文章编号: 1000 - 8934(2017)05 - 0092 - 06

由大数据引起的对因果与相关的讨论

齐磊磊

(华南理工大学 科学技术哲学研究中心,广州 510641)

摘要: 大数据带来了因果与相关的显明观点,共鸣者众,争鸣者也不乏其数。因果与相关本来就是哲学上的老问题,老问题新讨论,大数据视域下重新审视两者的关系是对哲学或者大数据哲学的有益补充。相关关系可以细分为(决定论)因果、统计因果与非因果关系。相关关系包含了决定论的因果关系,决定论的因果关系必定是相关关系; 统计因果找到了协调传统科学哲学的方法论与大数据方法论的中间桥梁,是大数据研究的一个中间驿站; 大数据视域下通过统计因果相关可以推测集体的和个体的因果关系,但不能给出明确的证明依据; 可以借助统计因果相关对因果与相关进行区别并联系起来。

关键词: 因果; 相关; 函数; 规律; 大数据

中图分类号: N031　　文献标识码: A

一、引言

笔者于 2015 年写过一篇论文“大数据经验主义”[1],提出了大数据经验主义的概念、概括并批判了他们的基本观点。论文发表以后,国内多位学者与我交流,其中最有启发性的建议是: 对因果与相关的讨论再细致些,要抓住这个问题深入说清楚。为了更好地说明因果与相关的关系,本文将相关进一步细分为: 决定论因果、统计(概率)因果和非因果相关三个部分进行讨论,尤其是引入统计因果相关这个较少有人提及的概念,它既可以清楚地表达因果与相关的区别,又是两者之间联系的纽带,这样的论证进路具有一定的新颖性。为此,首先沿着从“函数”到“相关”①再到“规律”最后到“因果”这样一个自然过渡的发展路线开始。

二、函数、相关、规律与因果

在现实世界中,任何事物都不是孤立存在的,而是与其他事物具有千丝万缕的联系。对于这种相互关系的研究,有一个从朴素的直觉表达到精确的数学描述的过程。最先给出统一描述的来自数学上的“函数”概念,这是 17 世纪数学从对运动的研究中引出的一个基本概念。伽利略在近代力学的开山之作《两门新科学》中用文字和比例的语言表达函数关系,全书中比比皆是,“只差把文字叙述表为符号形式这短短的一步了。”[2]44 随后,苏格兰数学家詹姆士·格雷戈里(James Gregory)在他的论文“论圆和双曲线的求积”中给出了相对比较明显的定义,但范围太窄。牛顿在他的微积分研究中用“流量”(fluent)来表示变量间(包括无穷小量间)的关系。1673 年,莱布尼兹在一篇手稿中使用“函数”表示任何一个随着曲线上的点的变动而变动的量。1714 年,在莱布尼兹的著作《历史》中,用“函数”一

收稿日期: 2017 - 2 - 24

基金项目: 国家社会科学基金一般项目“语义模型与表征模型研究”(14BZX025)、2016 年度广州市哲学社会科学“十三五”规划课题“大数据哲学中理论与因果问题研究”(2016GZGJ57)。

作者简介: 齐磊磊(1978—),女,山东临淄人,博士,华南理工大学马克思主义学院副教授,主要研究方向: 科学哲学与复杂系统哲学等。

①对于“相关”,通常有几种不同的表述方式,如相关性、相关关系、相互关系、关系,本文不作刻意区分。

词来表示依赖于一个变量的量。伟大的数学家欧拉 1734 年引进了函数的数学记号 y = f(x) ,这临门一脚踢开一般数学函数定义的大门大概花了 100 年。

在莱布尼兹以及其后的数学家看来,只要事物之间在物理上存在着严格的确定的关系就可以用函数关系表示。随着离散数学和集合论的创建,函数的概念变得更加广泛。因为,集合论为刻画事物之间形形色色的联系提供了一种数学模型——关系,它仍然是一个集合,以具有那种联系的对象组合为其成员。比如人与人之间有父女关系、师生关系等; 计算机程序间有调用关系、状态转换关系等。"集合论中关系不是通过描述关系的内涵来刻画这种联系,而是通过列举其外延(具有那种联系的对象组合的全体) 来刻画这种联系。"[3]87所以,集合论中通常使用具有相互联系的对象的有序对的集合来表示关系。这样,一个数学集合,只要有个映射,即一个有序对,映射过来就是个关系。塔尔斯基从数理逻辑上对关系的表述也很广泛。他的定义是: "事物 x 与事物 y 有 R 关系,简写成: xRy"。[4]93其中 R 指的就是相关关系,它说明相关是指只要两个变量有关系,不管它稳定不稳定,明确不明确都是指具有这个关系,这与莱布尼兹等人对函数最初的定义不同。函数的定义演变为: 只要有两个变量之间发生联系,它就是一个函数,自己也可以是自己的函数。比如 A = A 是个相关,具有可自反的关系; A = B 也是个相关,具有相等但不是因果的关系; 兄弟也是个关系,这个关系就是 R,R 也称之为函数。所以说,相关关系可以表示为一个数学函数,函数就是一个映射,变量之间存在一个映射,无论是人为的、天然的或者数学上把它们连接在一起就变成一个有序对,有序对就是一个关系。于是非数学的关系也纳入到相互关系的研究之中。

数理逻辑展示了许许多多的关系形式,关系或函数在逻辑上实际是一个概念,在离散数学或集合论中也是一个概念。但是,在集合论中和数学中,函数或关系不一定是规律。因为函数是指对于任何两个变量之间的关系,而规律或者定律一定要限制在一个物理系统或物理实体之间,只有具有这样关系的事物才属于一定的自然类,这种意义上来讨论关系它就属于一个自然律。那么现在的问题是: 相互关系在什么条件下会成为自然规律呢? 也就是说,要谈论相关关系,这里还需要说明一个自然律和自然类的问题: 在事物或现象之间的关系中,有一些是必然的有一些是偶然的,有一些是属于一定自然类的本质关系的,有一些是非本质的,有一些是比较普遍的、重复出现的,有一些是个别的、转瞬即逝的。我们只把那些本质的、必然的、反复出现的相互关系叫做规律性和自然规律。因此,关系要成为自然规律,前提条件就是这种关系必须属于一个自然类,并成为这个自然类中的一个本质的联系。如果它们之间不存在本质的联系,它们也不能成为规律。不过,在规律中,有一些是决定性的,有一些是统计性的。统计性规律虽然不确定,但它仍按某种规律在一定的范围内变化。变量间的这种相互关系和函数关系,称为具有不确定性的相关关系,即我们通常所说的"统计相关关系"。如目前我国在"全面放开二孩"后的生育率与人均 GDP 的关系就是典型的统计相关: 人均 GDP 高的地区,生育率往往较低,但二者没有惟一确定的关系,这是因为除了经济因素外,生育率还受教育费用、父母的时间精力、婚嫁成本以及风俗文化和其他随机因素的共同影响,随着这些因素的变化,生育率虽然不完全确定,但却有一定的规律可循。

这样就可以将相互关系划分为三种形态: 决定论自然规律(白色区域) 、统计性规律(黑色区域) 和偶然性关系(灰色区域) 。

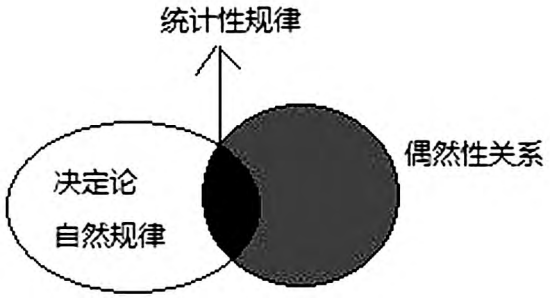


图 1　相互关系的三种形态

对相互关系和规律性进行了分析,是否就说明通过规律性完全可以解释这些现象,发现事物间的因果关系? 不一定! 因为数学上的函数与规律并不意味着变量之间就一定存在任何"因果"关系,虽然在普通的语言中"函数"这个术语往往带有因果的含义。例如 "对在常温下一个封闭容器内的气体,波义耳定律叙述为压强 p 和体积 v 的乘积是一个常数 c(这个值和温度有关) : pv = c. 用这个关系可以把 p 或 v 解出来,使 p 和 v 中的任何一个可看

作是另一个变量的函数，如 $p=\frac{c}{v}$ 或 $v=\frac{c}{p}$. 这里并不含有体积的变化是压强变化的‘原因’的意思，正如不含有压强变化是体积变化的‘原因’的意思一样。函数只是数学家所关心的两个变量间联系的方式。”[5]282

三、因果与相关

“事出有因”似乎是人们与生俱来的对外界事物探索的一种本性，万事万物都要追问“为什么”。为什么会有四季的轮回？为什么木头可以燃烧？为什么时间一去不复返？……对“为什么”的探求过程几乎伴随着人类的整个历程。遗憾的是，自从亚里士多德提出四因论以来，历经两千多年的努力，科学家在解释自然现象产生的原因，哲学家也把它上升为“因果”并作为一个基本的概念进行讨论，但却都没有对因果性或因果关系给出一个统一的描述。当代哲学的因果理论门类众多，张华夏教授在新著《科学的结构》中将概念分析与经验分析这两条因果理论研究进路整合起来，同样，本文讨论的因果既不单指“休谟—马奇学派的条件因，也不仅仅是指洛克—马顿—邦格的作用动力因”[6]214，而且还包括概率因果等等诸如此类由于研究的学派不同而提出的不同的因果理论。也就是说这里的“因果”实际上是一个概述性的词语，包含了科学与哲学上所讨论的任何一种对因果关系的表述，可以笼统地表示为：因果关系指的是事件之间的一个序列，如果事件 A 引起事件 B，则事件 A 是原因，而事件 B 是结果。这里“引起”一词可以依不同学派作不同的解释，可以解释为 A 是 B 的充分/必要条件；也可以解释为有一种因果力（能量或其他守恒量）从 A 传递到 B 的实体使 B 出现；也可以理解为 A 以一定的概率导致 B 出现。这第三种“引起”，被称为“概率因”。在此我们要补充一个统计概率因：A 是 B 的概率因，可定义为：$A=\mathrm{pro}(B)$. 其条件是：$P(B\mid A)>P(B\mid \bar{A})$.[7]202 其中 pro 为概率因的记号，p 为概率的符号，概率的取值范围在 0 至 1 的区间里。这个式子表示 A 对 B 的概率相关性：A 可能是 B 的原因，因为它提高了 B 的概率。这个概率因与大数据分析中的统计相关有密切的联系，下文会重点讨论。

按照这样的表示，相关关系也可以进一步具体化。A 和 B 相关，指的是事件 A 和事件 B 至少存在下列 6 种情况：（1）事件 A 直接引起事件 B；（2）事件 B 直接引起事件 A；（3）事件 A 引起事件 B 随后事件 B 引起事件 A；（4）事件 A 引起事件 C，而 C 又引起了事件 B；（5）事件 A 以一定概率引起事件 B 和事件 C；（6）非因果相关，例如数据收集。

显然，情况（1）－（4）所描述的 A 和 B 之间的相关实际上是决定论的因果关系的各种表现，它们依次是：直接因果、反向因果、循环因果（或因果反馈）、间接因果（或因果可传递性）。通常所说的因果性，指的是上述四种情况。（5）所描述的因果关系是统计概率性的，如图 2 所示。这种情况涉及到的正是萨尔蒙提到的概率因。一个原因引出两个结果，原因与结果之间存在概率因果关系，但两个结果之间没有因果关系。这里的概率指的就是前提条件，有这个条件就有这个概率，如果有这个条件的概率比没有这个条件的概率大，那么这个条件就是它的概率原因。例如，某段时间，冰激凌的销量（C）和中暑人数或个人中暑的可能性都会增加（B），两者之间表现出相关关系，如果我们依靠冰激凌销量升高的信息进行中暑预防，有时是有效的，有时会有很大的误差，因为二者之间并不是因果关系，正确的途径是找到它们共同的原因：天气炎热（A）。然而，天气炎热与个人中暑之间的因果关系是概率性的，A 与 B 之间以及 A 与 C 之间的关系是统计因果相关。而 B 与 C 之间的关系则完全是非因果关系，尽管它们之间有共同增长的关系。这样我们可以推出一个结论：相关关系包含了确定性的因果关系，确定性的因果关系必定是相关关系，如情况（1）－（4）；相关关系不一定是确定性的因果关系，如情况（5）、（6）；相关关系可以提供可能性，通过概率统计因果相关用于推测集体的和个体的因果关系，但不能给出明确的证明依据。例如张三是否中暑，还要有医学证据。

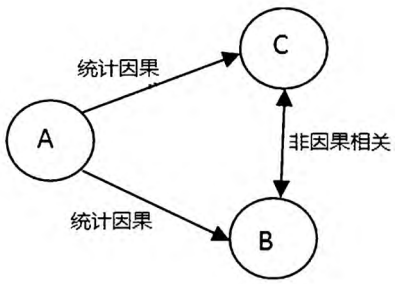


图 2　相关关系之统计概率因果关系

分析了因果与相关之间的关系之后，接下来讨论大数据时代对因果和相关的各种说法和看法就会更清晰。

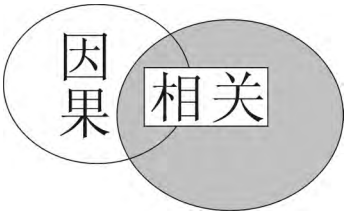
四、大数据视域下的因果与相关

大数据权威发言人舍恩伯格认为，“知道‘是什么’就够了，没必要知道‘为什么’。在大数据时代，我们不必非得知道现象背后的原因，而是要让数据自己‘发声’”[8]67“他对相关性的至高地位进行了有力的辩解‘寄希望于识别因果关系是一种自得其乐的幻想，大数据必将打破这种幻想。’”[9]159《连线》杂志主编克里斯·安德森有着更为激进的看法“相关取代因果，科学的进步甚至无需一致的模型、统一的理论和任何机理上的解释。”[10] 笔者在“大数据经验主义”一文中指出这样的看法是片面的。黄欣荣教授对此写了一篇商榷文章。他认为：“在大数据时代，由于数据的暴增，寻找数据间的相关性比因果性更重要，大数据主义承认事物的因果性，但更应该把握事物的相关性。”[11] 董春雨教授等人撰文提出“之所以出现相关优于因果，相关取代因果等的极端看法，实际上是他们没有认清二者之间的区别与联系。”[12] 随后他们分析了因果与相关的区别与联系，总结说“厘清因果性之于相关性的关系和意义，是大数据哲学探讨中必须深究的问题之一。”[12] 戴潘博士通过分析大数据知识发现的分类树算法表明“大数据所主张的相关关系来取代因果关系，其实并不是要抛弃因果关系，通过分类树算法的分析，可以发现其中所蕴含的因果结构。”[13] 对比分析这些观点，引起关注、讨论与商榷的主要焦点还是因果与相关的关系问题。

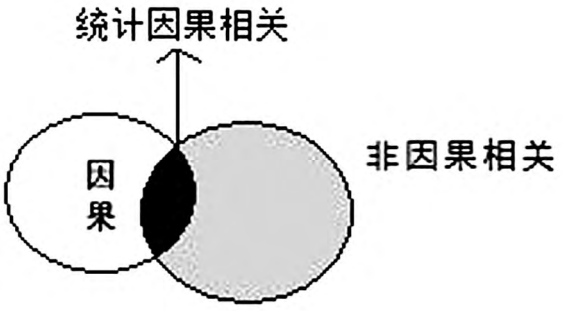
从大数据的立场出发，因果与相关，哪个更重要？相关关系是否可以替代因果？回到上面我们对概念的讨论，上一小节我们将相关关系具体分为6种情况时的结论有：相关关系包含了因果关系，因果关系必定是相关关系。按照这样的表达，相关关系既然包含了因果关系，那么大数据研究者们提出的“只要关注相关关系就够了”这样的说法也是正确的。但也有另外一种情况，即情况(5)(6)所表示的：相关关系不一定只是(决定论的)因果关系，它也包括统计因果与非因果相关。所以，只有在情况(5)(6)的前提条件下，从大数据出发对因果关系与相关关系所进行的讨论或争论才是有意义的。

进一步地追问：如果承认只有在情况(5)(6)的前提下才能对大数据视域下的因果与相关进行讨论，那是否就意味着情况(1)-(4)这些因果关系的各类表现不再是相关关系？显然不行。从逻辑上说，因果关系必定是相关关系，情况(1)-(4)是可以被列入相关关系的归纳式定义中的。尽管这样的一种定义与大数据发言人的表述有相异之处，但只要我们共同约定他们的论点前提在情况(5)(6)的条件下是成立的，就是站得住脚的。

再回到“大数据经验主义”一文。文中我们为了说明传统的科学方法论和大数据经验主义在因果与相关关系上的分歧，给出了一个简单的图示。[1] 仔细推敲图中用一个维恩图表示的“因果与相关的缠结”，截取下来如图3(a)所示。此图中因果与相关有个交集，结合上面的分析，这个交集表示统计因果相关，它既是因果又不是因果。说它是因果，主要指的是事物之间常常有一个概率的关系，是一个概率统计的因果。说它不是，主要是因为我们通常所说的因果关系，主要讲的是决定论的因果关系，一般不涉及概率统计因果。所以，因果与相关的交集部分不是拉普拉斯的决定论。拉普拉斯的决定论是世界上你知道一切的原因，未来的一切都可以了如指掌。但实际上并不是如此，这样就有一个概率因的问题。通常，概率因可以有两种解释，一种解释是它有多少概率成为原因，另一个是概率本身是一个结果或者原因，概率本身本来就是个概率，或者说它的出现本来就是个概率。这是从本体论上说，如果说有多少概率成为原因，那么实际因果的充分条件就必须找出来。统计概率因果的概念之所以必要，第一是能使我们从拉普拉斯决定论中解放出来；第二是使我们的传统科学哲学的方法论与大数据方法论协调起来，避免争论。大数据并非拉普拉斯决定论，大量问题是统计方法论，需要用统计因果相关来进行说明。另一方面，有些非本质主义者会重视概率因果，这样做的好处在于说明原因时不是太死板，而是说有多少概率成为原因。因此，整个并集区域包括三个部分，如图3(b)所示：(决定论的)因果集(白色区域)，统计因果相关集(黑色区域)和非因果相关集(灰色区域)。上面定义中的情况(1)-(4)是属于(决定论的)因果集，情况(5)属于统计因果相关集，情况(6)属于非因果相关集。



(a)



(b)

图3　因果与相关的缠结

图3(b)与图1的关系是:统计因果相关是一个不确定的关系,但是自然律有个重复性,它不是说不确定,而是在自然界中是反复出现的。自然律是和事物的本质或者是新本质主义的自然类、它的倾向性或者是它的本质联系、它的实体发生一种比较稳定的关系。所以自然律不一定是因果律,非因果律也可以是自然律。

这样,经过以上分析,相关与因果中间交集部分是统计因果。有了这个统计因果,不管是确定性的还是不确定性的关系,统计因果都可以帮我们分类说明。相关不一定是确定性的,也不一定是不确定性的,任何相关都是一个函数,或者像塔尔斯基所说的关系,统计因果相关当然也是个关系。同样,大数据视域下对因果与相关的讨论依然可以借助统计因果相关将两者联系并区别开来,而由大数据引起的关于因果与相关的争论也由于统计因果的细分而清晰起来。例如2009年谷歌利用搜索"发烧"、"头痛"、"咳嗽"等特定词条频率的增加预测了禽流感的案例中,特定词条与禽流感之间有时是一种非因果相关。因为有些人没有任何症状,但看了新闻报道或者其他原因,他也会搜索这些特定词条,这并不代表禽流感会出现。同时,有一些人确实是因为有了相关症状后去网上搜索。所以网上搜索特定词条的人可能与禽流感相关,但这是一种概率性的,至于概率是多少,需要进一步去数据分析与统计。即使概率极高,仍不是实际原因的充分条件。虽然政府最终会去查找引起禽流感的真正原因,但运用大数据分析出的导致禽流感的概率条件不再使我们只拘泥于拉普拉斯式的决定论因果中,统计概率因果找到了协调传统科学哲学的方法论与大数据方法论的中间桥梁,是大数据研究的一个中间驿站。

在大数据时代,海量数据带给我们诸多恩惠,大数据将"大"有可为,但不管是商界还是科学研究中,"放弃对因果性的追求,就是放弃了人类凌驾于计算机之上的智力优势,是人类自身的放纵和堕落,如果未来某一天机器和计算完全接管了这个世界,那么这种放弃就是末日之始。"[8]IX所以,为了将来的人类仍然还是人类,让现在的我们继续保持与生俱来的天性,在大数据时代,运用大数据去追问"为什么"。

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][美]莫里斯·克莱因. 古今数学思想(第二册)[M]. 朱学贤,申又枨,叶其孝,等译. 上海:上海科学技术出版社,2002.

[3]王元元,张桂芸. 离散数学导论[M]. 北京:科学出版社,2002.

[4][波兰]塔尔斯基. 逻辑与演绎科学方法论导论[M]. 周礼全,吴允曾,晏成书,译. 北京:商务印书馆,2010.

[5][美]莫里斯R. 柯朗,H. 罗宾. 什么是数学:对思想和方法的基本研究[M]. I. 斯图尔特修订. 左平,张饴慈,译. 上海:复旦大学出版社,2008.

[6]张华夏. 科学的结构:后逻辑经验主义的科学哲学探索[M]. 北京:社会科学文献出版社,2016.

[7]Wesley C. Salmon. Causality and Explanation[M]. Oxford University Press. 1998.

[8][英]维克托·迈尔-舍恩伯格,肯尼思·库克耶. 大数据时代[M]. 盛杨燕,周涛,译. 浙江:浙江人民出版社,2013.

[9][美]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.

[10]Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete[J]. Wired 16, issue 7, July, 2008. https://www.wired.com/2008/06/pb-theory/

[11]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.

[12]董春雨,薛永红. 从经验归纳到数据归纳:特征、机制与意义[J]. 自然辩证法研究,2016(5):9-16.

[13]戴潘. 基于大数据的科学研究范式的哲学研究[J]. 哲学动态,2016(9):105-109.

A Discussion caused by Big Data on Causality and Correlation

QI Lei – lei

(Research center for Philsophy of Science and Technology, South China University of Technology, Guangzhou 510641, China)

Abstract: Big data brought about the obvious points on causality and correlation, which are agreed by many scholars, but there are a lot of scholars have different views on them. The relation of causality and correlation is an old problem of philosophy. But there are new discussions on the old problem. It is a useful complement to philosophy or big data philosophy to re – examine the relationship between causality and correlation from the perspective of big data. The correlation can be broken down into determinism causality, statistical causality and non causality. The correlation contains the determinism causality and the determinism causality must be the correlation. Statistical causality found the middle bridge to coordinate the relationship between the methodology of the traditional philosophy of science and the big data methodology, which is a middle station to research big data. Using statistical causality can infer the causality between the collective and the individual, but it can't give clear evidence. It can be distinguished and linked causality and correlation by statistical causality.

Key words: causality; correlation; function; regularity; big data

(本文责任编辑: 董春雨)

(上接第 91 页)

On the Limitation and Transcendence of Big Data Thinking

DIAO Sheng – fu[a], YAO Zhi – ying[b]

(a. School of Economic Management and Law, b. School of Marxism, Foshan University, Foshan 528000, China)

Abstract: When highly valuing the great importance of big data thinking, we need to maintain a rational mind to treat its limitations seriously, which refer to misunderstanding of the "whole data model", the anxiety of quantitative thinking and excessive worship of correlation; so it is necessary to take into account both the whole and parts, integrate quantitative and qualitative approaches, emphasize the correlation's complementary role of them, and then achieve transcendence of big data thinking.

Key words: big data thinking; limitation; complementarity; transcendence

(本文责任编辑: 董春雨)