哥梯尔的“协议道德”理论评析 [On Gauthier s Theory of Morals By Agreement]

Item Type	Article
Authors	陈, 真
Publisher	河北省社会科学院
Rights	With permission of the license/copyright holder
Download date	2026-07-02 07:45:30
Link to Item	http://hdl.handle.net/20.500.12424/185674

陈真：哥梯尔的“协议道德”理论评析

陈真

哥梯尔的“协议道德”理论评析

陈 真

美国韦恩州立大学哲学系

哥梯尔(David Gauthier)在《协议道德》（*Morals By Agreement* (New York: Oxford University Press, 1986)。）一书中，试图从个人利益出发，即从自利理性的原则出发，推导出道德的原则。这样的道德当然完全符合自利理性。具体地讲，哥梯尔极力想证明：第一，有理性的个人在相互交往中，遇到类似“囚徒困境”情景中的次佳化问题时，愿意接受公正的、不偏不倚的限制性条款（即道德的原则），用以限制个人无止境地追求个人利益，从而避免次佳化问题，并实现共同的利益。第二，一旦达成限制性条款或协议，各方遵守条款或协议是符合理性的，即符合各方的个人利益的。第三，在形成关于如何分配合作利益的协议时有两条原则必须遵守：其一，洛克条款，即禁止在谈判之初损人利己，使对方处于不利的谈判地位，而使己方处于有利的谈判地位。其二，最小最大量相对让步原则，即关于分配合作利益的协议或公正的限制性条款仅当它使各方的最大相对让步减至最小时才是可接受的。第四，这些限制性的条款就构成了和日常道德不同的理想道德的原则。就道德能否从自利理性中推导出来的问题而言，前两项任务最重要。那么，哥梯尔是否完成了前两项任务呢？

哥梯尔称他所采纳的理性理论为“最大限度的理性概念”（the maximizing conception of rationality）。按照这一概念，如果采纳行动 A 能够最大限度地实现行动者甲某所选择的目的，甲某就有充分的理由采取行动 A，即甲某采取行动 A 是符合理性的。(参见David Gauthier, *Morals by Agreement*, 22。)哥梯尔对甲某所选择的内容毫无限制，故可以包含许多不同的理性原则。甲某可以选择个人利益作为其行动的最终目的，也可以选择共同利益作为其行动的最终目的。这两个目的是不同的，故所代表的理性原则也不同。而哥梯尔想证明行动者选择共同利益作为最终目的更合乎理性。众所周知，在“囚徒困境”的情景中，不合作、告密、违反协议最符合行动者的个人利益。哥梯尔认为，在这种情景下，选择个人利益为最终目的的行动者会面临一个问题：合作、遵守协议能够产生“合作盈余（a cooperative surplus）”（即合作所产生的利益大于或不小于不合作给各方所带来的利益）(参见*Morals by Agreement*, 141。)，而不合作、违反协议却不能。哥梯尔认为这足以让追求个人利益的行动者感到不安。这就迫使他们达成协议，同意对个人无限追求个人利益的行为加以限制。哥梯尔的论证可以表述如下：

前提一：如果合作能够产生合作盈余，理性的最大限度追求者就有充分的理由接受合作的协议，以限制他们无限追求个人利益的行为。

前提二：合作能够产生合作盈余。

结论：理性的最大限度追求者有充分的理由接受合作的协议，以限制他们无限追求个人利益的行为。(参见 David Gauthier, “Morality, Rational Choice, and Semantic Representation: A Reply to My Critics,” *Social Philosophy & Policy*, Vol. 5, No. 2 (1987): 175-176。)

这是一个有效的论证，即前提真，结论一定为真。对前提一，我毫无异议，因为如果合作能够保证，合作盈余也能得到保证，这将符合每个行动者的个人利益，理性的最大限度追求者当然有充分的理由合作。问题是前提二是否为真。合作盈余的实现取决于合作能否实现和保持。合作能否实现和保持取决于行动者有无充分的理由遵守协议或保持合作。如果行动者并无充足理由遵守协议，保持合作，那么合作就无法实现。而合作盈余就只能是“水中月，镜中花”。因此，关键在于遵守合作协议能否得到保证。如果能，遵守协议就符合理性，则合作以及合作盈余就能实现。反之，遵守协议则不符合理性，合作或合作盈余就无法实现。

霍布斯解决协议遵守问题的办法是设立一个具有绝对权威和力量的君主或政府来迫使各方遵守合作协议。哥梯尔不满意这种解决方案，他想证明：单凭理性的力量，无需任何外部的力量，就足以证明遵守协议的合理性。为此，他引进了一个新的理性概念，即“有限制的最大限度”（constrained maximization）。哥梯尔认为有两种“最大限度”的理论：直截了当的最大限度（straightforward maximization）和有限制的最大限度。它们可以分别定义如下：

直截了当的最大限度理论认为：甲某有充分的理由采取行动 A，当且仅当，采取行动 A 能最大限度地、最大可能地实现甲某基于个人利益的选择。

有限制的最大限度理论认为：甲某有充分的理由采取行动 A，当且仅当，采取行动 A符合合作协议，并且协议各方遵守协议，即使采取行动 A 并不能最大限度地实现甲某基于个人利益的选择。

直截了当的最大限度理论是西方公认的理性决策理论，有限制的最大限度理论则是哥梯尔自己的理性理论。遵守协议的合理性只能建立在后者的基础上，而非前者。后者与哥梯尔最初提出的理性的最大限度原则有何不同呢？理性的最大限度原则有两种解释：一种是利己主义的解释，即行动者最大限度追求的是她自己的个人利益。直截了当的最大限度理论采取的是利己主义的解释。另一种是非利己主义的解释，即行动者最大限度追求的是她的人生理想。如果哥梯尔选择第一种解释，他将无法推导出为什么要对最大限度的追求个人利益的行为加以限制，因为限制和最大限度的追求个人利益是不相容的。他只能选择第二种解释。非利己主义的最大限度理论对人生的理想又有两种可能的解释：一种对人生的理想的解释将限制性的条件与个人的良心、道德感等“内在约束力”联系起来。一种对人生理想的解释并不把良心、道德感等因素考虑进去，限制性的条件纯粹源于个人的人生理想。哥梯尔的人生理想指的是后一种人生理想。对于人生理想的具体内容，哥梯尔没有任何规定。人们可以将直截了当的最大限度理论包含在自己的人生理想之中，也可以将有限制的最大限度理论包含在自己的人生理想之中。哥梯尔必须证明，为什么人们必须将有限制的最大限度理论包括进自己的人生理想之中，即为什么有限制的最大限度理论优于直截了当的最大限度理论。

哥梯尔的证明大体如下：有限制的最大限度追求者，即奉行有限制的最大限度理论者，比无限制追求最大限度者，总体上，更能最大限度地实现其期望的功利值。其原因在于一个人的行为倾向或品行是可以为人所知的，即对他人来说是透明的。因而，有限制的最大限度追求者在社会上更受欢迎，能吸引更多的合作者，从而比无限制最大限度追求者更能从合作中获得好处。哥梯尔认为，甚至在“一次性的囚徒困境”情景中，遵循有限制的最大限度的理性原则也是合乎理性的。(参见 David Gauthier, “Uniting Separate Persons” in *Rationality, Justice and the Social Contract*, ed. David Gauthier and Robert Sugden (New York: Harvester Wheatsheaf, 1993), 185-187。)问题是，在“一次性的囚徒困境”情景中，如果违反协议能给行动者带来更多的期望功利值，行动者为何还要遵守协议呢？

假定有两位农场主甲和乙处于自然状态下（即处于没有警察、军队等维持治安的国家机器的状态下）。由于他们经常受到贪婪匪徒的威胁和袭击，甲和乙达成一个防御协定：当一方遭到袭击时，另一方应该及时援助。假定甲方遭到匪徒袭击，乙方面临两种选择：或遵守协

定，援助甲方，或违反协定，按兵不动。作为一个有理性的人，乙应该怎样做才最合乎自己的利益呢？假定他们从行动中所得到的益处可以用期望功利值（expected utilities）来衡量，则其行为的后果以及可能给各方带来的利益可以用下图来表示：

		甲	
		遵守协定	违反协定
乙	遵守协定	6 6	10 0
	违反协定	0 10	1 1

图中数值代表了定约人不同行为（遵守或违反协定）对各自产生的期望值（即各自可能得到的好处）。数值越高，则益处越大，因而定约人更有可能采取相应的行动。其中每个方框中左下角的数值代表了乙可能得到的益处，每个方框中右上角的数值则代表了甲可能得到的好处。

如图表所示，如果甲乙双方都遵守协定，则各自所获的期望值为 6。如果甲方违反协定而乙方遵守，则甲方所获的期望值为10，而乙方则为 0，反之亦然。如果双方都违反协定，则双方所得到的功利值为 1，低于双方都遵守协定所得到的功利值 6。

如果这只是“一次性的囚徒困境”情景，则甲乙都应选择违反协定才是合乎理性的。因为，不论另一方采取什么行动，违反协定的好处总是大于遵守协定的好处。如果对方遵守协定，则违反协定可使己方获得最大的好处，即获得功利值10。如果对方也违反协定，则己方违反协定可以避免最糟糕的情形，即一无所获的情形出现。尤其是，如果己方后于对方行动，则违反协定肯定最符合己方的利益。

让我们假定甲某在上一次袭击中遵守诺言，帮助了乙某抵抗了袭击。又假定，这一次匪徒们袭击甲某。再假定，乙某知道政府将很快派出军队清剿匪帮，这样乙某再也不需要甲某的帮助，因为匪徒即将被消灭。那么，按照有限制的最大限度原则，乙某遵守协议，援助甲某是合乎理性的，因为乙某知道甲某在上次袭击中帮助了他，虽然乙某这样做并没有最大限度地扩大个人的利益或个人的期望功利值。那么，为什么乙某这样做是合乎理性的呢？哥梯尔认为，这是因为保持有限制追求最大限度的品行是合乎理性的。但为什么保持这种品行是合乎理性的呢？哥梯尔证明如下：

前提一：如果这一次乙某保持遵守协议的行为倾向（即品行）是不符合理性的，那么上次在他遭受袭击前保持遵守协议的倾向也是不符合理性的。

前提二：如果上次在他遭受袭击前保持遵守协议的倾向是不符合理性的，那么乙某不通过自己的行为倾向让甲某知道他是一个有限制的最大限度追求者是合乎理性的。

前提三：如果乙某不通过自己的行为倾向让甲某知道他是一个有限制的最大限度追求者是合乎理性的，那么乙某通过自己的行为倾向使甲某在上次袭击中不遵守协议援助他也是合乎理性的。

前提四：但是，乙某通过自己的行为倾向使甲某在上次袭击中不遵守协议援助他是不合乎理性的。

结论：因此，这一次乙某保持遵守协议的行为倾向是符合理性的。(同上，186。)

哥梯尔关于在“一次性囚徒困境”情景中保持有限制追求最大限度的品行是合乎理性的论证是难以让人信服的，因为前提一为假。理由在于：这一次乙某是处于“一次性的囚徒困境”情景中，而上一次乙某是处于“重复性的囚徒困境”情景中。这一次，即使乙某遵守协议，表现出有限制地追求最大限度的倾向，他也无法从将来的合作中获得好处。而且，即使他不遵守协议，他也不会因此而受到惩罚。故他没有理由表现出遵守协议的倾向。相反，这一次改变他的有限制地追求最大限度的倾向是合乎理性的。而上一次，乙某通过遵守协议表现出有限制地追求最大限度的倾向是合乎理性的，因为这样做他可以从将来的合作中获得好处，而且可以避免不遵守协议可能受到的惩罚。

在“重复性的囚徒困境”情景中(决策的情景只是一系列的“囚徒困境”情景中的一环，而当事人或行动者并不知道究竟有多少类似的情景还会发生)，亦即正常的情况下，哥梯尔关于有限制地追求最大限度的理论似乎更为成功。有人也许会认为，即使是在“重复性的囚徒困境”情景中，如果甲某是一个无可救药的“急功近利者”，而乙某并不知道甲某是否可以救药，对乙某而言，遵守协议，保持有限制地追求最大限度的倾向，也不符合理性。但这一反驳假设了两个条件：其一，游戏中只有两个行动者；其二，其中一个是无可救药的“急功近利者”（即所有其他的交往者都是无可救药的“急功近利者”）。如果游戏中的交往者超过两个，并且并非所有的交往者都是无可救药的“急功近利者”，则保持有限制追求最大限度的品行或倾向优于无限制地追求最大限度的品行或倾向。

让我们在农场主的例子中再加上另一个农场主丙某。假定丙某是一个有限制的最大限度追求者，他还没有加入甲某和乙某的防守协议。又假定甲某是一个无可救药的“急功近利者”，但乙某是一个有远见的理性主义者，即有限制的最大限度追求者。乙某可以通过一到二次合作的代价弄清甲某是一个无可救药的“急功近利者”。这样，乙某就会放弃与甲某的合作，转而寻求新的合作者。在我们设定的情况下，即寻求同丙某的合作。丙某非常愿意同乙某签订防守协议，而不是甲某，因为乙某已经证明自己是一个有限制的最大限度追求者，而甲某则证明自己是一个不值得信任的、无限制追求自己利益的利己主义者。这样，乙某和丙某就会达成新的防御协定，并且通过相互合作，分享合作的好处。而甲某则被排除在合作之外，无法分享合作的好处。这样，从长远的观点来看，乙某和丙某从相互合作中所获得的好处要远远大于甲某从利己主义的行为中所获得的好处。总的说来，一个人有充足的理由将有限制的最大限度原则包括在自己的人生理想中仅当在实现他的人生理想的过程中所得到的好处（即与有限制最大限度追求者打交道所获得的好处）大于他的损失（与急功近利主义者打交道所遭受的损失）。社会中有限制最大限度追求者越多，一个人就越有理由成为一个有限制最大限度追求者。哥梯尔认为，有限制最大限度追求者从合作中所得到的好处大于所得到的损失，故保持一个有限制最大限度追求者的倾向或品行是合乎理性的。

哥梯尔的这一结论得到了某些计算机实验的支持。1979年，有一位名叫罗伯特·艾克斯洛德（Robert Axelrod）的政治科学家，主持了一系列的比赛，以探讨合作的逻辑。他要求人们提交能够在游戏中彼此对抗的计算机程序。在一次比赛中，有一位据说是世界上最了解“囚犯两难”问题的政治科学家，名叫阿拉托·拉帕帕特（Anatol Rapoport），提交了一个“针锋相对”（Tit-for-tat）的程序。这个程序规定：第一次与对方玩游戏时都与对方合作。从第二次开始，只是重复对方上一次的行为。艾克斯洛德要求参赛者击败“针锋相对”的程序。62个程序参加了比赛，也许包括“急功近利者”或“纯道德者”的程序，但最后的胜利者仍然是“针锋相对”的程序。

艾克斯洛德在解释“针锋相对”程序成功的原因时说：“‘针锋相对’的程序之所以取得如此的成功是因为她将与人为善，以牙还牙，宽恕原谅和清楚明白结合起来。她的‘与人为善’使她避免了不必要的麻烦。她的‘以牙还牙’使对方不敢轻易背信弃义。她的‘宽恕原谅’有助于恢复相互合作。她的‘清楚明白’使对方能够明白她的态度，从而有助于建立长期的合作。”(Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984). 转引自 Matt Ridley, *The Origins of Virtue* (New York: Viking Penguin, 1997), 60-61。)

综上所述，哥梯尔关于道德和遵守道德的合理性在“重复性的囚徒困境”情景中得到了很好的证明，而“重复性囚徒困境”更接近日常实际情况。虽然如此，哥梯尔的理论仍然有其局限性。

首先，哥梯尔将所有道德原则都归结为“协议原则” 或“协议道德”，但并非所有的道德原则都可以归结为“协议道德”。即使他的论证完美无缺，也只能证明一部分道德的合理性。约翰·罗斯(John Rawls)提出的假设性社会契约论在一定的程度上可以克服这一缺陷。

其次，即使这一部分道德原则的合理性也是有条件的，即大部分其他交往者必须是有限制的最大限度追求者或可以教育好的无限制的最大限度追求者。如果绝大多数的交往者都是无可救药的无限制最大限度追求者，则哥梯尔的证明也不能成立。

再次，他的理性原则并非如他所希望的那样是唯一的理性原则。相当一部分当代西方哲学家认为，道德的合理性直接建立在与行动者无关的理性原则或理由的基础上，他们称这样的理由为“中立于行动者的理由”（agent-neutral reasons）。公众利益、公平原则都可以看作是合理性的原则，因为它们都有其价值。实际上，哥梯尔在论证道德的合理性时，有时求助于共同利益，有时求助于不偏不倚的立场（impartial position），如相等理性（equal rationality）和阿基米德支点（the Archimedean point）等，但这些不同于哥梯尔的理性原则的假设无法用哥梯尔的理性原则加以辩护，也无法还原成与行动者相关的理性原则。这些都说明哥梯尔的理性原则不可能是唯一的理性原则。

（本文发表在《河北学刊》2004年第3期。此次上网改正了原文个别误译之处，同时将“囚犯两难”改为“囚徒困境”，将“精明理性”改为“自利理性”。）
作者电子信箱：chenzhen@njnu.edu.cn

/