doi:10.13582/j.cnki.1672-7835.2021.04.005

认知盲点理论视角下的纽科姆疑难探析

史红继
(南京大学 哲学系,江苏 南京 210023)

摘　要:自诺齐克提出纽科姆疑难以来,关于其基本症结及出路一直存在激烈的争论。运用索恩森提出并为孔斯等人所改进的认知盲点理论,不仅可以揭示纽科姆疑难的实质,而且可为消解因果决策理论(CDT)与证据决策理论(EDT)之间的冲突指明新的路径。在纽科姆故事场景下,尽管因涉及选择行为实现方式的思考,使得 EDT 优于 CDT,但这并非意味着 EDT 可以完全解决该疑难。在将决策行为视为完美干预的情况下,纽科姆问题最佳解决方案是两盒选择,但该选择只是基于决策因果结构的特性,而非 CDT。

关键词:纽科姆疑难;行为决策;认知盲点;悖论
中图分类号:B81　　**文献标志码**:A　　**文章编号**:1672-7835(2021)04-0033-07

“盲点(blindspots)”这一术语常见于生理学,用以指代视网膜上无感光细胞的部位,因此处仅有神经纤维而无感光结构,不能感光成像,故称为盲点。盲点在视野内所占面积微小,且临近部位的活动可对其代偿,因此平时并不为人所觉察,也不会对人的正常视觉造成障碍。索恩森(R.A. Sorensen)根据上述生理学盲点的特征加以推广,提出了“认知盲点”概念:一个命题 p 是关于给定命题态度 A(比如知道、相信)和一个给定认知主体 a(在时间 t)的一个盲点,当且仅当,p 本身是无矛盾的,但 a 对于 p 不能持有命题态度 A。其中所谓“不能”受限于纯逻辑规则、物理规则以及心理学规则等。例如,“下雨了,但 a 并不知道(相信)”这个复合命题本身是无矛盾的,然而若 a 不能对该命题持有“知道(相信)”态度,否则就会产生矛盾①。“天正在下雨但 a 不知道”,对 a 来说,就构成一个认知盲点。令 p 代表“天正在下雨”,Kap 表示 a 知道 p,则有:

(1)Ka(q &¬ Kaq) 假设
(2)Kaq&Ka¬ Kaq　1,知识在合取上的分配律
(3)Kaq&¬ Kaq　2,知识衍推真

索恩森基于认知盲点的概念开发了一种悖论的解决方案——认知盲点理论。该理论虽被揭示出一些重要缺陷,但经过孔斯(R. C. Koons)等人的改进已逐步趋于完善。然而,认知盲点理论在学界长期论争的纽科姆疑难研究中的运用,尚未得到深入研讨并达成共识。本文试图表明,将经过改进后的认知盲点理论运用于纽科姆疑难研究,不仅可以揭示纽科姆疑难的实质,而且可为消解因果决策理论(CDT)与证据决策理论(EDT)之间的冲突指明新的路径。

一　纽科姆疑难的内部困境

1969 年,诺齐克(R. Nozick)在《纽科姆问题和两个选择原则》一文中首次提出了纽科姆疑难,由此引发了关于该疑难的激烈探讨。作为合理决策行为分析中的一个经典案例,纽科姆疑难的表述存在多种版本,本文使用塞恩斯伯里(R. M. Sainsbury)的如下表述:

> 面前有两个盒子 A 和 B,您或者可以把两个盒子都打开,或者只打开 B,您只能获得您所打开的盒子中的东西。
>
> 假设有一个超级生物,它以往对您的行动的预言总是准确的,现在它又遵

收稿日期:2020-12-27
基金项目:国家社会科学基金重大项目(18ZDA031)
作者简介:史红继(1992—),男,河南商丘人,博士生,主要从事现代逻辑与逻辑哲学研究。
①Sorensen, R. A. “Uncaused Decisions and Pre-decisional Blindspots”, *Philosophical Studies*, 1984, 45(1): 51-56.

循如下方式行动完毕:

它已在盒子 A 中放入了一千元现金。并且,

如果它预料您将只打开 B,则它在 B 中又放入了一百万元现金。

如果它预料您将两个盒子都打开,则它就不在 B 中放任何东西①。

相当一部分学者认为,纽科姆疑难由于存在着"超级预言家",所以不应作为日常生活中的决策行为分析案例。然而,这个预言家的设定在该案例中其实无关紧要,如以色列学者盖夫曼(H. Gaifman)所表明的,构造合理行动疑难完全可以去掉超现实假设,即可以运用一系列现实的"公共知识"在"合理行动"或"合理选择"的问题域中构造出严格的逻辑悖论②。因此,本文不再探究"超级预言家"的设定本身是否合理。

诺齐克提出纽科姆疑难的初衷是比较两个判定"合理行动"的不同原则:"最大期望效益原则"(简称"MEU 原则")和"占优原则"(简称"DP 原则")。前者指主体的目标是为了从该行动中获得所期望的最大效益。后者指实施行为 α 是合理的,那么它能够满足两个条件:(a)无论可能发生什么事情,对于行动者而言,实施 α 的后果不会比采取当前可选择的其他行动的后果更坏;(b)在实施 α 的过程中,至少存在一个可能的后果,它要好于选择实施其他行动所得到的后果③。

不过,诺齐克于 1991 年在普林斯顿大学所做的唐纳讲座中表示,在过去的思考中,他未想到要用证据的或因果的决策理论来充分且系统地发展相竞争的各种版本的决策理论。于是,他在随后完成的《合理性的本质》一书中主张:"我们暂时可以把我们的讨论限制在最大化(条件)期望效用的两个原则内,这两个原则分别是由因果决策理论和证据决策理论所表述的。"④因此,不需要基于 DP 与 MEU 两个原则,而是代之以证据决策理论(Evidential decision theory,简称 EDT)与因果决策理论(Causal decision theory,简称 CDT),就可以从纽科姆问题的情境中推导出相冲突的行动选择:

证据决策理论行动 α 的预期效益 = 在行动 α 被实施的情况下,可能状况 S 发生的概率,乘以"可能状况 S 下行动 α 的效益",并将不同可能状况下的这个值加总。

因果决策理论行动 α 的预期效益 = 条件句"如果行动 α 被执行,则导致可能状况 S 发生"的概率乘以"可能状况 S 下行动 α 的效益",并将不同可能状况下的这个值加总。

从 EDT 角度考虑,在两盒选择的情况下,B 中被预言家放入一百万的概率接近 0;而在只选择 B 的情况下,B 中被预言家放入一百万的概率接近 100%,两者概率并不相同,此时明智的行动应当是只打开 B。不过,从 CDT 角度出发,由于无论如何行动,预言家都已经做出了在 B 是否放入一百万的决定且已实施结束,那么就会出现:"如果只选择 B,B 中有一百万"的概率与"如果做两盒选择,B 中有一百万"的概率是相同的。此时,只选择 B 的行为并不能导致 B 中出现一百万的概率增加,因此两盒选择更为合理⑤。

分析纽科姆问题所提及的两个理论路径,可以发现 MEU 与 DP 原则的差异在于,MEU 原则比较的是两个行动间的预期效益的差异,但 DP 原则比较的是在每一个可能情况中两个行动的预期效益的差异⑥。不过,仔细研析不难发现,两盒选择行为的背后存在着这样一个机理:不论何种选择,其行为都不会改变已经确定的事实。从这个角度来看,DP 原则并未完全展示出该观点,而 CDT 可以更好地表述"逆因果"的独立性,进而认为应该两盒选择。此外,EDT 则显然具有更偏爱于只选择 B 的决策的理论特征。所以,才有许多哲学家赞同诺齐克的观点,我们可以将纽科姆疑难的实质归结为证据决策原则与因果决策原则之间的冲突,而它的解决方案则落脚于判断哪种决策原则更为理性。但是,笔者后文的分析将会表明,情况可能比想象中更加复杂。

①Sainsbury, M. *Paradoxes* 3*nd edition*. Cambridge University Press, 1995, p.69.

②Gaifman, H. "Paradoxes of Infinity and Self-applications, I", *Erkenntnis*, 1983, 20(2): 131-155.

③Sainsbury, M. *Paradoxes* 3*nd edition*. Cambridge University Press, 1995, p.74.

④诺齐克:《合理性的本质》,葛四友译,上海译文出版社 2016 年版,第 71-75 页。

⑤王一奇:《另类时空图书馆,假设性思考难题及其解决方案》,台湾大学出版社 2019 年版,第 28 页。

⑥王一奇:《另类时空图书馆,假设性思考难题及其解决方案》,台湾大学出版社 2019 年版,第 27 页。

二　认知盲点理论的改进与应用

索恩森在论及产生悖论的根源时指出,悖论之所以产生,是因为我们把盲点当成了一种正常的陈述。据此,他提供了一个解悖思路:就置信悖论而言,一旦我们意识到盲点必须以一种特殊的方式处理,悖论就不再出现①。然而事实上,在一些情况下,并非把盲点当成了"正常"状态才让我们"获得"悖论,而是在自然语言体系甚至人类认知系统中,盲点本身就是直觉有效的。例如,塔尔斯基(A. Tarski)对说谎者悖论的细致分析表明,人们在推导的过程中不自觉地使用了"T 模式",即 X 是真的,当且仅当 p(其中 X 是语句 p 的名称)。索恩森若强行将人类直觉可信的 T 模式转化为他所认为的"不正常"状态,是难以令人信服的。

虽然索恩森的认知盲点理论并未能完全展现他所宣称对悖论消解方案的效用,但幸运的是,孔斯在研讨盖夫曼悖论以及塞尔顿的声誉悖论时,运用了索恩森的认知盲点理论,并对此理论进行了优化。孔斯表明,盖夫曼悖论、塞尔顿的声誉悖论以及纽科姆疑难等,看似相互独立,但实则具有相似的结构和机理②。因此,笔者将基于认知盲点理论的这种优化版本,对纽科姆问题进行精确塑述,进而为后续的分析奠立基础。

在谈及说谎者悖论时,索恩森认为自我指称对其既不是必要条件,也不是充分条件,说谎者悖论的建构可以建立在雅布鲁悖论的基础上,由一系列无限的句子构成,且不涉及自我指称。同时,他基于认知盲点理论所提出的解悖方案亦不涉及自我指称,因为他坚信自我指称在悖论中没有重要作用③。不过,乔格灵(T.B. Jongeling)和凯策尔(T. Koetsier)已经证明了索恩森的这个观点是错误的,并得出结论:一个悖论,要么是无限的,要么至少包含一个自我指称周期。这个结论很容易直观地理解,如果推理链条是有穷次的,它以具有明确真值的语句而结束,从终端结论出发,其他语句的真值可以通过反向推理来确定;只有推理链条是无限的时候,才会出现问题。所以有穷次的悖论总是涉及自我指称,即便这些悖论性质不同。很明显,那些导向矛盾行为的盲点,总是涉及自我指称。盲点"p 和'a 不知道 p'"可以理解为 a 的知识,我们称之为 a 的知识存储,这意味着 p 不属于 a 的知识,或者 p 不是在 a 的知识存储中的知识。一旦 a 接受了该语句,它就变成了自我指称,因为它涉及到 a 的知识储备,而它本身就是知识的一个元素④。纽科姆疑难的有穷次的悖论的拟化形式,决定了它也具备该特性。缘此,笔者将利用盖夫曼式构造直观地展现这一点,并以此为基础进一步分析纽科姆问题的根源所在。

在索恩森认知盲点理论的观照下,我们可以这样理解 EDT 与 CDT:对某些支持 EDT 的人来说,CDT 是盲点。该盲点具有这样的属性,即 CDT 是真实可靠的;同时,支持 EDT 的人可能不知道它是真实的。换言之,一旦一个支持 EDT 的人在他的推理中接受了 CDT,他将不得不停止相信他据以推断 CDT 的信息,但他也无法断定这些信息是错误的。不过,也还存在着这样一种可能,即涉入纽科姆情境的人在行动开始之前,就已经处于与决策相关的盲点中了,这也是孔斯所指出的索恩森盲点理论的不足之处。孔斯认为,认知盲点理论不应局限于只有当行动者采取行动后,才使得参与一方处于盲点的情形,在行动之前,行动人可能就已经陷入了盲点状态⑤。

借鉴孔斯对塞尔顿声誉悖论的相关讨论,我们可以通过如下过程将纽科姆疑难塑述为一个严格悖论——两盒选择是不合理的,当且仅当,它不是不合理的:

令 J_ip 代表主体 i(i=e,c)对命题 p 的合理信念,其中 J 代表相信算子。并用 Q 表示如下虚拟命题:假如支持 EDT,那么 CDT 是不合理的。为了更清晰地刻画纽科姆疑难的形式,我们假设参与人相信 $Q \leftrightarrow \neg J_eJ_eQ$,也就是说,CDT 是不合理的,不应该取两盒选择,除非他相信参与人相信 Q。如果支持 EDT 的人也相信:支持 CDT 的人相信 J_eQ,当且仅当,支持 EDT 的人也会合理地相信它,那么支持 EDT 的人就会得出 $Q \leftrightarrow \neg J_eJ_eQ$。如孔斯所言,任何一个形如 $J_e(Q \leftrightarrow \neg J_eJ_eQ)$ 的命题都与合理信念理论中的一个高度可信的公理或

①Jongeling, T. B. Koetsier, T. "Blindspots, Self-reference and the Prediction Paradox", *Philosophia*, 2002, 29(1-4): 377-391.

②Koons, R. C. P*aradoxes of Belief and Strategic Rationality*. Cambridge University Press, 1992, pp. 13-39.

③Sorensen, R. A. "Yablo´s Paradox and Kindred Infinite Liars", *Mind*, 1998, 107(425): 137-157.

④Jongeling, T. B. Koetsier, T. "Blindspots, Self-reference and the Prediction Paradox", *Philosophia*, 2002, 29(1-4): 380-381.

⑤Koons, R. C. P*aradoxes of Belief and Strategic Rationality*. Cambridge University Press, 1992, p.38.

规则不相容[①]。通过利用盖夫曼的相关理论,可以更为直观地表明:由于一些特殊情境的限制,当主体持有 $J_e(Q \leftrightarrow \neg J_eJ_eQ)$ 的信念,则会与其自身信念库中的合理信念不相容。

盖夫曼在说谎者悖论的研究中,提出一个新颖的解决方案。他构造了一种算法为语句殊型或其所谓的"指针"(pointers)之网进行赋值。一个殊型集是一个有向图(directed graph),其中的节点都是语句殊型,而有向边则代表调用关系(calling relation)[②]。为了简单起见,假定认知主体的信念可以映射在语句的语法结构上,按照盖夫曼式的结构,殊型 p、q 与类型 $[J_cp]$ 的子殊型就形成一个闭环(如图 1 所示),这表明纽科姆疑难具有自我指称的特性。

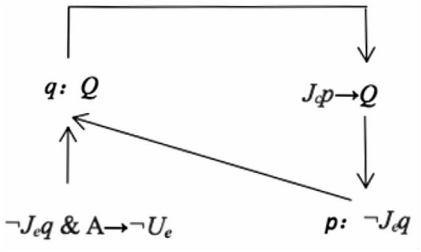

图 1　纽科姆疑难殊型图

另外,如图 1 调用关系所刻画的形式所示,纽科姆 CDT-EDT 模型的思想殊型形成了一个有向网。图中顶行的殊型都是支持 EDT 的人的潜在信念,底行的殊型则是支持 CDT 的人的潜在信念。EDT 类型的人倾向于接受类型为 $[J_cp\rightarrow Q]$ 的殊型,即 CDT 的人亦相信 Q,因为主体都是理性的人,应和他们一样。然而 CDT 类型的人倾向于接受 $[\neg J_eq\&A\rightarrow\neg U_e]$ 的殊型,该殊型表示不支持 EDT 相信的 Q,且倾向于认为理性主体 a 不会得到 EDT 类型行动的最大结果 U_e(1 001 000 元),因为预言家总是准确的。因此,支持 EDT 的这个主体被阻遏接受虚拟条件句 Q,而支持 CDT 的主体则被阻遏识别出支持 EDT 类型的该主体不接受 Q。支持 EDT 类型的人与支持 CDT 类型的人都受困于情境导致的盲点,没有哪一种类型的人可以接受对方的决策,即便预先获知了对方的结论以及所得出结论依据的信息亦如此。

正如孔斯所指出,按照认知盲点理论的分析程度,我们尚难以对声誉悖论盖棺论定[③]。而笔者认为,在纽科姆问题上亦是如此。尽管从索恩森—孔斯的认知盲点理论出发,可以发掘出现存于该推理链条上的自我指称(闭环推理),但是这并未指明最终的解悖方案。不过值得肯定的是,作为决策理论的经典案例,纽科姆疑难不仅涉及预言判断的因素,也涉及有穷次的行动决策,这意味着,结合认知盲点理论,我们的确可以在一定程度上刻画出"疑难"之所在。

三　纽科姆疑难的实质与冲突的化解

(一)纽科姆疑难的悖论实质

如前所述,纽科姆问题之所以让我们深感疑惑,是因为面对决策场景时,我们的认知出现盲点:当接受 EDT 时,只选择 B 具有最大效益;而认识到 CDT 时,又可合理排斥逆向因果,进而得出两盒选择才有最大效益。在此,我们有必要进一步追问,CDT 与 EDT 两者背后的信念或者推理机制是否内在地要求可以在任何时候被同一主体应用?倘若两者在某些情况下并不能够被同一主体使用,而只是允许在不同情况下分别使用,那么对于纽科姆疑难而言,着眼于确定更为合理的决策原则的解悖思路似乎将迎来曙光,因为它只要求我们聚焦于纽科姆问题本身即可。

纽科姆疑难作为语用悖论群落中的经典案例之一,其拟化形式的构造虽非严格意义上的悖论,但是以往的悖论研究所总结的规律仍可以在方法和思路上启发我们。张建军曾总结和提炼了逻辑悖论的"三要素",即(1)公认正确的背景知识;(2)严密无误的逻辑推导;(3)能够建立矛盾等价式[④]。笔者认识到,对悖论构成要素的这种揭示,既有助于澄清悖论的成因和分类依据,又指引解悖的理路,甚至也可以用来评估解悖方案的完备程度。在此理论背景下重新审视纽科姆问题,我们可以做出如下分析。

首先,根据前面的讨论,要素(2)和(3)似乎很容易满足。令在只选择 B 情况下,B 中有一百万的概率为 $\alpha(50\%<\alpha<100\%)$,那么在两盒选择情况下,B 中有一百万的概率为 $(1-\alpha)$。按照

①Koons, R. C. Pa*radoxes of Belief and Strategic Rationality*. Cambridge University Press, 1992, p.27.

②Koons, R. C. Pa*radoxes of Belief and Strategic Rationality*. Cambridge University Press, 1992, p.101.

③Koons, R. C. Pa*radoxes of Belief and Strategic Rationality*. Cambridge University Press, 1992, p.38.

④张建军:《逻辑悖论研究引论(修订版)》,南京大学出版社 2014 年版,第 7 页。

EDT,由于 1 000 000 * α>1 000+1 000 000(1-α),选择只打开 B 可获得最大效益;按照 CDT,不论何种情况,由于预言者已完成放置,B 中放有一百万的概率与 B 中未放的概率都是相同的,令其为 β,则 1 000 000 * β<1 000+1 000 000β,此时两盒选择可获得最大效益。因此,经过严密推导可以得出矛盾等价式:只选择 B 是合理的,当且仅当,它不是合理的。如前所述,借重索恩森—孔斯的认知盲点理论,为做逼近性刻画,我们已经得到 $J_e(Q \leftrightarrow \neg J_eJ_eQ)$ 这一与合理信念理论中的公理和规则不相容的命题形式①。

然而,如"三要素"的理论洞见所指明的,要素(2)与(3)的成立,应明确是在要素(1)的基础上而言的,即 CDT 与 EDT 对于应用到纽科姆问题上,是"公认正确的背景知识"。"公认正确的背景知识"这一要素不仅显示了纽科姆悖论之形成的前提,亦包含了认知共同体所使用的逻辑法则,否则"严密无误的逻辑推导"以及后续的"能够建立矛盾等价式"就无从谈起了②。

事实上,大多数悖论或拟悖论的出现都是由于使用了认知共同体可能未完全意识到但却已"集体默认"的公共信念,纽科姆问题亦是如此。这一点从"疑难"形成的最初源头亦可以省悟到:纽科姆问题之所以形成了"自我指称"的闭环,其实源于我们最初默认了 CDT 与 EDT 可以同时应用到同一个行动,以及两者期望值本应该相同。通过之前的分析表明,将预设的背景知识当作"完全无误"会促使认知盲点产生。然而,背景知识只是代表某一特定阶段认知共同体认识所达到的水平,那些相对于特定认知主体的"公认正确的背景知识"未必一定是正确的,是可以修正的。这也是纽科姆问题具有"可解性"的基本依据。

(二)CDT 与 EDT 之冲突的消解

如果按照第二节所述的 CDT,那么需要预设:行动结果的实现不会与选择实现的行动的方式有所关联,而仅仅被行动本身所影响,这样才能够保证 CDT 所衍生出的模型是唯一的。当然,这种预设单纯地将决策行动的思考视为行动与结果之间的因果效应③,不仅忽略了决策行动的多重可实现性(multiple realizability)④,而且还忽略了每个行动的不同实现方式,这将会引入不同的影响因子,进而造成不同的决策因果结构⑤。可以说,CDT 只考虑了一阶决策,即在从事决策时,不将实现行动的方式作为考虑因素之一,而纯粹考虑行动本身,进而认为逆因果关系不成立、EDT 不合理。而正如希契科克(C. Hitchcock)所指出,主体在行为层面上是否真的可能进行干预是存在争议的,尤其是在纽科姆型案例中因果模型已经指定了诸如意向和决策因素之类的心理状态进入决策的方式⑥。换言之,主体在进入决策场景时,理性决策不只是思考行动本身从而在诸多行动选项中选择哪一个,同时也会思考行动的实现方式。比如,假设我们计划从厦门到台湾旅游,有坐大巴、坐轮船、坐飞机 3 个行动选项,每个行动选项分别有拼车(船、机)、包车(船、机)2 种乘坐方式,那么理性决策的问题就转化为从 3 * 2 种行动做法中找出最有利的,而非只从 3 个行动选项中找出最有利的。同理,当置身于纽科姆决策场景之时,并不是先决定理性行动完毕之后,再考虑决定实施行动的方法,而是在对行动做出理性决策时,同时需要考虑实现行动的方式,因此,整个过程是一个二阶决策。

伊根(A. Egan)曾用"泰德邦迪的怨念"建构了 CDT 的反例,他宣称 CDT 并不能捕捉到何为"会有最好结果的行动",而只是获得了"现有的因果结构下的最好期望值行动"⑦。更进一步来说,当被考量的行动与未来状况 S 之间,存在着共同的因果影响因子(common causal factor,简称 C)时,则可能状况 S 的概率,会因 C 的存在而有所变化。比如,在"泰德邦迪的怨念"的例子中,保罗是否是心理变态这个共同影响因子的概率,会影响到他是否决定按下按钮,即便他本身知道按下与否的行为并不能"因果地"影响他是否为心

①Koons, R. C. *Paradoxes of Belief and Strategic Rationality*. Cambridge University Press, 1992, pp.23-27.

②关于明确悖论构成的第一要素中包含认知共同体所使用的逻辑法则的重要性,参见张建军:《再论广义逻辑悖论的基本构成要素》,《南国学术》2018 年第 1 期。

③p 对 q 具有因果效应,当且仅当,对于 p 中的变元的值做出改变,且其他条件不变的情况下,q 中某些变元的值有相应的改变。

④同样的行动目的,实现的方式可能有多种。

⑤王一奇:《另类时空图书馆,假设性思考难题及其解决方案》,台湾大学出版社 2019 年版,第 286 页。

⑥Hitchcock, C. "Conditioning, Intervening, and Decision", *Synthese*, 2016, 193(4): 1157-1176.

⑦Egan, A. "Some Counterexamples to Causal Decision Theory", *Philosophical Review*, 2007, 116(1): 96.

理变态的概率①。

当主体被置于决策前的特定环境,且行动结果只与行动本身有关时,使用 EDT 所计算的可能状况 S 发生的条件概率与使用 CDT 的计算结果相同,这符合人们以往的观点:CDT 与 EDT 两个基本决策原则在计算行动期望值无差异,都可使用。不过,当行动结果受到行动本身与额外存在的共同因子的影响时,CDT 作为决策原则将不能发挥积极作用,但 EDT 却是可以的。因为,EDT 所计算的值不仅局限于行动本身对结果的因果效应,还囊括了引发行为的思考或原因在因果效应上带来的所有可能结果的总和。例如,豪斯曼(D. M. Hausman)曾给出这样的例子:

> 核电站的工程师们关注这样的问题:如果蒸汽管会爆炸,那么核反应堆会随之关闭②。

当利用蒸汽管爆炸来预测核反应堆关闭时,若是存在着共同因子,比如发生了地震 ,或是阴谋破坏, 或是压力过强等使得蒸汽管爆炸,并且一定程度上影响到核反应堆的关闭,那么,EDT 会将共同因子作为考量因子之一,加入到对蒸汽管爆炸所产生的所有可能结果的条件概率计算之中。因此,相较于只局限于特定的决策因果结构的 CDT,EDT 在任一个恰当的决策因果结构中,都可以正确地捕捉到产生最好结果的行动③。

意识到 EDT 所计算的值并不必然等于仅基于行为而引发的可能状况 S 的期望效益,乃是十分关键的。这意味着,在纽科姆场景中,主体理性决策的行动实现首先需要甄别完美干预行动是否可能,而对选择之效果的思考,则使得非完美干预的行动实现个例存在。对选择之效果的思考,就是一种假设性思考,即如果选择的行为方式不同,是否影响预言家在 B 中放(或不放)一百万元。假设性思考是人类更好地适应大自然不可或缺的能力,也是甄别不同行动在环境中是否优劣的有效手段。宇德科夫斯基(E.Yudkowsky)和索雷斯(N. Soares)认为:“因果决策理论的主体之所以失败,是因为他们的想象打破了行动与环境之间的太多关联。”④而主体的想象欲要实现行动与环境之间的关联,就离不开假设性思考。而且,从日常决策案例中不难发现,假设性思考发生在行动之前,说明它不是关于行动与未来行动结果在时间线上逆因果的产物。下面对假设性思考结构的刻画也将揭示这一点。

假设性思考是一种想象力的操作,人类通过想象将某个“行动方式”作为前件放入到信念系统中,然后在那种想象的情境下,生成“假设 m”信念相对应的经验现象 M,再与后件 n 的经验现象 N 作对照。其中 m 代表某个行为方式,n 代表未来可能状况;M 与 N 分别代表过往行动中的一些经验现象。在此,笔者借助克拉克(Clark)和马歇尔(Marshall)从刘易斯的书中提取出的模式,来更好地刻画假设性思考的结构。我们假设某一命题φ是某特定团体成员共同知道的心智状态,且简化模型使得团体成员只有 A 与 B 两人。那么,

A 和 B 交互地知道φ,当且仅当,某种事态 G 成立,同时:

> (1)A 和 B 都有理由相信 G 成立。
>
> (2)G 向 A 和 B 显示,双方都有理由相信 G 成立。
>
> (3)G 向 A 和 B 显示φ⑤。

此时,A 与 B 代表认知主体,事态 G 代表“假如经验现象 M,则经验现象 N”,它可连接到现实经验的信念系统中,φ代表假设性思考“假如 m,则 n”。若主体在行动前,将φ作为选择行为方式的信念,而φ的可信程度来源于现实生活中以往行动经验中的统计数据,我们当然可以合理地认为,引发行动结果的影响并非源于“超时空扭转”。

通过运用二阶决策的假设性思考刻画纽科姆疑难,我们会得到:$P(S|do(B)) \neq P(S|B)$,其中 $P(S|do(B))$ 代表对“只打开 B”做 do 运算后获得一百万元的概率。do(B)代表干预其他因素对选择 B 的因果影响,保证决策行动——选择 B 在因

①王一奇:《另类时空图书馆,假设性思考难题及其解决方案》,台湾大学出版社 2019 年版,第 267-281 页。

②Hausman, D. M. *Causal Asymmetries*.Cambridge University Press Cambridge, 2011, p.121.

③王一奇:《另类时空图书馆,假设性思考难题及其解决方案》,台湾大学出版社 2019 年版,第 280 页。

④Yudkowsky, E. and Soares, N. “Functional Decision Theory: A New Theory of Instrumental Rationality”. *arXivpreprint arXiv*:1710.05060 (2017).

⑤Clark, H. H. and Marshall, C. R. “Definite Reference and Mutual Knowledge”, In *Elements of Discourse Understanding* (ed. Aravind, K. and Ivan, A. Sag *et al*). Cambridge University Press, 1981, pp.10-33.

果框架中是独立变元,而无任何先决条件。P(S|B)代表在“只打开 B 盒子”的情况下,盒子有一百万元的概率。之所以出现这样的结果,是因为 CDT 只是单纯考虑行为本身对结果的因果效应;而相形之下,无论行动方式是否为行动与结果之间的因果影响因子,EDT 都可以作为主体的决策理论。考虑到行动方式在因果效应上对结果带来的影响,CDT 应用于纽科姆问题将不再适宜。这意味着,我们可以在纽科姆疑难的情境中通过排除 CDT 的合理应用而消解 CDT 与 EDT 之间的冲突。

四　余论

如果 EDT 可以作为应用于纽科姆案例的适当原则,那么接下来的工作便是在这一案例中确定合理的干预位点。但笔者同样看到,在纽科姆问题的设定并未编码我们行为的确定特征的情况下,为了确保行为受到规范性理论的约束,理性的建议是将干预点设在一个人的生理状况的下游和决策行动的上游。然而,这种包含了多个可能的干预点的方式,凸显出一种理性的悲剧:理性行动未必是具有理性特性的主体所执行的行动,一个非理性的主体可能由于性格、意向等特征,恰好选择了理性的行动,而这些问题只有在多个位点声称可以干预时才会出现①。刘易斯(D. Lewis)认为,这场悲剧只是不完美世界中的一部分。有时,世界对非理性的人的奖励并不意味着他们是理性的②。无论如何,在对某个变量进行干预,以便于获知因变量与结果之间的因果效应的过程中,所选择的干预位点不应该是行为,因为理性主体不会在行动之时才试图真正理性。如格林(P. Greene)所指出的,这场悲剧的出现,应使我们重新思考我们的决策理论。如果行动的时刻是唯一可以想像的干预点,那么传统的因果决策理论是可信的。但事实上,其他的干预点是可能的③。因此,在纽科姆故事的背景下,理性的决策只是依赖于恰当的因果结构。在把决策行为视为完美干预的情况下,我们应该做两盒选择。不过,该结论的成立,并非与 CDT 有关,而是与决策因果结构的特性有关,该特性表示纯粹决策因果结构下的行动,无任何先决条件,即决策行动在因果模型中是独立变元,并且这里的干预充其量也只是近似地代表真正的干预。

Analysis of Newcomb's Problems from the Perspective of Epistemic Blindspots Theory

SHI Hong-ji

(Department of Philosophy, Nanjing University, Nanjing 210023, China)

Abstract: Since Nozick raised Newcomb's problem, there has been a fierce debate about its basic crux and its way out. Applying the Epistemic Blindspots Theory proposed by Sorensen and improved by Koons, et al. can not only reveal the essence of Newcomb's problem, but also point out a new way to resolve the conflict between causal decision theory (CDT) and evidence decision theory (EDT). In the context of the Newcomb story, although EDT is superior to CDT due to considerations involving the realization of choice behavior, this does not mean that EDT can completely solve the problem. When the decision-making behavior is regarded as a perfect intervention, the best solution to Newcomb's problem is to choose two boxes, but the choice is based only on the characteristics of the decision-making causal structure, not CDT.

Key words: Newcomb's problem; behavioral decisions; epistemic Blindspots; paradoxes

(责任校对　朱正余)

①Easwaran, K. “A classification of Newcomb problems and decision theories”, *Synthese*, 2019,196(4):1-20.

②Lewis, D. “Why ain'cha rich?”.*Noûs*, 1981, 15(3): 377-380.

③Greene, P. “Success-first Decision Theories”, In *Newcomb's Problem*(Classic Philosophical Arguments) (ed. A. Ahmed). Cambridge University Press, 2018, pp.115-137.