大数据主义与大数据经验主义
——兼答黄欣荣教授

齐磊磊
(华南理工大学 科学技术哲学研究中心,广东 广州 510641)

摘 要:大数据经验主义是基于哲学的视角提出的一个具有学术渊源的概念,与目前流行的大数据主义和而不同。大数据经验主义立场显明,认为大数据时代不需要理论;相关性替代了因果性。大数据主义的态度则较为温和,认为理论是处理大数据整个过程中的基础;大数据的相关性为寻找科学规律提供了帮助;相关性是表象,因果性才是事物的本质,获得相关性的目的是为了更好地寻找因果性。

关键词:大数据;大数据主义;大数据经验主义;因果;理论;规律

中图分类号:N02,B017 **文献标识码**:A **文章编号**:1008-7699(2018)02-0016-06

谷歌、IBM、脸谱网等创新公司与互联网、云计算等互动技术,推动整个数字世界进入到大数据时代。作为对大数据时代的一个哲学反思,笔者 2015 年 7 月曾经在《哲学动态》上发表一篇论文"大数据经验主义——如何看待理论、因果与规律"[1],文中首先梳理了从 16、17 世纪的洛克、牛顿、贝克莱、大卫·休谟坚持的经验主义到卡尔纳普、石里克的逻辑经验主义再到以南茜·卡特莱特为首的新经验主义的观点,然后结合大数据的概念与哲学意义上对大数据的理解以及新经验主义的观点,提出了大数据经验主义的概念;在对大数据经验主义的基本观点进行概括后,用较大篇幅对大数据经验主义的观点进行哲学分析。

论文发表以后,国内学界许多读者关注到这篇论文并以不同的方式与笔者交流,比较赞赏的观点主要集中在问题的敏锐度与哲学分析的力度上,当然也有持不同意见者,如黄欣荣教授专门写了一篇文章《大数据如何看待理论、因果与规律——与齐磊磊博士商榷》。[2]在学术研究的过程中,除了学术上的共鸣之外,难能可贵的当然还有学术上的争鸣。认真研读黄教授的商榷文章,争论的焦点汇集为:大数据经验主义与大数据主义是不是一回事?本文主要围绕这个问题展开。

一、大数据经验主义与大数据主义

黄教授在商榷的文章中首先肯定了笔者提出"大数据经验主义"这一概念的意义:"她在文中提出了大数据经验主义的概念,并系统提炼了大数据经验主义的科学哲学观点,这是大数据哲学的重要提炼和概括。"[2]在提出这个看法之后,黄教授话锋一转,开始讨论他的不同观点。黄教授在接下来对不同观点的阐述中第一句话就导致产生了我们商榷的焦点问题。他说:"大数据经验主义是一种新经验主义(以下简称为大数据主义)。"[2]对于这样的一个"简称"方式,可能是无意为之,但是笔者提出的"大数据经验主义"在黄教授的商榷文章中包括题目在内的所有使用之处都被简称为"大数据主义"。

在"大数据经验主义——如何看待理论、因果与规律"一文中,笔者之所以提出大数据经验主义这个

收稿日期:2017-12-27
基金项目:华南理工大学中央高校基本科研业务费资助(XZD06)
作者简介:齐磊磊(1978—),女,山东临淄人,哲学博士,华南理工大学马克思主义学院副教授.

概念,有一个哲学,尤其是科学哲学的学科背景。在科学哲学视域下,从经验主义到逻辑经验主义再到新经验主义,贯穿其中的核心是"经验"的概念与意义。正是有了这个"经验"的存在,笔者才会链接到当下的大数据时代几位权威发言人的观点,才会创造性地提出"大数据经验主义"的这个说法。所以,提出"大数据经验主义"概念是对时代特征进行哲学反思的产物,具有可追溯的学术渊源。同样,在此基础上,笔者概括出的大数据经验主义的三个基本主张也是基于科学哲学的视角,是对大数据几位权威发言人观点的提炼与总结,这样的概括其核心的主张也是基于传统上对"经验"的解读与结合,"经验"一词可谓是"大数据经验主义"这个概念的灵魂。因此,"大数据经验主义"这个提法具有自身的独特性,是不该简称也不能简称,当然也是不能用其他概念代替的。

除了黄教授的这个简称,商业界确实也存在着"大数据主义"这个概念。为《纽约时报》撰稿长达20年的史蒂夫·洛尔(Steve Lohr)在2015年出版了*DATA-ISM: The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything Else*一书(中译本翻译为《大数据主义:一场发生在决策、消费者行为以及几乎所有领域的颠覆性革命!》)[3]他在书中引用了专业研究机构的数据、统计了大数据的规模与速度,说明我们的这个世界在大数据和云计算的互动中迅速进入到一个大数据构筑而成的数字世界。史蒂夫·洛尔认为对大数据的研究价值"更重要的问题是如何运用、如何理解这些数据。"[3]8基于这样的主旨,作者以大量企业和商界的案例阐述大数据何以成为"主义":比如重点关注那些处于数据科学领域前沿的年轻企业家和具有悠久历史的公司,重点介绍他们的相关经历,从他们的职业生涯的变迁来揭示数据技术与方法不断发展的步伐,抽象出大数据主义思想的典型代表,最后回归主题,大数据主义正在兴起。显然,史蒂夫·洛尔所谈及的"大数据主义",主要围绕一个比较宽泛的经济领域,调查那些具于前沿意识的项目与创意,与数据公司的科研人员、企业家共同讨论大数据理论。

那么,史蒂夫·洛尔所讨论的"大数据主义"与我们所说的"大数据经验主义"是一回事吗?当然不是!原因有二:其一,从学科领域与研究的侧重点来说,史蒂夫·洛尔的"大数据主义"是从经济学领域出发研究若干个商业案例与前沿科技公司,侧重的是大数据对人们思维与生活方式上的变革与影响,主要关注数据决策和数据应用方面。其二,主要是从因果与相关的表述角度。史蒂夫·洛尔想要"厘清大数据中的相关关系与和因果关系"时,他发现相关关系可以为商业、医学等应用领域提供有效的预测工具,但不能因此否定因果性。对于很多人曾经认为"对于大量商业决策而言,有相关性就能得出令人满意的结果"[3]163,史蒂夫·洛尔引用了IBM人工智能专家戴维·费鲁奇的反对观点:"商业战略与政策制定等决策领域面临更大的风险,仅凭相关性是绝对不够的。……未来的人工智能除了会数据分析以外,还要对因果关系产生有启发性的认识,包括理论、假设、现实世界的心理模型、事情的原委等,两者必须更密切地相互配合。"[3]164或许是受史蒂夫·洛尔的影响,黄教授简化的名称"大数据主义"与史蒂夫·洛尔的书名完全相同,同样黄教授为了支持他的商榷立场,在他的文章中也引用了史蒂夫·洛尔上面的这一段话。仔细分析,《大数据主义》中的这一段话所表达的观点恰恰反对的是笔者所提出的"大数据经验主义"对因果与相关关系所表述的意思。也就是说,从因果与相关的立场上,大数据主义是反对大数据经验主义的观点的。在这个意义上,将大数据经验主义直接简称为大数据主义是不恰当的。由此以大数据主义的立场进行商榷实际上在一定程度上支持了笔者对大数据经验主义的批判。

从另一个角度,笔者提出的"大数据经验主义"是史蒂夫·洛尔所说的"大数据主义"的加强版,或者称为强硬的"大数据主义"。这样的一种表述同样也来自史蒂夫·洛尔的《大数据主义》。史蒂夫·洛尔曾经介绍到摩根士丹利的前首席经济师理查德·伯纳,称他是一位有条件支持大数据的拥护者,在提到大数据作为金融显微镜的作用时,史蒂夫·洛尔说:"强硬的数据主义者认为无须任何理论,也无须借助为世界运行方式建立模型,单凭相关性就可以解决一切问题。"[3]161对于这种"强硬的数据主义者"的观点,理查德·伯纳说:"我认为,说相关性足以说明问题的人都应该反思。""在他看来,数据与理论(或者经

济行为模型)对于了解经济社会而言都必不可少。伯纳补充说,当前的这种争论在经济学史上早已有之,可以追溯至加林·库普曼斯在1947年发表的论文'缺乏理论的计量'。库普曼斯是一位荷兰裔美国经济学家,后来获得了诺贝尔经济学奖,他在这篇文章里对商业圈中的强硬'经验主义'方法进行了抨击。"[3]162这样,"大数据经验主义"除了有自身的学科背景之外,在经济学领域也找到了相应的理论源头,它们共同的焦点都指向了从"经验主义"的角度讨论因果与相关,只不过来自哲学领域的"大数据经验主义"是对强硬"经验主义"的拥护,而来自经济学领域的"大数据经验主义"是对强硬"经验主义"的批判。

退一步说,即使《大数据主义》只主张在商业或者经济领域,只需要进行数据分析就可以做出决策,那么作为对"大数据主义"的表述,这样的说法也没有任何问题,正如笔者界定的"大数据经验主义"有他自身的特征一样,"大数据主义"也有自身的特征。史蒂夫·洛尔在提出了"大数据主义"的名称后,在厘清大数据中的相关关系和因果关系时,即使提出这样的观点,即认为相关关系可以为商业、医学等应用领域提供有效的预测工具,因此而否定因果性,这也是他提出的"大数据主义"的题中应有之意,与其它背景下提出的"大数据经验主义"没有对比的基准。

至此,我们已经基本回答了商榷的焦点问题。相对大数据经验主义的观点和立场,大数据主义是如何看待理论、因果与规律的呢?我们接下来进行详细讨论。

二、大数据主义如何看待理论、因果与规律

区别于"大数据经验主义"相对比较极端的观点,"大数据主义"采取温和的态度来看待数据相关与理论、因果的关系。他们认为,数据的相关性是为了寻找数据规律以助于发现因果关系。这正是"大数据经验主义"一文中笔者所坚持的立场:"我们不否认大数据方法论,但并不赞同目前大数据时代引领下的这种大数据经验主义的神化观点。"[1]笔者当时使用的是"大数据方法论",所要表达的观点实际上与后来的"大数据主义"的观点一样,与黄欣荣教授商榷文章的立场也是保持一致的。由于"大数据经验主义"一文中已详细地讨论过大数据经验主义如何对待理论、因果与规律,对于安德森等人言辞过于激烈的论调,属于大数据经验主义的观点,此处不再赘述。对于舍恩伯格的有些观点,细心的读者会看到,我们此处也会涉及到一些。① 这恰恰表明:大数据经验主义与大数据主义是截然二分的,但很多人对大数据的观点和看法并不是一成不变的,他们在大数据"忽如一夜春风来"的冲击下或许发表了一些过激的言论,被归为"大数据经验主义"之列,但随着认识的深入、实践的应用以及冷静地思考,对大数据的态度会发生改变(比如他们会更为正确地看待大数据与理论、因果、规律之间的关系),进而转向"大数据主义"。下面详细讨论大数据主义对待理论、因果与规律的态度。②

大数据的风云人物舍恩伯格与库克耶反对安德森"理论终结"的说法。他们认为:"'理论的终结'似乎暗示着,尽管理论仍存在于像物理、化学这样的学科里,但大数据分析不需要成形的概念。这实在荒谬。"[4]93进而,他们表达了大数据与理论关系的看法:"大数据是在理论的基础上形成的。比方说,大数据分析就用到了统计和数学理论,有时也会用到计算机科学理论。……建立在这些理论上的大数据分析模式是实现大数据预测能力的重要因素"[4]94从这些言论上看,舍恩伯格与库克耶把理论看作是主体部分,大数据的产生离不开理论的支撑,对大数据的分析以及具体应用(如预测)也都是以理论为基础的。

谈到大数据整个处理过程,舍恩伯格与库克耶的观点更为显明:收集大数据时,理论影响着我们如何

① 舍恩伯格与库克耶合著的《大数据时代》一书中,许多观点与立场也不是非常明确、清晰,甚至有些前后并不一致。

② 理论、因果与规律,三者具体表达的虽然有差别,但相对于大数据,它们又是一个"统一战线",所以本文将三者看作一个整体而未作刻意区分,根据具体情况提及其中某个或某些个,有时也将"因果与规律"涵盖在理论之中,以"理论"为代表。

做出相关的决定;分析大数据时,我们使用什么样的分析工具也依赖于理论;分析大数据最后的结果时,同样也离不开理论的指导。因此,他们的结论是:"大数据时代绝对不是一个理论消亡的时代,相反地,理论贯穿于大数据分析的方方面面。"[4]94 具体来说,我们可以先设定一个问题,使用大数据来分析、验证计算机借助算法生成的若干可能性假设,而不是依靠经验或实验逐个验证,这样的方式去除了对既有认知的阻碍,从统计学的角度提高了精确性。仔细分析两者的区别:使用计算机的算法程序产生的大量数据可以验证问题的所有可能的答案,最后选取其中最优的一个;而传统的经验试错法有可能会丢失某些关键的数据而造成解答的偏差。但同时我们还要考虑这样一个问题:有时候数据并不是越多越好,如果不加选择地随意使用大数据则会存在一些潜在的风险。比如当有人为了某种目的而恶意提供虚假的数据,如果使用者直接采用而不作理论上的分析,那势必会产生错误的结论或做出糟糕的决策。

大数据以理论为根基,"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。"[4]94 与小数据时代不同,大数据可以帮助研究者找到以前所发现不了的规律与因果联系,除了在商业、科学等诸多领域带来的大的变化,大数据为更好地认识世界提供了更多的方式与可能。目前的这个世界变得更加复杂,随之带来的不确定性远超我们的想象。因此,当人们使用大数据探索世界时,他们可能会获得更好的理解,相应地会提高解决问题的能力和决策水平。人们寻找因果关系是一种与生俱来的能力或习惯,我们随时准备着从因果关系的角度来认识世界,大多数情况下,人们只有真正地解释与理解世界内部究竟是怎么一回事时,才会感到欣慰。虽然实际发现的因果关系并没有想象中的多,甚至有些是错误的①,但这并不是只要相关性而放弃寻找因果关系的理由。

因此,大数据主义者是比较温和地看待理论、因果与规律。除此之外,大数据主义者也认为:"大数据的发展可能会改变经济和社会生活,可能会改变科学研究的途径,甚而改变人类的思维方式。"[5] 如今,大数据处理技术会对来自各方面的大量信息进行分析,当你在网上搜索时,大规模数据库可以满足我们的访问,帮助我们做出更好的决策,譬如你在网上购书,系统会给出百分之多少的人也浏览过这本书,百分之多少人购买,有哪些书与其搭配购买。就像这样,我们的很多行为都被数据化。购物、社交、爱好等等都被大数据分析,这些数据潜移默化地改变着这个社会,改变着人们的行为习惯与思维方式。

三、大数据在理论、因果与规律中的位置

基于对大数据与科学理论关系的思考,很多学者对传统的科学发现模式产生了新的看法,认为"科学始于数据"。黄欣荣教授详细梳理了科学哲学中曾出现的科学发现模式中的几种范式,并以此为基础得出了这样的结论:"在大数据时代,知识的发现可以从数据开始,不再需要预先做出理论的假设。"[2] 黄教授此处使用"可以"而不是某种更强硬的语气,不是强调一定要从数据开始,这样的表达方式表现了大数据主义较温和的态度:知识的发现可以从假设与模型开始,也可以从数据开始,前者为主,后者为辅,或者说后者是前者的有益补充。

关于这一点,吉姆·格雷(Jim Gray)作为计算机专家,从科学记录的角度倡导了"科学研究的第四范式"(也有人称作"数据密集型科学"),更为全面地分析了科学发现可以从大数据开始。黄教授赞同格雷对科学发现模式所作的系统的四种分类,前两种范式(经验(实验或试验②)范式和理论范式)是科学哲学历史上两大流派的核心观点;第三种计算范式,即大规模的计算机模拟,它的出现是由于 20 世纪中期,

① 深层的研究显示,通常我们对因果关系的快速直觉是完全错误的。参见维克托·迈尔-舍恩伯格,肯尼思·库克耶:《与大数据同行:学习和教育的未来》.赵中建,张燕南译,华东师范大学出版社 2015 年出版,第 47 页。

② Tony Hey, Stewart Tansley, Kristin Tolle 合著的《第四范式:数据密集型科学发现》一书中译者翻译为"试验"。

“支撑试验和理论的计算技术的同时增长，加大了传统科学记录的压力。不仅底层数据在持续增加，模拟和试验的产出也变成大型而复杂的数据集，它们只能总结性地出现（不能完整地被记录）在传统出版物中。”[6]184 在这样的情况下，计算技术成为产生大量数据的工具，大量的数据推动了科学理论的发展，计算数据的记录用来补充实验方法的传统描述。它所处的位置等价于传统中的实验数据，大量的实验数据在理想状态下是可以带来更好的理论规律或科学假设，推动科学理论的发展。科学理论的世界发生了变化，随着收集的数据或模拟产生的数据爆炸式地增长，“从计算科学中把数据密集型科学区分出来作为一个新的、科学探索的第四范式颇有价值。”[6]xi 因此，形成于新的发展形势下的第四范式并没有要取代前三个范式的意图，相反还成为加强大数据与理论密切关系的粘合剂：“在一定意义上，格雷的第四范式提供了一个集成框架，使前三者（范式）相互作用，相得益彰”[7]181，即格雷自己所说的：“模拟、理论和试验在大量数据背景下必须携手合作。”[6]181 这样的描述恰恰说明目前大数据在科学理论中的地位。

从科学记录的角度对大数据引起的第四种研究范式的分析，大数据主义的看法可以用天文学上的一个案例形象地表达：正如开普勒利用布拉赫对天体运动的大量观测数据中发现了行星运动三定律一样，对大数据的分析引发产生了若干新的理论，“在对所采集并仔细保存的实验数据进行挖掘和分析的基础上建立起新的理论，也正是第四范式的一个重要特征。”[7]前言 iii

通过以上分析，在笔者看来，大数据与理论的最根本的关系可以归结为：大数据帮助发现理论。这种帮助作用并不仅仅只停留在“假设-模型或实验”阶段，大数据的助推作用贯穿于发现理论的多个环节与过程中。但是，就像拉卡托斯的“研究纲领”所要表达的意思，大数据在科学理论发现中的这种积极作用并没有改变“研究纲领”中的“内核”。也就是说，目前科学理论的发现过程中，虽然大数据起到了重要的作用，但并没有取代其他范式建立一种以大数据为中心或者是以大数据为起源的研究范式，而是仍然遵从于以问题为导向的理论研究。如果“科学起源于数据”，那就会陷入漫无目的地收集数据的海洋，即使能够做到大数据主义所主张的全数据分析，那么在收集这些全数据时也要针对一个明确的问题，不然全数据收集就会陷入自己的悖论之中，是不可能完成的。

如果把“科学起源于大数据”当作是一种研究方法，那么理想中或逻辑上的全样本分析实际上是一种完全归纳。这种完全归纳如果可以实现，就会更容易探明因果关系。按照科学方法论，科学的归纳在于寻找因果关系，进而提出规律或理论。所以，那些通过大数据的分析只关注相关性就可以的研究者实际上是“用大炮打蚊子”，不是说不可以，实在是浪费了我们的“大数据时代”。于是我们可以说：不以寻找因果关系为目的的大数据研究是不彻底的。

四、结语

利用大数据得出理论、因果与规律，实际上像传统科学研究的过程、步骤一样，只不过是用于分析的数据量的大小的差别，相应地会有不同的研究方法或者可能更接近于真实的结果。除此以外，并没有更大的神秘。

大数据只是帮助研究者更好地发现理论、因果与规律，是假设-模型-理论中的一个有效的发现方法，处于辅助地位而不能代替它们中的任何一个。在大数据使用的“婴儿期”，类似像大数据的拥护者所断言的：“我们正处在一个认识论的革命之中，因果分析和理论生成会被现代主义方法论毫不留情地取代”[7]，以及只要数据不要理论、只要相关不要因果这样的言论为时过早。

让商界的归商界，学术的归学术。在使用大数据时，目前的资料文献大都集中于商用案例的应用描述或分析。商界注重应用，学术注重研究，由于各自侧重点的不同，导致对大数据的态度会不一样，但仔细分析，商界的使用最终也是要回归到数据的分析，最终还要借助理论进行，最后还要究其原因。大数据

中经常被使用的案例有一个是关于2009年谷歌成功预测了禽流感:通过大数据的统计,集中在一段时间内某一地区的人们搜索“发烧”“头痛”“咳嗽”等特定词条频率大量增加,谷歌公司由此断定在这个地区会引发禽流感。这个事件也让大数据包括谷歌公司名声大振。但遗憾的是,这样的原理却在2011—2013年间推出错误的结论,出现“大数据,大偏差”的窘境,究其原因主要是因为对大数据只关注相关性而忽略了理论与因果关系的讨论,这样得出的规律用铁的事实告诉我们是不恰当的。

忽如一夜智能启,千数万数汇集来。我们这个时代,恒河沙数的数据势如破竹,我们唯有正视它、利用它才不会成为时代的弃儿。但同时,我们也不能唯数据论,把数据当作替代理论和因果的尚方宝剑。新的时代要有新的思维与方法,培养大数据的理念与思维,不仅要顺大数据之势而谋,还要应大数据之势而为,学术研究中应该正确使用大数据并使其最终服务于理论、因果与规律的研究。

参考文献:

[1]齐磊磊.大数据经验主义——如何看待理论、因果与规律[J].哲学动态,2015(7):89-95.

[2]黄欣荣.大数据如何看待理论、因果与规律——与齐磊磊博士商榷[J].理论探索,2016(12):33-39.

[3]史蒂夫·洛尔.大数据主义[M].胡小锐,朱胜超,译.北京:中信出版集团,2015.

[4]维克托·迈尔-舍恩伯格,肯尼思·库克耶.大数据时代:生活、工作与思维的大变革[M].盛杨燕,周涛,译.杭州:浙江人民出版社,2013.

[5]李军.大数据:从海量到精准[M].北京:清华大学出版社,2014:40.

[6]HEY T,TANSLEY S,TOLLE K.第四范式:数据密集型科学[M].潘教峰,张晓林,等译.北京:科学出版社,2012.

[7]CHANDLER D. A world without causation: Big Data and the coming of age of posthumanism[J]. Millennium: Journal of International Studies,2015(3):833-851.

Big Data-ism and Big Data-Empiricism

——Answer to Profess Huang Xinrong

QI Leilei

(*Research Center for Philosophy of Science and Technology*, *South China University of Technology*, *Guangzhou* 510641, *China*)

Abstract: Big data-empiricism is a concept of academic origin based on the philosophical perspective, which is harmonious with the prevailing big data-ism yet different from it. The position of big data-empiricist is clear that the big data age does not require theories and correlation substitutes for causality. The attitude of big data-ism is more moderate in that the theory is the basis to deal with the whole process of big data. They think the correlation between big data helps to find a scientific rule. Correlation is the appearance and causality is the nature of things. The purpose to access correlation is to better find causality.

Key words: Big Data; Big Data-ism; Big Data- empiricism; causality; theory; regularity

(责任编辑:黄仕军)