加以印证。例如，某个城市房地产的数据不一定要从房地产部门直接获得，也可以从与房地产相关的部门获得，如从具有水、电、燃气等统计数据的部门获得。现代城市中的房屋要有水、电、燃气三通。要查看这个城市房地产的开发情况，可以查看这个城市一段时期以来安装水表、电表、燃气表的数量，以及水、电、燃气交费的情况。如果安装水表、电表、燃气表的数量多，开通和交费的用户少，说明这里的房屋积压多，空置比例高。我们有了水、电、燃气的数据，就可以从水、电、燃气的交费情况中发现线索。比如，有多少房屋是用来居住的，有多少房屋是用来投资的，然后进行有针对性的问卷调查。

想一想，我们在平时购物和支付的经济活动中，在社交媒体的文字书写和转发中，甚至在打字和语音输入中，不知留下了多少信息，它们在电脑的服务器中被储存下来，通过数据处理，能够被用作各种各样的宏观和微观的统计分析。商家会利用你网上购物成交的记录和浏览商品的记录，让电脑自动估算出你的购物倾向和习惯，向你推荐商品。过去有一种说法：人在做，天在看。在大数据时代，“天眼”就是数据网络，人好像无时不刻不处于无形眼睛的监察。我们在不知不觉中留下数据，而这些网络数据可以在你毫不知情的情况下被加以利用。

本文不是想具体说明大数据的技术应用，而是想从理论上探讨大数据技术应用对社会科学方法论研究的意义。前者是技术的问题，后者是社会科学哲学的理论问题。畅销书《大数据时代》[①]非常生动地描述了大数据的技术应用和由此带来的重大变革，其中谈到：“当数据处理技术已经发生了翻天覆地的变化时，在大数据时代进行抽样分析就像在汽车时代骑马一样。一切都改变了，我们需要的是所有的数据，‘样本＝总体’。”[②]从社会科学方法论的角度看，该书除了以上论点外还有如下三个论点值得我们关注和反思：(1)发现关联物，找到相关关系，是预测的关键；知道“是什么”就够了，没有必要知道“为什么”；在大数据时代，我们不必非得知道现象背后的原因，而是让数据自己“发声”。[③] (2)执迷于精确性是信息缺乏时代和模拟时代的产物，依靠大数据的统计概率，接受不精确性，我们才能打开一扇从未涉足的世界的窗户。[④] (3)行业专家和技术专家的光芒，都会因为统计学家和数据分析学家的出现而暗淡，因为后者不受旧观念的影响，能够聆听数据发出的声音。在大数据时代，专业知识以及对这些知识的理解变得不重要了，很多工作可以由统计学家和数据分析学家来做。[⑤]

以上，第一个论点涉及因果关系与相关关系的哲学问题，第二个论点涉及决定论和概率论的哲学问题，第三个论点涉及统计分析和意义理解的哲学问题。尽管《大数据时代》一书通过许多生动的例子刻画了与上述论点相关的大数据时代的思想方法的特征，但我觉得其论证不够深入全面，可能会引申出似是而非的结论。本文试从社会科学哲学方法论角度考察上述论点，以期把思想方法上的问题说得更透彻一些。

二、因果关系与相关关系

因果关系与相关关系是哲学上的老问题。休谟主张，我们关于因果关系的知识，“产生于当我

① 同类的书很多，例如，吴军：《智能时代：大数据与智能革命重新定义未来》，中信出版集团股份有限公司2016年版；李开复、王咏刚：《人工智能：李开复谈AI如何重塑个人、商业与社会的未来图谱》，文化发展出版社2017年版；伊恩·艾瑞斯：《大数据思维与决策》，人民邮电出版社2014年版。我的引证选择《大数据时代》这本书，因为它较为集中地描述了“大数据时代的思维变革”，从而引发我考虑与其相关的社会科学方法论的问题。

② ［英］维克尔·迈尔－舍恩伯格、肯尼思·库克耶：《大数据时代》，浙江人民出版社2013年版，第27页。

③ ［英］维克尔·迈尔－舍恩伯格、肯尼思·库克耶：《大数据时代》，浙江人民出版社2013年版，第2、67－94页。

④ ［英］维克尔·迈尔－舍恩伯格、肯尼思·库克耶：《大数据时代》，浙江人民出版社2013年版，第2、45－66页。

⑤ ［英］维克尔·迈尔－舍恩伯格、肯尼思·库克耶：《大数据时代》，浙江人民出版社2013年版，第180页。

们看到一切特殊的对象恒常地彼此联结在一起的那种经验”①。如果人们发现有两组现象,后一组总是跟着前一组发生,通过心理联想,我们认为,前一组是后一组的原因。两组现象之间存在相关关系是“事实”;认为前一组是后一组的原因,是我们对“为什么”的解释。世间的因果关系或许根本不存在,所存在的只是现象间的相关关系。休谟的观点是否合理,历来争议很多,其关键取决于我们对什么是原因的意义的理解。事实上,原因这个词语是多义的。亚里士多德提出“四因说”,用以回答“为什么”的问题。一个大理石雕像为什么会产生出来?它是由雕刻家的意图(目的因)、所采用的大理石(质料因)、所设计的形象(形式因)和所施加的雕刻力量(动力因)结合起来产生的。

现代科学寻找事物发生的规律,现在我们也常把规律解说为原因。当问:苹果为什么会掉在地上?回答:因为万有引力。在此,万有引力的规律被理解为原因。但规律是不是应该被理解为原因?这是一个值得商榷的问题。按照《大数据时代》作者的看法,一组数据不是另一组数据的原因,这里存在相关关系,它不等同于因果关系。我觉得这一看法有合理之处。举例来说,我们不能把水、电、燃气的数据理解为房产景气指数的原因,也不能把房产景气指数理解为水、电、燃气的数据的原因。它们只是两组数据之间的相关关系。我们发现,在一定条件下,一个城市的水、电、燃气的数据可以用作反映该地房产景气指数的参考数据,因为它们之间存在相关关系,但这种相关关系不等同于因果关系。由此看来,休谟的观点似乎在大数据时代得到印证。但因果关系的思考方式是不是应该被抛弃呢?问题可能没有那么简单。因为人类仍然要询问:数据为什么会产生出来呢?怎样去理解数据间的相关关系呢?这里还是要寻找因果关系。计算机可以发现两组数据之间的相关关系,但计算机还不能理解为什么会发生两组数据之间的相关关系。回答为什么的问题,还是要借助于因果关系的解释。如果我们把投资买房看作原因,就能理解为什么这些房子不是用来住,由此造成空置房多而水、电、燃气用量少的相关关系的数据。

找到两组现象之间的相关关系,就是找到了规律。规律可以加以量化,科学追求对规律的量化。近代科学用公式表示规律,这是近代科学区别于古代科学的显著特点。用数学公式表示规律,建立在发现两组现象之间存在数量关系的基础之上。从因果关系这个词的严格用法来看,把规律解释为原因,似乎有不妥之处。我们谈到因果关系的时候,通常说前因后果,但规律本身没有时间上的先后关系。我们通常认为,原因与结果是两个实体性的东西或两个发生的事件之间的关系,一个在前,另一个在后,一个导致另一个的发生,但规律并不是实体性的东西或发生的事件。那么规律是不是绝对不可以被理解为原因呢?这取决于对“原因”这个词的意义的理解。正如亚里士多德的“四因说”把“形式因”解说为原因一样,把“规律”解说为原因在一定语境下是可以理解的,在日常用语中已经成为习惯用法。

随着近代自然科学的发展,发现相关关系并用公式表达其规律成为科学探究的热衷目标。发现相关关系被认为是科学的事情,仅仅给出动机之类的原因被认为还停留在前科学的水准。自然规律被认为是必然的规律,是不能用动机之类的原因来说明的。如何才能把研究自然科学的发现规律的方法用于研究社会科学呢?显然,人的行为是有动机的,人被认为有自由意志,并可以进行自由选择。鉴于社会是由人组成的,自然科学的发现规律的方法似乎很难用于社会科学。但是能不能发现人的行为方式之间的相关关系呢?一定社会层次和团体中的人有没有规律性的行为倾向呢?这些行为倾向受什么因素制约呢?如果能找到这种相关关系的数据,对社会现象进行量化的研究也非不可能,而大数据时代为这种量化研究提供了可能性。人的行为留下大量的信息数据,这些信息数据能被计算机和网络收集和综合起来,从而能发现人的行为倾向之间的相关关系。例如,电商发现某一客户买了奶粉,就在网页上向她/他推荐尿布、婴儿车等,因为电脑记录能发现相当多的用户有这类连贯性的倾向,尽管电脑并不知道该客户是孕妇还是其他什么人。

大数据时代营造了这样一种可能性,计算机的设备能自动记录和发现人的行为间的这些数据

① [英]休谟:《人类理智研究》,商务印书馆2009年版,第21页。

关联，商家能利用这些数据为他们获得利益，某些机构能利用这些数据对社会中的人进行操控。至于造成这些数据的原因是什么，他们常常是不关心的。这就是《大数据时代》一书所说的，发现关联物，找到相关关系，是预测的关键；知道“是什么”就够了，没有必要知道“为什么”。然而，让我们比较一下，这对社会科学理论会造成什么样的影响？早先的社会科学理论大多基于因果关系来解释相关关系。拿迪尔凯姆的“自杀论”来说，他发现社会的整合程度与自杀率存在一定的相关关系，因为社会生活既意味着个人有一定的个性，又意味着个人准备放弃这种个性。一个社会的整合程度过于松散，个人缺乏社会群体对他的关心，自杀率会高一些。反之，一个社会的整合程度过于严密，个人承受不了社会群体对他的监控和指责，自杀率也会高一些。由于有了这种原因说明，迪尔凯姆的社会理论包含着寻求社会合理化的意向，去营造一种个人自由程度和社会整合程度处于“均衡状态”的社会结构。① 再如，生产力与生产关系之间存在相关关系，资本倾向于追求利益的最大化，资本的剥削是工人阶级受奴役的原因，工人阶级为获得解放，必须改革社会制度。倘若在大数据时代，由于计算机在从事发现和运用相关关系方面的重大作用，而看不到或故意抹杀社会理论在阐明因果关系中的价值和意义的维度，就会走上偏路。早先的社会理论不是价值中立的。一种社会理论一旦产生，这不是纯粹对相关关系的说明，而会对人的思想产生影响，对社会改造产生影响。如果认为发现和应用相关关系都是机器的事情，人就成了机器支配下的物品了。

我在此并非批判《大数据时代》这本书，我认为该书在谈到有关“没有必要知道为什么”这个论点时，主要是在着力刻画大数据时代的现象，而且说的还比较谨慎，它在这一章的结尾处补充说：“大数据绝不会叫嚣‘理论已死’，但它毫无疑问会从根本上改变我们理解世界的方式。很多旧有的习惯将被颠覆，很多旧有的制度将面临挑战。”②确实，在大数据时代，很多发现相关关系的工作和提出行动建议的工作，已由计算机来承担。在这种情况下，如果我们不考虑社会理论的价值观念和人生指导意义，不去追问为什么的问题，沉湎于预测和操控，就会存在被彻底物化的危险。

三、决定论与概率

近代科学理论大都按照决定论模式建立。决定论主张，自然界中的万事万物按照必然规律运行，自然规律以全称命题方式表达。比如，所有物体之间都存在万有引力。既然是必然的，那就不存在反例；对于一个全称命题，只要有一个反例，它就不能成立。如果一个科学家发现一个按照决定论模式建立的科学理论中的反例，那么他就可以否定这个科学理论。波普尔的证伪主义的科学纲领由此建立，他把能不能被证伪作为区分科学与非科学的标准。

能不能把自然科学中的这种决定论模式应用到社会科学中来呢？波普尔本人是否定的。他认为社会现象不是必然的，历史没有决定性的规律可循，“相信历史命运纯粹是迷信，对于人类历史的过程不可能以科学或任何其它理性的方法加以预言”③。但是还有一些社会科学家主张，自然科学的模式通过一定改造可以应用于社会科学。其思路大致有两种：(1)概率论，(2)理想型。按照概率论的思路，即使自然现象也不全是决定论的，微观世界中的粒子，宏观世界中的气象，都不能按照决定论的模式来建立理论；只要能够发现其中的一些概率，达到预测或预报的目的就足够了。社会现象极其复杂，但不等于没有概率性的相关关系可寻。在人的经济利益、社会层次、思想方式和行动倾向之间，存在一定的概率的相关性。社会科学的任务是发现它们之间的相关性的概率关系。按照理想型的思路，即便自然科学也要考虑理想条件和理想类型。例如，一个物体在没有阻力的理想条件下是匀速运动的，尽管在现实世界中不存在没有阻力的状况；世上物体的形状千差万别，但几何学所研究的是理想的点、线、面和图形之间的必然关系。这就是说，科学理论的理想模型是可以

① 参见[法]迪尔凯姆《自杀论》，冯韵文译，商务印书馆1996年版，第346页。

② [英]维克尔·迈尔－舍恩伯格、肯尼思·库克耶：《大数据时代》，浙江人民出版社2013年版，第94页。

③ Karl Popper, *The Poverty of Historicism*, London: Routledge, 1960, p. viii.

按照必然的、决定性的关系来建立的，但这种理想模型并不等于现实世界中的实际情况。自然科学以自然规律的必然性为预设，社会科学鉴于人的主观能动性不做这样的预设，但社会科学可以把现实的人加以典型化，进行理想分类，建立它们之间的理想型的相关关系。社会科学中理想模型可以作为现实的社会研究的参考。这就意味着，在一定意义上，社会科学中的理想模型如同物理学中的绝对真空的理想条件下物体运动规律和几何学中的理想图形间的关系一样。

英国实证主义哲学家穆勒首先把概率论的方法用到社会研究中来，他写道："一旦我们知道人类中的大多数、某个民族、某个阶级中的大多数将如何思考、感受和行动时，就足以说这样的命题相当于普遍的命题了。对于政治和社会的科学来说，这就足够了。正如我们早先已经指出的，在社会的探索中的一个近似的概括，对于大多数实践的目的来说相当于一个精确的概念；其之可能，仅当所要加以断言的个人被无偏见地选择时，仅当所确定的是大众的性格和集体的行为时。"①后来的实证主义的社会学家基本上都是沿着穆勒的这条思路进行的，至多在细节上做一些改进。

理想型的研究方法则要归功于韦伯。韦伯社会理论的基本思路是康德式的：一方面主张人有自由意志，人的行为受动机支配；另一方面又主张，人的社会行动是可以划分类型的，人的有意志的行为受人的行为模式的制约，类似于康德所说的人的认知方式受到人的认知范畴的制约。这意味，虽然人的行为不是处于决定论的规律的支配之下，但依然有线索可循。韦伯认为，人的行为与动机之间关系可以分为不同的类型；如果确定了这些类型，就有线索可循了。韦伯区分了如下四种类型：(1)目的合乎理性的，即通过对外界事物的情况和其他人的举止的期待，并利用这种期待作为"条件"或者作为"手段"，以期实现合乎理性所争取和考虑的作为成果的目的；(2)价值合乎理性的，即按照伦理的、美学的、宗教的或作其它某种阐释的价值规范采取行动，不管是否取得成就；(3)情绪的或情感的，即所采取的行动受到自己的情绪或情感的强烈影响；(4)传统的，即按照约定俗成的习惯采取行动。② 实际上，人的社会行为往往是混杂的，很难见到这四种纯粹的类型。有的人可能一时情绪激动，采取了某种行为，但很快平静下来，按照合理的方式进行补救。有的社会行为是目的合乎理性、价值合乎理性和传统习惯综合的结果。但韦伯进行这样的分类，有助于按照人的社会行为的典型的类型来进行推断，对此加以解释和理解。

孔子说："学而时习之，不亦说乎！"如果我们问一个正在被家长和教师逼迫下读书的孩子，他说出的真实感受可能与孔子的相反：学而时习之，不亦苦乎。因此，我们不妨把孔子说的"学而时习之，不亦说乎"理解为一种理想型。孔子还说："其为人也孝弟，而好犯上者，鲜矣；不好犯上，而好作乱者，未之有也。"这前一句是高概率的判断，这后一句是全称判断。"孝弟"是一种行为类型。属于"孝弟"行为类型的人，"好犯上"的社会行为是很少见的。"不好犯上"也是一种行为类型。属于"不好犯上"行为类型的人，造反作乱的社会行为是没有的。孔子虽然没有提出理想型的社会理论，但他的思路，也是按照理想型来思考的。他划分人的行为倾向的类型，找出它们之间的相关关系，由此确立儒家的"君子务本"，而"孝弟"是"仁之本"的社会政策和政治目标。

现在我们来看大数据时代社会研究方法的特点。有人认为，大数据为决定论的社会研究方法提供了可能性，因为数据全了，依据可靠了，结论就确定了。有人进而主张，大数据为计划经济开创了真正的可能性。以前的计划经济弊端多，是因为数据不全；现在能依据全面的数据来安排经济，计划经济的前提具备了。我觉得以上观点似是而非，是出于对大数据时代数据采集和运用的误解而得出的结论，其理由如下：

(1)社会生活是动态的，大数据时代所收集的数据是动态的数据。我们用已经收集到的数据去预测将来，这是用过去预测将来；即使过去的数据再全，也不能完全预测将来的情况。数据全了，依据就可靠了，这一说法只是基于有限的或必然的情况才能成立。在有限的情况下，当我们收集到全

① John Stuart Mill, *System of Logic*, London 1895, p. 554.

② 参见[德]韦伯《经济与社会》(上卷)，林荣远译，商务印书馆1998年版，第56页。

部数据,根据完全的归纳法,确实可以得出完全可靠的结论。决定论的前提是物体运动规律的必然性,但社会生活中的人的行为没有这种必然性。社会生活总是变化着的,是向着未来开放的,这不是一个封闭的系统,因此已有的数据再全也没有用,完全归纳法和决定论的研究模式在此行不通。在社会研究中基于过去预测将来,只能是概率性的。

(2)在大数据时代,企业能够根据网上的预订来安排生产计划。比如,一家企业发布新款手机,它根据网上的预定情况,来安排生产数量。从企业的角度来看,这是计划经济。但这种计划经济,是企业依据市场的供求关系的计划经济。在此,数据从市场来,再用于市场的销售;企业能依据这样的统计数据尽可能避免盲目生产。但这种企业的计划经济与国家指令式的计划经济不是一回事情。国家指令式的计划经济是在非市场或弱市场前提下的计划经济,是不考虑或几乎不考虑市场需求变化的计划经济。同样,国家能根据互联网所提供的大数据进行宏观调控。这是对市场经济的宏观调控,与指令式的计划经济是不同的。

《大数据时代》的作者断言,执迷于精确性是信息缺乏时代和模拟时代的产物,只有依靠大数据的统计概率,接受不精确性,我们才能打开一扇从未涉足的世界的窗户。我觉得这句话只说对了一半。首先,依靠统计概率,接受不精确性,在穆勒的时代,即在 19 世纪,就已经在社会研究中流行了,尽管那时还缺乏大规模收集数据和处理数据的有效手段。这里所说的"模拟时代",在我看来,指设计理想型和模拟理想型的研究方法流行的时代。理想型本身是精确的,按照理想型去设计表格,进行调查和统计数据的流程也可以做到精确化;理想型的现实运用是对理想型的模拟,这里就可能存在偏差和误用了,甚至会出现所设计的理想型是异想天开,根本不适用。在大数据时代,大量的信息以混杂的方式涌现出来,如果我们追求精确性,就无法利用这些数据。但我们可以通过模糊算法的计算机程序,对这些信息进行模糊的归类和整理,从中能找到具有概率性的不同类型的数据间的相关关系,为我们利用这些数据提供可能性。而且,大数据间的相关关系具有这样的特点,尽管数据是模糊的,随着数据收集得越多,正确性的概率就越高。在这一意义上,我觉得《大数据时代》作者的说法是可取的。但我们如何才能理解大数据间的相关关系呢?这里依然要靠理想型。没有理想型,大数据的模糊算法的程序设计不出来,大数据间的相关关系无法被我们所理解和运用。人的认识要通过概念来把握,概念在一定程度上可以说是理想型。有人说,概念之间只有"家族相似",家族相似是模糊的。此话不错,但如果没有理想型,你怎么知道它们之间相似呢?只有树立一个样板,你才知道它们与样板之间的差别和相似。这里,精确性和模糊性是一组相对的范畴,不知道精确也就不知道在多大程度上是模糊的,反之亦然。再说,没有定性,你就无法定量,量是在定性的分类基础上的量。定性要借助理想型,而确立理想型,对于构建理论很重要。尽管理想型的理论在现实的社会研究中只有参照意义,但离开了理想型的理论,就难以理解事物和数据的意义。这就是说,尽管大数据时代开辟了一条模糊地利用数据的途径,但如果没有通过理想型的理论对大数据中的相关关系的意义的理解,就不知道如何去应用这些相关关系。

四、人工智能与意义理解

目前人工智能主要涉及两个方面:识别与决策。在这两个方面,人工智能都取得了很大的成绩。机器的人脸识别已经做得非常好,现在已经达到比人眼的识别更准确和高效。计算机下围棋已经战胜人类顶尖高手。计算机为什么能够做到这一点呢?我觉得主要在于计算机的程序模仿了人的认知方式,再加上计算机所使用的光电感应技术和信息储存技术要比人的生物的感觉反应和大脑的记忆更加灵敏,并且容量也更大。不过,这后一点不是关键,人早就借助显微镜发现人所看不见的微生物,借助胶卷储存信息,这不是质的飞跃而是量的提升。使我们感到惊讶的是,计算机似乎有了像人一样的识别和决策的能力。为了解答这个问题,让我们比较人和机器各自是采取怎样的方式进行识别和决策的。对于识别来说,存在着正确还是不正确的问题;对于决策来说,存在着成功还是不成功的问题。人工智能正是在这方面找到了模拟人的方法而大大提高了准确性。

让我们先来讨论“识别”中的人工智能问题。例如,我们前面有一个人的两组照片:一组是同一个人的同一张照片的不同比例的放大;另一组是同一个人的不同年龄阶段的照片。对于人而言,这种识别很简单,我们往往一眼就能够看出它们是同一个人的照片。但对于机器来讲就不那么简单,这要经历一个“学习”的过程。机器通过它的光电感应设备获得一系列参数,面对这些参数它存在一系列的选项。如果它的第一个选项是按照相片的大小进行识别,它就错了;如果它按照相片中各要素间的同比例关系进行识别,那么它在第一组相片的识别中就成功了。对于第二组相片,由于每张相片中增加或减少了一些要素,如成年人长出了胡子,老年人白了头发,就要以眼睛、鼻子、嘴巴间的主要要素间的比例关系进行识别,而悬置非主要的要素。这是一个从试错中学习的过程,是一个逐一排除错误选项和保留正确选项的过程。计算机在识别学习的过程中是怎么知道自己犯错误的呢?严格地说,计算机自己并不知道,这是人告诉它的。人给出不同的图像和标准答案,让它在试错中得到训练,让它通过递归的方式最终找到正确的途径。

现在我们来讨论大数据时代人工智能的预测和决策的问题。在大数据时代,人工智能预测和决策的准确性,在很大程度上得益于数据的丰富性。预测是按照相关关系做出的,只要具有相关关系的数据库中的资料足够丰富,预测的准确率就会相当高。谷歌翻译为什么正确率高呢?因为它把世界上很多著名的图书馆中收藏的已经没有版权的书籍及其译本输入到数据库中,再加上通过搜索引擎把网页上的资料及其译文输入到数据库中,建立起海量的语言资料数据库,在此基础上统计具有相同以及大致相同译文的句子,其相同的数量越多,相同率越高,翻译的正确率就越高。过去的机器翻译缺乏这样的条件。在数据库小和语言资料少的情况下,只能建立句型和词典的数据库,按照句型选择相对应的词汇。由于语言的句型很丰富,语词的用法多种多样,机器翻译的正确率就很低。

现阶段的人工智能的决策能力的提高,也得益于大数据。我们来考察博弈类游戏中人工智能是如何进行决策的。如果博弈的规则确定,数量有限,是一个封闭的系统,人工智能在原则上能够穷尽一切算法,这意味着人工智能在原则上能够找到最佳的决策途径,人是无法战胜人工智能的,至多只能平手。这正如计算器计算数学题一样,人不及人工智能快速和准确。围棋在原则上也是一种规则确定的封闭系统,但围棋要计算的量非常大,目前的计算机还不能穷尽围棋的一切算法,特别是在时间限定的比赛中,只能在有限的可能性中做出最佳的选择。人类棋手是如何寻找最佳选择的呢?一是靠直觉,二是靠计算,三是靠定式。就计算而言,计算机要比人类棋手强。如果说人类棋手在1分钟内能算清5步,那么目前的计算机至少有在1分钟内计算清楚8到10步的能力。定式就是已经计算清楚的局部的棋谱,人类棋手借助定式来减少计算量,确定正确的选择。计算机能储存更多的精密计算过的棋谱,而且记忆更加快速和正确。以前的计算机围棋程序在“直觉”方面比人类棋手弱。人类棋手能依靠直觉减少选棋的范围,在很短的时间里找到行棋的大致正确方向。计算机在过去缺乏这种直觉的能力,搜索大量无意义的局面,导致有意义的局面没时间计算或者计算深度不足。这就是为什么过去的计算机围棋程序不能战胜人类职业棋手的原因。如何才能训练计算机在最短的时间内找到最佳的行棋方向呢?计算机所能做的是依靠对弈中生成的棋谱进行学习。假如说有“扩大实地”、“做活”、“联络”、“断”以及它们之间各种兼顾的行棋的方式,人类棋手凭直觉知道当下哪一种选择较好。这是一个手段效率评估的问题,需要考虑的因素很多,而且要结合棋局上的具体情况,不能给出一个固定的评价指数。因此,这对于计算机来讲是一个难题。为了培养计算机的这种“直觉”的能力,我们让两台计算机进行对弈,把在规定时间内找到较好手段获得成功的棋谱记录下来,也把选择了较差手段而导致失败的棋谱记录下来,让计算机在不断试错中学习,最终计算机学会如何在最短的时间内找到最佳的行棋方向。这种计算机的行棋选点方式的学习与计算机人脸图谱识别方式的学习在思路上是一样的,就是一开始让计算机任意形成各种要素间的可能的连接方式,然后通过试错,让它初步淘汰差的选项,从而知道哪一种连接方式

及其组合更好。这需要依靠计算机的深度学习及其算法[①]来完成这一任务。当然,围棋的深度学习要比人脸识别的深度学习更加复杂,其复杂程度在于局势在不断变化,采取的手段要因地制宜。机器学习的优点是,把难以用确定方式表达的复杂逻辑放在数以百万计的多层神经网络系数里,通过海量的大数据把这些系数训练出来。让两台计算机在对弈中进行深度学习,要比计算机学习人类棋谱的方式更加好,因为前者是在相同的设计程序下的学习,有助于按照程序进行较为周全的比对和在试错中进行系统地学习。计算机的这一能力的提高取决于优化的学习方式和训练时间的长短。据说,以前的 Alphago 是靠人类棋谱学习的,而 Alphago Zero 完全不用人类棋谱,通过机器对弈中自身产生的数以千万计的棋谱进行训练,极大地提高了其行棋方向选择能力。这就是为什么 AlphaGo Zero 完胜 AlphaGo Master 的原因。

现在我们讨论人工智能与人的意义理解问题。人的意义理解问题归根到底是人的生活问题,人在生活中理解人生的意义。计算机不直接参与人的生活,不是真正意义上的人的生活世界中的一员,还不能像人那样真正懂得生活的意义。人的生活有价值导向,人能感受到生活中的悲欢离合、酸甜苦辣,因此有情感,有激情,有忧愁。前面谈到韦伯把人的社会行为分为四种:(1)目的合乎理性的,(2)价值合乎理性的,(3)情绪的或情感的,(4)传统的,即按照约定俗成的习惯采取行动。目的合乎理性的社会行为指,在目的已经确定的情况下,如何采取有效的手段去达到目的。在这方面,大数据时代的计算机能够帮助人们去选择有效的手段。但人的目的本身是可选择的,人按照对生活意义的价值的认识来选择目标,按照伦理规范对自己的行为进行道德评价。计算机不能理解这些意义。如果说选择有效的手段去达到目的是工具理性的话,那么计算机所能做的是工具理性范围内的事情。计算机能够在一定程度上模仿人的工具理性。在此有两个关键点:一是对事物的识别;二是采取达到目的的手段。识别有正确与错误,手段有成功与失败。人的认知理性是在试错中找到正确的方向,失败是成功之母。波普尔把"试错法"当作科学理性的基本方法。他写道:"阿米巴和爱因斯坦的区别在于,尽管他(它)们都在使用尝试和排除错误的办法,但阿米巴不喜欢出错,而爱因斯坦却对错误感兴趣:他怀着在发现错误和排除错误的过程中学习、提高的愿望,有意识地寻找自己的错误。科学的方法就是批判的方法。"[②]人工智能的成功,在很大程度上取决于把这种试错的思路用于程序设计中。计算机程序的"如果—那么"的递归的语言,其实就是"试错法"的语言。假如有 A、B、C 三个选项,其中只有一个选项是正确的,计算器的递归程序所能做的就是,如果 A 假则 B,如果 B 假则 C。当然,如果选项非常多,并不那么确定,就要优化选择的算法,并通过概率计算。在规则确定的情况下,计算机能识别什么是正确,什么是错误,什么是失败(输),什么是成功(赢)。但计算机并非真正知道正确和错误、失败和成功的意义。说到底,认知来源于实践,只有生活中才能理解生活的意义。计算机只是辅助地参与了人的生活实践,犹如过去人使用的镰刀、斧头和算盘参与人的生活实践一样。因此,我们不能说计算机真能理解意义。

在大数据时代,人工智能是不是只需要依靠数据统计就能解决预测和决策问题呢?《大数据时代》的作者特别强调海量数据在这方面的作用,认为专业知识和对这些知识的理解变得不重要了,甚至调侃:"在谷歌的机器翻译团队中,这些工程师们都不会说他们翻译出的语言;类似的还有,微软机器翻译部门的统计学家们在茶余饭后的谈资就是说每次一有语言学家离开他们的团队,翻译的质量就会变好一点。"[③]我不想否认在某些情况下会出现这样的事情。语言学家的结构主义的语言学原理曾被设想为机器翻译的基本思路,但这条结构主义的原理在人类语言活动中要通过结合

① 有关机器学习及其算法的方式很多,如:随机森林算法(random forests)、梯度提升算法(gradient boosting)和惩罚线性回归算法(penalized linear regression)等。有关介绍这类算法的书也很多。我自己感到鲍尔斯的《Python 机器学习:预测分析核心算法》(人民邮电出版社 2017 年版)尽可能地用简单的术语来介绍算法,同时通过具体例子让读者很快深入了解模型构建背后的原理,这对于理解机器学习的算法与人的认知方法的关系是有帮助的。

② [英]波普尔:《客观知识》,上海译文出版社 1987 年版,第 75 页。

③ [英]维克尔·迈尔—舍恩伯格、肯尼思·库克耶:《大数据时代》,浙江人民出版社 2013 年版,第 181 页。

"语境"的"转换—生成"才能发挥作用,这在机器翻译中很难用计算机编程的算法实现,毕竟计算机难于把握"语境"。所以,语言学家的诸如此类的语言学见解在实际编程中不得不被束之高阁。目前人工智能机器翻译主要是依靠已有译文的大数据,所以统计学家和数据分析学家在此显得更加有用。但就总体而言,我认为专业知识以及对专业知识的意义的理解,在人工智能中依然必不可缺。首先,选用什么样的数据库对于正确率非常重要。谷歌机器翻译团队选用世界上许多著名图书馆珍藏的大量优秀著作及其优秀译本为数据库的资料,所以翻译质量很高。对于网上搜索到的翻译资料,也需要专家对其翻译质量加以甄别。试想一下,如果不加甄别地拿网上搜索到的中文论文的英文摘要作为翻译的数据库,那么翻译的质量就不会太高,因为这些英文摘要的翻译质量参差不齐,不符合英语习惯的中式翻译很多。甄别哪些英文摘要的翻译质量高,这需要仰仗专家的专业知识。在设计评估效率参数和优化机器学习流程时,也要具备一定的专业知识。总之,人工智能是为人的实践需要服务的,不理解所需解决的问题的意义,不懂得如何应用,就不能设计出满足人的需求的人工智能的程序。人工智能的设计和应用,离不开人对社会生活中的意义的理解。

懂得了这一点,就能懂得为什么人的决策不能完全依赖计算机。拿经济决策来说,在互联网加大数据的条件下,一个终端设备的生产厂商能够及时和准确地知道世界各地零部件供货的情况,并知道哪家工厂生产的零部件质量最好和价格最优,它能优化采购和进行组装,从而获取最大的利润。人工智能的专家能根据这些数据设计出一套程序,为这样的经济决策服务。这一设计的核心理念就是把追求利润的最大化当作经济中的"理想型",从而便于量化的计算。然而,这一程序设计得再好,也不能够防备如下情况,某些掌握核心技术的零部件的生产厂商出于政治或其他原因突然停止供货,这就超出了"理想型"的范围。因此,经济决策不能仅仅依据工具理性的追求利润的最大化,还要有价值导向,企业家要有远大的志向和一定的牺牲精神,宁愿一时利润少也要花大本钱和大力气攻克核心技术。

数据库再大,程序设计得再好,也是依据已经积累的过去的资料来预测将来,而将来是开放的,人类社会是一个开放的系统,因此这样的决策和预测总是存在风险,总是会失灵的。人是有理想的,人在生活中产生新的价值观念,批判过去引领未来。因此,人归根到底不能跟着机器走,机器的决策永远不能够取代人的决策。

(责任编辑:轻 舟)

Discussion on the Methodology of Social Science in the Age of Big Data

Zhang Qingxiong

Abstract: The Age of Big Data illustrates the changes brought about by big data through many vivid cases. As far as the thinking method is concerned, the author believes that big data requires people to change the inquiry into causality and pursue correlation instead; people are asked to change their demanding of accuracy and to pursue confounding; the importance of professional experts and their knowledge can be reduced by the presence of statisticians and data analysts. I think these statements, while charactering new features and trends in the era of big data, are not sufficiently thorough and may cause some misunderstanding. Among them, the first argument involves the causation and correlation; the second argument involves the determinism and the theory of probability; the third argument involves the statistical analysis and understanding of meaning. These are all philosophical questions. My essay tries to examine the above arguments from the perspective of social science methodology, with a view to making these problems more penetrating in philosophical reasoning.

Keywords: Big Data; Causality; Correlation; Statistical Analysis; Understanding of Meaning