在自利与利他之外

——论罗尔斯“原初状态”道德视角的超越与困境

王　嘉

内容提要　罗尔斯在论证原初状态时，曾指出古典功利主义道德视角在自利与利他问题上的自相矛盾。与之对照，罗尔斯的原初状态道德视角的设置虽然在形式上超越了自利与利他，但在此视角之下却无法做出具有普遍效用的道德判断和选择，也无法真正解决自利与利他之间的难题。这一结论对于深入理解纯形式化的道德思想实验具有重要意义。

关键词　原初状态　自利　利他　超越　困境

王　嘉，南京师范大学公共管理学院讲师　　210023

自《正义论》问世以来，尽管不少研究者对原初状态（original position）持完全肯定的态度，但更多的研究者针对这种道德视角提出了各种各样的质疑和批评，其中既包括著名学者也包括一般研究者。笔者对原初状态这种道德视角的批评性分析，是从罗尔斯在《正义论》中批判的古典功利主义道德视角出发，刻画出原初状态思想实验的核心特征，并据此具体地“操作”此思想实验。

一

古典功利主义[1]“公正的旁观者”是罗尔斯在《正义论》中拿来和原初状态相比较的道德视角，也是本文首先要论证的原初状态所要超越的对象。罗尔斯归纳了公正的旁观者的几个基本特征：同情（sympathy）、公平（impartiality）、对相关知识的掌握以及具有想象力的认同能力。与之相对，原初状态下的各方则是相互冷淡而非同情的，而且缺乏对自身自然资质和社会地位的知识。罗尔斯认为两者的不同关键在于前者是同情的，而原初状态下的各方则是相互冷淡的。因此罗尔斯把“公正的旁观者”也称为“同情的观察者”，而且“一旦我们看到（同情的观察者的）定义中各部分被设计得给予同感（fellow feeling）以自由活动的余地，我们就能理解这一定义的要点。”[2]

本文为江苏省高校优势学科建设工程资助项目，江苏省高校哲社重点研究基地资助项目，南京师范大学人文社会科学青年科研人才培育基金项目（13QNPY02）。

[1]罗尔斯在这里所说的古典功利主义者是指斯密和休谟，他们也是18世纪英国情感主义（sentimentalism）道德学派的代表，这个学派还包括沙夫茨伯里和哈奇森等。

[2]John Rawls, *A Theory of Justice*, Cambridge: The Belknap Press of Harvard University Press, 1999, p.163.

同感和同情是罗尔斯所说的古典功利主义道德视角的出发点，罗尔斯认为从这样的道德视角出发是有问题的。首先，从同情的或仁爱的视角出发会导致利他主义，甚至是彻底的利他主义。而一个彻底的利他主义者所考虑的只是他人，满足了他人的偏好，就等于满足了自己。即便不导致彻底的利他主义，这样的道德视角和功利主义的另一基本信条也是冲突的，即：使自身的利益最大化。罗尔斯认为这是一个令人意想不到的结论：“如果说平均功利原则是一个试图最大限度地改善自己前景的有理性的人（他不厌恶冒险）的伦理标准，那么古典的功利原则就是完全的利他主义者的伦理。”[1]也就是说，由同情作用导致的利他主义特征和功利主义原则带来的利己主义在功利主义的框架下同时出现了，功利主义的道德视角既要从利他主义的角度出发，又要以自利为出发点。罗尔斯将这种状况称为“令人惊讶的对照”[2]。这是功利主义道德视角的一个无法克服的矛盾。

罗尔斯对“公正的旁观者”的另一批评是其非人格性（impersonality）。罗尔斯认为，古典功利主义的道德视角没有在人和人之间做严格的区分。这是因为同情作用的想象力使公正的旁观者能够站在每一个个体的地位上进行考察，将他人的欲求都当成是自己的欲求。“在古典功利主义的观念中，一个人进行选择时就像确确实实地经历了每个个体的体验，……并对体验的结果进行总结。”[3]罗尔斯的这一批评实际上批判的正是公正的观察者的同情特征，即，古典功利主义假设的道德视角虽然是站在每个个人的角度上考虑问题，但这种普遍化视角的尝试没有能够区分出个体自我的利益，它只具有涉他性利益关切的特征，所以罗尔斯说，“功利主义理论的失误之处在于它将公正误解为非人格性。”[4]相反，原初状态下的各方不是同情而是相互冷淡且仅关注自身利益的，因此原初状态下各方不是一个仅关注每个他人利益的没有自我人格的主体。

通过以上两个方面对“公正的旁观者”的批评，罗尔斯将原初状态的理论优越性凸显出来。正是在和古典功利主义道德视角的这种比较中，我们才能更清楚地看到“相互冷淡”、“相应知识的屏蔽”、“仅关注自身利益”这些设定的意义所在。如果单纯就原初状态这种 道德视角在超越自利和利他上的普遍化尝试而言，原初状态在设置上的确非常巧妙地克服了罗尔斯揭示出的古典功利主义道德视角的矛盾。

二

罗尔斯在引述“公正的旁观者”时提到的古典功利主义者指的是休谟和斯密。在休谟和斯密的伦理学中，“公正的旁观者”是道德主体视角的出发点。以同情作用为基础的“公正的旁观者”所秉承的理论传统渊源于近代英国的情感主义道德学派（sentimentalism）。情感主义道德学家通常把同感（fellow feeling）、同情能力（sympathy）视为道德主体超越自身利益进行涉他性关切的道德动机。他们认为同情不仅是人们在日常生活中实际具备的心理能力，而且人们的利他主义实践和真正的道德生活都必须诉诸于这种与生俱来的涉他性欲望或情感。早期情感主义的涉他性关切发展到休谟和斯密这里，就是“公正的旁观者”。“公正的旁观者”诉诸一个非自身利益关切的视角，它的核心特征是将视角从自身“移出”，并“移入”他人的地位去思维，从而作出一个不被自身利益或利己主义所左右的判断或行为。

从《正义论》的表述可以看出，罗尔斯对这种道德视角并不完全反对，因为和原初状态一样，公正的旁观者也试图提供一个普遍化的观察点或视角。但罗尔斯非常明确地指出了原初状态和公正的旁观者的关键不同：公正的旁观者是具备同感（fellow-feeling）或同情（sympathy）的，而在原初状态中，

[1][2][3][4]John Rawls, *A Theory of Justice*, Cambridge: The Belknap Press of Harvard University Press, 1999, pp.164-165, p.165, p.165, p.166.

各方是相互冷淡而非同情的[1]。这一区别带来的结果就是,“在将所有的欲望融成一个欲望系统之中的意义上,导致了非人格性。”[2]这一结果使得公正的旁观者无法合理地讨论具体自我的利益,可以想象这一状况可能导致一种完全的利他主义。

在罗尔斯的分析框架内,导致公正的旁观者的上述困境的主要原因就是其同情特征。而同情正是情感主义伦理学家用来对抗利己主义或自爱的道德视角的核心理念。霍布斯从经验的和反经院神学的立场提出了人的自利本性,并将此作为伦理学的出发点。情感主义者则从反霍布斯的利己主义出发,提出具有利他特征的同情视角。在情感主义伦理学家这里,一个人的思维和行为是否道德,取决于他是否是同情的。反之,如果一个人的道德视角的出发点是完全的自利或自爱,那就是不道德的。情感主义者们显然不能接受霍布斯那种类似于人性本恶的宣言,也不认为社会的稳定和人们之间的和谐相处主要依赖于强制权力以及互利原则。人的本性中还有一种更为重要的东西能够体现人的道德动机,即同情。同情中包含着移情(empathy)作用,即将思维和行为主体的视角移入他人的地位,站在他人的处境上来思考、判断和行为。罗尔斯认为,这种视角虽然具有普遍化特征,但是很大程度上隐去了本应作为道德思维出发点的道德主体自身的利益诉求,并且有可能导致彻底的利他主义困境。换句话说,站在公正的旁观者的视角,主体的思维立场既不是自我的,也不是他人的,而是非人格的。因为这种道德视角体系既不是从自我出发的,而是假设或想象出他人的地位而“移入”的;也不是某个他人的,而是每一个他人的。

罗尔斯在《正义论》第30节中给出的上述简明论证揭示了以同情或利他作为道德(政治)视角的出发点所导致的逻辑困境。而以自利作为道德理论基础的伦理学,也有其不可克服的矛盾。如霍布斯的伦理学(政治哲学),将自利的人性假设作为主体思维和行动的出发点和终极目的,每个主体是否为他人考虑、和他人合作,是否遵守契约,是否遵守规则,归根到底是看这样做是否于己有利。但诚如布莱恩·巴利所言:“如果我们假设,目的就是一己私利,我们就可以说,当一己私利是通过不遵守规则增进的,就没有理由坚守由协议产生的规则了。”[3]也就是说,按照霍布斯的理论,如果主体是为了自己的利益最大化才和他人合作、遵守规则,那么假如在某种情形下不遵守规则也可以使自身利益最大化,主体就完全有理由不遵守规则而使自身利益最大化。这样一来,以自利为出发点的契约论体系就无法为社会成员如何恪守规则提供逻辑自洽的证明。

实际上同情视角和自利视角共同的困境在于,不管是以同情还是以自利作为道德(政治)哲学的首要原则,这一首要原则都有可能与与其相反的倾向发生逻辑上的冲突。即,如以同情为首要原则,就无法合理地解释自身利益(如罗尔斯所论证);如以自利为首要原则,就无法在整个理论系统中做到不与互利合作相冲突(如巴利所论证)。在这两种视角下,利己和利他没有办法在整个理论体系中无可辩驳地相容。

在道德视角的选择上,罗尔斯采取的方案是从自我出发,而不是从他人出发。更确切地说,是从相互冷淡的理性自利的自我出发。在这一点上,罗尔斯回到了霍布斯的自利假设。但是无知之幕的设定让罗尔斯没有止步于霍布斯,理性自利加上无知之幕的设置,使罗尔斯既超越了霍布斯,也超越了公正的旁观者。因为各方是理性自利的,所以其出发点是自身利益,从而保证了原初状态下的各方不会变为非人格的利他主义者。同时各方所具备的正义感以及对自身的地位和背景的一无所知,保证了原初状态下的各方不可能从自身的具体地位出发而成为自私自利者,而是“被迫”站在每一种可能的(主体)处境上进行考虑选择。原初状态的这一精巧设计使罗尔斯在利己和利他之间取得了

[1][2]John Rawls, *A Theory of Justice*, Cambridge: The Belknap Press of Harvard University Press, 1999, p.163, p.164.

[3]布莱恩·巴利:《作为公道的正义》,曹海军、允春喜译,〔南京〕江苏人民出版社2008年版,第42页。

形式上的制衡。或者说，相互冷淡、理性自利和无知之幕的多重保证使罗尔斯的理论体系避免陷入同情或自利假设可能遭遇的上述困境。

正是在这个意义上，在原初状态之下，似乎出现了一种奇诡的状况（两个方面）：首先，理性自利的设置使原初状态下的各方既不会成为利他主义者，也不会成为只考虑自身特殊利益的自私自利者；同时，原初状态下的各方既是从自利出发的，又是“被迫”从他人利益（利他）出发的。这种状况的前一个方面可以说是否定性的方面，上文已做了解释，即原初状态的设置对利己和利他皆有限制。后一个方面是肯定性的方面，即原初状态的设置又同时具备利己和利他的特征。肯定性的利己特征是通过相互冷淡、理性自利的假设体现出来的，而肯定性的利他特征则需要从无知之幕本身来分析。

我们知道，在无知之幕之下人们都不知道自己的社会地位、阶级出生、天资禀赋和自然能力，也不知道自身的善观念和生活计划，甚至不知道自己的心理特征，如冒险还是保守，悲观还是乐观。这样的设置实际上意味着，在每个可能主体身上，每一种可能背景出现的概率都是一样的。由于不知道自己具体的特殊背景，所以必须将每一种可能的背景都纳入权衡，才能作出最合理最佳的选择。在这种情形下，思维主体要做的，就是选择一个处在任何地位上的人都能合理接受的原则。从这个意义上说，原初状态下的主体思维立场是特定形式的利他主义，因为各方在不知道自身背景的情况下被迫考虑的是每一种可能背景下主体的利益。而且我们完全有理由认为这也是另一种形式的同情或移情，因为考虑每一种可能背景下主体的利益，就相当于将思维立场“移入”每一个可能的主体背景地位。只不过这种“移入”不是由道德情感意义上的同情心理所驱使的，而是由形式化的道德视角设置来实现的。这就像罗尔斯所说的：“相互冷淡和无知之幕的结合达到了跟仁爱一样的结果。因为将这些条件的结合迫使原初状态中的每一个人都将他人的利益纳入考虑。”[1]

这样，原初状态的设置不仅“保证了与初看起来在道德上更吸引人的假设的同样效果”[2]，而且避免了上文提到的可能由仁爱或同情视角带来的那些困境。换句话说，原初状态不仅兼顾到了利己和利他，还规避了可能由利己和利他假设带来的逻辑上的不相容。不管是“己”的利益，还是“他”的利益，在原初状态的框架内都不相冲突地被考虑到。因此，这一道德视角可以说是真正“普遍化”的，即，不是着眼于某一特殊群体的利益，而是着眼于每一个体的普遍利益。这一普遍化视角不仅使罗尔斯在形式上超越了公正的观察者和霍布斯自利假设，同时也在形式上超越了利己和利他。这就是本文要论证的第一个问题：罗尔斯原初状态道德视角的超越。

三

接下来自然转入本文要讨论的第二个基本问题，即，在此普遍化的视角之下，道德主体能够做出怎样的选择？道德主体在形式上超越利己和利他的普遍化视角之下，能不能也做出超越利己和利他的具有实质意义的选择？笔者认为，即使罗尔斯能够设置出一个超越利己和利他的道德视角形式，也无法在此视角之下得出一个超越利己和利他的实质性原则。在此，有必要实际操作一下罗尔斯的思想实验，来看看我们在原初状态下会做出怎样的选择。

上文已经指出，由于不知道自己的具体处境，原初状态迫使思维主体考虑每一种可能处境下的利益。也就是说，我们在原初状态下，必须站在每一种可能的境况下去考虑问题。当然，不可能将每一种境况都考虑一遍，也没这个必要。为了简化起见，我们只选择最有利者和最不利者进行考察。首先我们从最有利者的角度来看。因为原初状态下的每一主体都是理性自利、相互冷淡的，而且不

[1][2]John Rawls, *A Theory of Justice*, Cambridge: The Belknap Press of Harvard University Press, 1999, pp.128-129, p.129.

存在一般意义上的同情等心理特征。因此设想我是一个最有利者,那么在社会利益分配问题上,我所期望的应该是一个有利于最有利者的制度安排。比如,用“最大最大化“原则规范的制度。即,社会制度的基本原则应该尽可能地适合甚至扩大最有利者的利益需求。其次,站在理性自利的最不利者的处境之上,应该选择有利于最不利者的社会制度安排。例如”最大最小化“原则,即罗尔斯的差别原则:社会和经济的不平等应该这样加以安排,以使它们适合于最不利者的最大利益。

现在,在这一简化了的思想实验过程中,产生了两种选择,即“最大最大化”原则和“最大最小化”原则。这两个选择是主体在不知道自身的具体背景而被迫在每一种(实际简化为两种)可能的背景下做出的。接下来的问题是,站在每一(实际是两种)处境上做出的这两种选择 ,将如何被选择? 即,在“最大最大化”原则和“最大最小化”原则之间,将如何做选择?

导向罗尔斯差别原则的一个重要的推理依据就是“最大最小化”原则。就上述思想实验来说,“最大最小化”原则应该是站在最不利者地位上的选择,“最大最大化”原则应该是站在最有利者地位上的选择。关键问题是 ,罗尔斯凭何将“最大最小化”原则而不是“最大最大化”原则作为导向最终结论的推理的依据? 对此,罗尔斯的解释是,如果不采取“最大最小化”原则,处在最不利者的地位上就可能遭致无法忍受的恶劣境况。因此,为了保证不导致可能产生的无法忍受的境况,我们就必须选择“最大最小化”原则来作为推理的依据。

但是,罗尔斯的这一关键性选择并不是一个具有普遍意义的“公正”选择,他的这一选择只是站在最不利者的地位上做出的。我们完全有理由认为,站在最有利者的地位上,“最大最小化”原则也会是无法忍受的 。这就像理查德·米勒在《罗尔斯和马克思主义》一文中指出的:“可以这样说,最有利者通常会认为,为了使最不利者的状况最大限度地提高而放弃巨大利益是无法忍受的。…除非采取强制,否则在任何剥削社会中的最有利者都不会放弃他们的特权。比如,没有哪个支配性的剥削阶级会主动放弃他们的统治地位。”[1]举一个实际的例子,美国篮球明星乔丹在20世纪90年代做球员期间,坚决站在球员工会一边,在劳资谈判中要求提高球员工资。对于这时的乔丹来说,压低球员工资是无法忍受的。但在2011年美职篮劳资谈判时,已是球队老板的乔丹却坚决地站在资方,始终要求压缩球员收入比例。因为对于身为老板的他来说,提高球员工资是不能忍受的。所以米勒说:“我将试图表明,罗尔斯的这些主张假定了一种对社会(利益)冲突的范围和后果的相当低的评估。”[2]

实际上在最不利者和最有利者各自的无法忍受的状况之间如何做出选择,这是一个原初状态本身无法回答的问题。原初状态只是给我们提供了一个站在每一种地位上考虑问题的视角,而如何在这些视角产生的结果之间做选择,则需要另外的标准。如果把问题转换一下,就好比面对两种不同的赌博规则,你会选择哪一种:在第一种规则下,你不会输得太惨,也不会赢得太多;在另一种规则下,你可能输得很惨,也可能赢得盆满钵满。这种问题实际上已经不再具有普遍意义,而涉及到个性化的心理倾向。不同的赌博规则之间做选择,取决于个人的心理偏好。同样,如果在最有利者和最不利者各自无法忍受的状况中做选择,也只能取决于个体的个性化倾向,并没有一个普遍性的标准可供衡量。如果某人选择上述第二种赌博规则,这并不是不合理的。同样,在原初状态下,某个具有“最大最大化”心理倾向的个体选择了“最大最大化”原则,也并不是不合理的,尽管他知道这一选择将决定他的生活前景。除非罗尔斯能够证明没有人具有这种心理倾向,否则他采用“最大最小化”原则进行推理就不具有普遍性,由“最大最小化”原则推理出的(被认为)会被各方普遍接受的差别原则也站不住脚。

---

[1][2]Richard Miller, “Rawls and Marxism”, Norman Daniels. *Reading Rawls: Critical Studies on Rawls´ “A Theory of Justice”*, Palo Alto CA: Stanford University Press, 1989, p.214, p.206.

思想实验操做到这里，可以看出，以普遍化视角为特征的原初状态设置，如果不引入个性化偏好，根本无从在各个视角下产生的选择中做选择。以本文的理论旨趣来说，这揭示出原初状态在利己和利他之间的困境：原初状态看上去是使每个人都被迫关注每个人的利益，从而在自利的前提下呈现出普遍利他的形式，但如果不考虑个性化偏好，原初状态只能向我们展示在每一种可能处境下所选择的利益诉求，而在这些利益诉求之间，却无法做选择。

换句话说，以原初状态为例的普遍化道德视角虽然为我们提供了一个形式上超越利己和利他的思维立场，但在具体的道德（政治）原则的选择和判断上，是无法超越利己和利他的。甚至可以说，我们仅在高度形式化的普遍化道德视角下无法作出任何实质性的选择和判断。以上述思想实验为例，如果不另行引入个性化心理偏好因素，仅在原初状态下我们根本无法论证分别站在最有利者和最不利者处境下做出的选择哪一个更值得重视。诚如米勒所言，站在上层阶级的角度我会认为平均化是无法忍受的，而站在下层阶级的立场上则会觉得“最大最大化”原则是无法接受的。在不考虑同情因素、将各方设置为相互冷淡的原初状态下确是如此。

原初状态，除了作为一个纯粹在形式上平衡利己和利他的设置之外，最重要的功能应该是提请人们站在他人的视角上去想问题、做判断，我们的道德立场应该是具有“人类“视角的，而不仅仅是”自我“视角的。在这个意义上，原初状态是形式化了的彻底“同情”。但正如罗尔斯所批判的那样，彻底的同情或利他是不可行的；同样，单纯的原初状态也是不可行的。因为彻底的同情是“在决定要做什么的时候，每个人都是在其他的每个人想要做的事中作选择。很显然这样什么也解决不了；事实上也并没有什么事情要决定。”[1]正如罗尔斯自己所言，正义问题起码要在两个或以上的人之间才会产生。在不讲概率的原初状态下，即便罗尔斯特别强调了各方是理性自利的，但由于所有人都不知道自己的具体状况，因此想象中的各方一方面表现为一个个“绝对隔离”的自我，另一方面又表现为没有真正的“彼此”之分。既然无从比较，或没有“彼此”，那么何来“正义”？从这个角度，我们不难看出已为许多研究者所批判的罗尔斯原初状态下的主体际性问题，即，作为一个个不分彼此或绝对隔离的“我”，如何去讨论“正义”或“公正”问题？

综上所述，罗尔斯的原初状态设置虽然在形式上超越了休谟和斯密的公正的观察者以及霍布斯的利己主义视角，但道德主体单纯在此视角下，仍然无法作出超越自利和利他的、具有普遍意义的选择。这一困境如同在《国富论》和《道德情操论》之间所表现出的“斯密问题”一样，也许远远超出道德（政治）哲学的范畴，需要用更加具体、更加实证的方法来探讨自我和他人之间的利益平衡问题，即正义问题。单纯形式化的普遍化视角的设置，对于解决自我和他人之间（即人与人之间）的正义问题并无帮助，更无法从中推导出有利于某个特定群体的差别原则，最终只是一种“超然”于自利和利他之外的思维“幻象”。

〔责任编辑：曾逸文〕

[1]John Rawls, *A Theory of Justice*, Cambridge: The Belknap Press of Harvard University Press, 1999, p.165.