**社会正义与差别原则**

——财富与风险分配公平的互惠性思考实验

【内容概要】

本文主要以罗尔斯的《正义新论》为线索，探讨当今世界的贫富悬殊和社会撕裂的问题状况以及相关的政治哲学-法理学思考实验。作者指出，差别原则是理解罗尔斯正义理论的关键，而互惠性观念则是差别原则的正当化根据。从这个角度来重新认识罗尔斯正义理论，可以发现分配曲线的形式化论证的背后其实潜伏着重构宪法共识、加强东西方文明之间就秩序原理和制度设计进行深入对话的契机。

Based on John Rawls's *Justice as Fairness: A Restatement*, this paper explores the current problematic situation of Gini Coefficient and social disintegration in the world as well as the related political-legal thought experiment. The author points out that the difference principle is the key to understand Rawlsian theory of justice, and the concept of reciprocity is justification basis of the difference principle. From this perspective, it can be found that behind the formalistic argument of the distribution curve lies the opportunity of reconstructing the constitutional consensus and strengthening the in-depth dialogue between Eastern and Western civilizations on the principles of social order and institutional design.

【关键词】

纯粹程序正义 公平分配 合作博弈 差别原则 互惠性 公共理性

pure procedural justice, equitable distribution, cooperative game, difference principle, reciprocity, public reason

季卫东\*

关于罗尔斯的政治哲学和法哲学主张，国内外的研究、讨论以及批评的文献已经汗牛充栋了。本文无意重复已有的论述，也不想面面俱到，而是仅仅聚焦罗尔斯的晚期著作——更强调分配公平性的《正义新论》，特别是其中的差别原则、互惠性概念以及OP曲线的形式化论证，从不同视角考察国家与法律理论创新的契机及其对中国社会治理话语体系重构的参考意义。

一、前提条件：理性的建构与合理的选择实践

罗尔斯的思想渊源，主要是穆勒式自由和宽容的思想以及在此基础上接纳的康德式构建主义。一般而言，他的学说因为致力于社会合作系统以及良好秩序的构建，所以难免些许设计理性的倾向。另外，因为注重自由、宽容以及相应的互惠性，所以设计理性又会留有合理选择和调整的空间，从而需要对“理性的（rational）”与“合理的（reasonable）”概念进行更加细致微妙的区别。

“理性的”与“合理的”这一分类在罗尔斯正义理论成形的早期阶段就已经出现了，[1]但似乎没有得到具体而明确的表述。在《正义论》这本书中，两者之间并没有明确的界限；罗尔斯仅在为数不多的场合同时使用“合理的”和“理性的”这两个概念，且基本上视它们为同义词。直到《政治的自由主义》发表时，这一对概念才有了精确的、完整的展开。[2]《正义新论》成书之际，罗尔斯引入西布里（Sibley）对这两个概念所作的区分，[3]并通过康德的纯粹实践理性与经验实践理性的分类以及完全自律与合理自律的区别等媒介，与政治意义上的人的两种道德能力的概念相呼应。在这里，人的第一种道德能力是指正义感觉，第二种道德能力是指善良构想；[4]前者属于理性的范畴，后者属于合理的范畴；前者主要涉及平等的自由原则，后者主要涉及结果不平等的条件、多元性的事实以及适当调整和修正的机制。罗尔斯曾经具体分析判断一个人是否合情合理的标准包括两个层面：首先，人们乐于提出那些能够被所有人都视为公平合作条款的原则，如果别人提出这些原则，他们也乐于肯认；其次，当其他人都遵循这些原则时，他们也会承诺履行这些原则，即使有可能损害其自身利益。[5]不难发现，这两方面的特点正好对应着自由而平等人的两种道德能力中“正的界说”与“善的界说”。[6]一般而言，正的界说以及相应的道德能力与“理性的”类型相关，而善的界说以及相应的道德能力与“合理的”类型相关。[7]

在做出以上区分之后，从“原初状态”出发建构道德以及正义理论的纯粹程序主义思想实验就延伸出两条逻辑进路。第一、理性选择理论，即证明那些掩藏在“无知之幕”背后、不知道自己属于哪个阶层、但却致力于自己利益最大化的人们，他们所作的决策、所立的合约是理性的，在揭开“无知之幕”后也可以通过“反思均衡”的伦理学方法来建构和重构道德，并通过道德稳定性形成正义观念。第二、合理选择的实践理论，即证明在撤出无知之幕后的现实条件下，可以通过“重叠共识”的政治学方法来建构合作体系，从而确保制度架构的稳定性。在这里，追求重叠共识的主语是不同的价值观。或者说人们通过沟通、合意以及承认来实现不同信仰、思想、伦理的兼容并蓄。因此，重叠共识与暂定的妥协不同，是基于道德根据的一种稳定的政治构想。值得注意的是，罗尔斯曾经多次强调，合理性优先于理性，合理性是对理性的限制条件；[8]这种有限理性的立场就使得设计合理主义带有某种进化合理主义的色彩，并且导致“反思均衡”有狭义与广义之别——狭义的反思均衡是指特定主张能一以贯之，或者说是一种正义观实现内在整合性，而在广义的或者说完全的反思均衡化过程中，人们还必须对公共正义的构想与其他可选择性正义构想进行整合。[9]

《正义新论》一再强调合理的优先于理性的，似乎理性逐步失去道德上的中立性而被吸收消化到合理的概念之中，并且还主张那种试图把整个正义原则的论证都建立在理性选择理论之上的做法是一个错误。[10]但实际上，在罗尔斯论证正义原则的过程中始终贯穿着一种目的理性的演绎逻辑方法，并大量借鉴经济学的形式化技术，[11]俨然把“效用最大化”作为理性的判断标准，[12]而合理的概念反倒更像是对理性在遇到疑难情况时的一种补充。在这里正义观似乎出现了理性的悖论。无论如何，合理与理性总是交错在一起并且显得密不可分，或者用罗尔斯自己的话来说，最终的正义理念是“合理而理性的”（reasonable and rational）人们的选择。[13]在这个理论体系中，处于原初状态的公民只能是理性的、自律的，不得不停留在目的理性的位置上；另一方面，只有当公民自觉适用他们在原初状态中自由选定的正义原则并以此为动机采取行动时，他们才能完全实现自律。因此，把理性与合理有机结合在一起的是罗尔斯的道德主义倾向，但却对外呈现出一种纯粹程序正义的形态。

二、价值排序：关于正义的思考实验

1. **原初状态与无知之幕**

罗尔斯试图从“原初状态”通过程序化的演绎式论证推导出正义原则。他的理想是找到一种深思熟虑并且坚定的信念，以建构一种在自由而平等的公民之间实现公平的社会合作体系。但实际上却只能进行推论，说明处于原初状态的公民会在政治层面的正义原则清单上进行什么样的选择。无论如何，罗尔斯旨在用上述信念或者推论衍生出一种恰当的分配原则，据此调整公民之间的不平等。[14]所谓原初状态，是把公民作为仅仅平等的道德人格来表现的一种代表装置，是一种确保缔结任何协议都是公正的、不受主观任意性以及力量对比关系左右的条件设定，由此也是对推论进行限制的选择装置。他借助原初状态这一思考实验装置来模拟参与协议签订的各方的反应，[15]其目的并非要对人们的行为进行解释，而是为了对道德判断进行说明，并解释公民所持有的正义感以便相互理解，进而寻求全体一致的同意。[16]

实际上，罗尔斯无法证明我们必须接受他关于原初状态的假定，而只是希望通过沟通来说服我们接受——即使不是立刻接受，在考虑之后也会最终接受——这种假定。他希望自己关于原初状态的构想能够确立一种在缔约当事人、公民以及我们自身当中都获得广泛承认的正义观，为此通过“无知之幕”的假定来化约对原则进行选择的过程的复杂性。当然，并不能认为原初状态的构想本身没有道德效力，只是原初状态所依据的概念群集在伦理上是中立的。[17]这里更重要的是观察视角的区别。实际上，持有原初状态的构想、讨论其依据的概念群集的只有关注元规则的“我们”。对“我们”而言，原初状态是具有道德意义的。然而对于那些身处原初状态之中的缔约当事人而言，原初状态在道德上则是中立的。在这里，为了避免直观的、实质性的道德判断互相冲突引发混乱和战争，罗尔斯把正义论的第一判断，即我们（公民）作为自由而平等的道德人格这个命题彻底形式化了。正是这样做的结果才出现了所谓原初状态。在这个意义上也可以说，原初状态构成了与我们的直观的道德判断相对应的一种完全形式化的公理体系。正因为如此，我们势必倾向于程序主义的正义论。罗尔斯曾经明确指出，“原初状态把纯粹的程序正义最大限度具体化了”。[18]

不言而喻，所谓程序正义是指不存在关于正确与否的独立标准，正确与否是由程序的结果来定义的正义构想。[19]在这里，程序上的正确性被转换成了结果上的正确性。基于同样的逻辑，原初状态的公平也被转换成了从正义原则演绎出来的公平，因而才有“作为公平的正义”这种说法。把原初状态视为纯粹程序正义的一个场景，就势必导致合理的自律这样一种模型。也就是说，当事人们从给定的选项清单无论选择什么原则都被认为是正确的。[20]因此，在作为形式化公理体系的原初状态当中进行合理计算的结果就可以被理解为程序正义的内容。换言之，程序的内容是由合理的计算构成的。由此可见，基于原初状态所能得出的正义理念也不是唯一的，需要进行选择。在理想状态下是通过演绎性论证从原初状态推导出正义原则，但现实情形是推论当事人将从政治原则的一长串选项清单中怎样进行选择。在《正义新论》中，关于选择的推论实际上采取的是两步比较法。首先把正义两原则与平均效用最大化原则进行比较，然后再把正义两原则与带有最大最小规则保障的平均效用原则进行比较。

概而论之，最大最小规则是关于原初状态下的当事人在选择正义原则之际应该考虑的各种论点的发现装置。罗尔斯对该规则的适用情形进行了辩护，指出在原初状态这种假设状况下引入该原则是一种合理的选择。但是，他也坦言最大最小规则只是一种启发式算法，用于辅助原初状态的立约人进行慎思。[21]因此，罗尔斯并没有穷举原初状态下能够成立所有的选项（也无法做到），尽管有很多选项被“无知之幕”的论证给排除在外，但仍有一些选项通过了原初状态的筛选。此外，不宜适用最大最小规则的情形大致可以举出三种：（1）各种状况的分布概率至少在一定程度上是已知的，（2）最坏情形在采取最佳选项时产生的结果（保证水平）是无法令人满意的，（3）未选择的选项导致的最坏结果并不会低于保证水平太多。[22]毫无疑问，这些情形无疑限缩了最大最小规则的适用范围。但无论如何，最大最小规则对正义两原则的正当化而言并不是决定性因素。[23]

一般认为，“无知之幕”是对公民在不清楚自己所属阶层的状况下选择使自己利益最大化的正义原则的思维方式的一种概括，以此作为实验的条件设定。无知之幕的假设不等于一种理性选择理论，但却有利于选择的论证以及在重叠共识的基础上进行正当化。但是，也有学者认为无知之幕的设定对于论证平等原则没有帮助，这一批判的代表人物便是德沃金，在政治哲学领域内他常被认为是与罗尔斯、诺齐克齐名的人物。德沃金认为，罗尔斯隐藏在无知之幕背后的深层理论乃是一种抽象的权利理论，基于抽象性的权利观。这一抽象权利既不指向某种特殊的个人权利，也不可能指向特定个人目标的权利。换句话说，虽然有对各种自由的具体权利，但是并不存在某一种对自由的抽象权利。[24]德沃金认为，在政治理论中，这一比自由更加抽象的一般性概念，便是“平等”。从平等中可以演绎出自由，却不能从自由演绎出平等。[25]在德沃金看来，罗尔斯正义理论的基础实际上是具有两种道德能力的人（做出人生的合理计划和拥有正义感）所拥有的平等权利，因为平等关怀和尊重权利已经暗含于原初状态（纯粹程序正义）的先决条件中，是无知之幕背后的立约人展开谈判的前提，而不是契约订立的结果。[26]德沃金还认为，实现分配正义的前提是资源平等，这种平等应该通过符合完全程序正义的拍卖市场模式来实现；在资源平等基础上的分配正义则应该通过作为纯粹交换机制的保险市场模型以及以平等关怀为原则的行政举措来实现[27]。值得注意的是，德沃金借助运气与选择这个关键性的概念区分，把责任观念和问责机制嵌入运气和风险分配之中，进而把选择自由、责任自负的自由主义原则也嵌入平等的正义理论之中，与罗尔斯的正义新论有其曲同工之妙。

1. **反思均衡与判断整合**

“反思均衡”是罗尔斯正义理论的一个独具特色的伦理学方法，也被理解为一种正义原则和道德观念的建构以及正当化机制，其本质在于判断的整合化。在《正义论》一书中，罗尔斯认为反思均衡，就是要让原初状态下选择的正义原则与我们直观的、经过深思熟虑的道德判断达到均衡。博弈论学者宾默尔（Binmore）对于反思均衡这一概念的理解，非常有启发性，他说：“这是发展科学模型时的一种基本方式，即理论模型与实证数据之间的互动”。[28] 换言之，如果理论模型和数据不相符，要么更改模型，要么对数据进行更仔细的检查。[29]反思均衡所使用的“数据”其实就是罗尔斯所提到的暂定的不动点（provisional fixed points）。[30]对于思辨性的科学探索（speculative scientific investigation）需要运用形式化方法（formal methods）来约束思考过程，能够一定程度上避免解释的任意性而造成空中楼阁。这对于道德哲学也适用，而且更加不能马虎。[31]

需要指出的是，反思均衡并不旨在寻找一个认识论上的真理，而是说在一定的历史时期内，我们的某些观念可能是非常带有确信属性的。[32]例如，罗尔斯将林肯关于奴隶制的判断作为一个不动点。这并非是说林肯的判断就是绝对正确的，而是说，该判断目前被人们内心所坚信，因此不妨在开始反思均衡时视之为正确判断的起点。[33]当然，无论对于起点本身我们有多大的信心，随后的反思过程都有其必要性。在这个意义上，反思均衡并不依赖一个不证自明（self-evident）的起点，也就是说它是一种非基础主义的（nonfoundationalist）论证。[34]值得留意的是，罗尔斯的理论其实是把正义两原则当作实现反思均衡的原则，为此必须调和超越性演绎与经验性归纳、理性与合理之间的关系。

 

图 1：反思均衡示意图[35]

不妨将图1中的点状符号理解为特定历史环境和社会中人们所持有的常识性判断，即“不动点”。其中，矩形点表示人们普遍内心确信是错误的判断，而圆形点表示普遍被相信为正确的观点，那么反思均衡可以理解为一种筛选机制——对那些使用正确/错误的符码对常识性判断进行分类操作的政治理论加以甄别。例如，图1中的虚线描绘了基于常识性判断所能达到反思均衡的区域，某一种政治理论如果落入这一划定的区域之内，便可以被视为通过了反思均衡的考验。图1（右）展示了一种未通过反思均衡的情况。图中的实线表示某一种政治理论，该理论将某一圆形点——不妨认为其代表了“劳动者有休息的权利”这样一种被普遍确信的判断——被错误地置入了直线的左侧，即直线所代表的政治理论与人们的常识性判断产生了冲突。换而言之，如果无法推翻常识性判断，那么该政治理论便没能经过反思平衡的检验。值得注意的是，常识性判断并非永远正确。图1（左）的反思平衡区域右下方有一个方形的灰色点便展示了这样一种罕见的情况，即某一个不动点与反思均衡状态中得出的理论发生了冲突，但是并未造成理论的修改，反而是引起人们对于该不动点判断的改变。从这个意义上说，我们可以理解为什么反思均衡是非基础主义的论证方式。也就是说，处于原初状态的人们在揭开“无知之幕”之后需要通过反思均衡的互动来整合价值判断、确立道德并且实现道德的稳定性，进而确立正义的观念。而在现实生活中，则是在均衡化互动的基础上以达成“重叠共识”的方式来形成社会合作体系、实现社会稳定性，进而建立和健全政治制度。上述两个稳定性的实现在相当程度上有赖于论证以及正义感的陶冶，为此必须把合理的道德心理学（reasonable moral psychology）也纳入正义论的视野之中。[36]

三、正义焦点：差别原则和财富分配

众所周知，罗尔斯理论的主旨是论述正义的两个原则，即平等的自由原则与关于应对结果不平等的例外性原则。后者又细分为两种设定条件限制的原则：（1）机会的公平平等原则，（2）“差别原则（difference principle）”。在这里，以经济发展的较高水平为前提，上述原则的优先顺序是作为宪法必要条件的第一原则以及机会平等原则、差别原则——差别原则的排序在最后。但是，仔细推敲其中的思想内容和演变路径，我们可以发现其实差别原则才是罗尔斯关于正义学说的核心观念或者生命线。那么，差别原则是矫正或者缩小贫富悬殊的原则吗？是回应最不利者的利益最大化诉求以缩小差距的原则吗？回答为是，但又不全是。例如足球或者棒球运动员以及艺术明星的巨额报酬并不违反差别原则，因为正义原则并不适用于私人团体或者结社组织。又例如在发展中国家出于生存权的考虑而限制某些自由权利也被认为没有违反正义的排序标准，因为对于饥寒交迫的人而言，自由显得太奢侈。当然，罗尔斯本人曾经指出过，把基本自由的优先性完全限制在繁荣程度较高且足以使它们被“实际”享有的社会，还是有所不足的。从原初状态的角度来看，缔约当事人也许会坚持认为，他必须相当富足，能够实际享受这些自由。由于存在着无知之幕，便设定了一个条件：为了使基本自由的优先性得到落实，每个公民必须在物质上相当富足，使他们能够实际享受这些自由——虽然没有明说，但论证中一直就存在着这层含义。

既然罗尔斯承认这个前提条件，即要想使第一条正义原则成为一条固定的优先原则，“需要公民的基本（物质）需要得到满足”，[37]那么我们就可以用四条原则来表述罗尔斯的“特殊的正义概念”。[38]这些适用于财富足以给全体成员提供基本物质需要的任何社会的原则，按照优先秩序排列如下：

1. 必須满足每个人的基本物质需要。
2. 在一个与人人自由的诉求相类似的、实现恰当平等的关于基本自由的方案中，每个人都应该享有平等的权利。
3. 在如愿获得社会地位的机会的分配之际，这个社会体系必须尽量增加处境最不利的阶层的机会。
4. 对经济的不平等必須做出这样的安排，以尽量增加最贫穷阶层的财富。

在这里，显然是用差别原则来解决社会和经济的不平等问题。关于差别原则的讨论，需要考虑它与其他原则之间的关系以及差别原则适用的对象（限制什么样的分配）、标准（按照什么尺度进行调整）、手段（采取何种方法再分配）以及正当化的理由。差别原则是狭义的分配正义原则，因此只针对基本善（primary goods）的下位具体分配问题，也就是说只涉及收入与财富的分配。为了向每个人提供基本需要，为了符合差别原则，就需要政府的干预和财富的再分配。[39]政府必须用“家庭津贴以及疾病和失业的专门支出”，或是通过消极的收入税，来保证某种“社会最低标准”。[40]机会的公平平等原则（或尽量增加机会）则要求政府通过公立教育制度或补贴私立学校的举措来保证教育机会的公平。

罗尔斯的正义两原则适用于基本善的全面分配，既包括基本的权利和自由、迁徙和择业的自由，也包括与职务和地位相伴的权威以及与责任对应的权力或特权、收入和财富、尊严等等，都是客观的、看得见摸得着的，但是这种公平分配的对象不包括主观的幸福感。差异原则仅仅适用于收入和财富的分配，含有以下两层涵义：第一、在机会公平平等的前提下承认竞争以及作为其结果的差别；第二、对结果的差别还要进行适当的调整，而调整的方式是让最不利者的利益得以最大化。在这里，如何判断最不利的阶层或群体、如何对不同阶层或群体进行比较、他们的人生期待收入如何确定等等就成为不可回避的问题，需要客观而精准的计算。罗尔斯试图通过OP分配曲线等经济学的分析方法和技术手段来改进差别原则的适用，类似的尝试当然也可以举出中国政府推动的精准扶贫的实例。

根据差别原则，我们应该对社会和经济的不平等做出一种适当的安排，能最大限度地使最不利阶层获益；[41]如果出现比这种安排更严重的不平等那就是不正义的。但是，比这更小的不平等又如何呢？例如通过劫富济贫的方式缩小贫富差距就是过分的行为，实际上也会让穷人更穷，显然不是最好的状态。[42]用福利经济学的话说，它不是帕累托最优；而采用差别原则所规定的分配方式，富人和穷人的命运都会得到改善，这既是帕累托最优，也是公平的，因此是“最佳的正义安排”。罗尔斯主张的差别原则是要兼顾经济效率与分配正义，为了达到两者之间的平衡，也为了更清晰地说明正义观念，他导入了关于分配正义的OP曲线的工具性分析框架。

四、OP分配曲线的分析框架

尽管OP曲线只是用于辅助分析而非一种形式化的论证工具，但它对于理解罗尔斯的正义原则、特别是根据差别原则进行分配的思路具有重要的意义。罗尔斯的OP分配曲线在区分更有利群体（More Advantaged Group, MAG）与更不利群体（Less Advantaged Group, LAG）的基础上建立差别原则适用场景的分析坐标，为差距的指数测量提供了包括最高正义直线、公平分配最大值、纳什点（Nash Point）、边沁点（Bentham Point）、封建点（Feudal Point）等基本概念或指标，[43]形成了一个理论分析的初步框架，既可以进一步探索数理化的分析，也可以进行更深入的理论说明。

下图是《正义新论》中用来诠释差别原则的一张经典图片。其中，横轴代表的是社会合作中的更有利者群体（MAG）的福利；纵轴代表的是更不利者群体（LAG）的福利。将社会合作形成以及变化过程模拟为仅仅两个群体之间的博弈，显然忽视了社会群体的多样性和复杂性，并且没有考虑群体相互之间的动态转化以及各群体内部的差异性。但是，这一简化后的模型仍然捕捉到了分配正义中的主要问题，同时避免了陷入过于复杂的数学证明。



图2：OP曲线[44]

**1. 社会合作中的效率和公平**

值得注意的是，罗尔斯运用OP曲线的分析框架时隐性地采用了博弈论中的合作博弈（cooperative games）来模拟社会更有利群体与更不利群体之间的竞争与合作关系。这反映了关于社会作为公民之间合作框架的理念，即“没有合作，就没有任何东西被生产出来”。[45]基于这样一种假设，相较于非合作博弈（non-cooperative games），采取联合博弈的模型确实更符合罗尔斯对社会的理解。[46]在罗尔斯看来，社会合作体制具有以下特征。首先，合作体制由公共规则决定，后者包括：组织生产互动、规定劳动分工、派定各种角色。其次，所分配的对象包括从产品中分配的应付的薪酬和收入。每一种特定形状的OP曲线对应于一种特殊的合作体制。OP曲线起点O是平均分配点，两个群体在此处得到了相同的报酬，但是社会生产效率极低，随着分配差异的产生，社会效率得到提高，P点向右上方延伸，并开始偏离了代表平等分配的45度斜线。

显而易见，唯有提供差别原则这样的公平基础作为条件，那些处境更好的或财富更多的群体才能期待他人来与之进行合作。[47]这似乎是在暗示，任何的合作条款如果被更不利群体视为不公平的，那么他们就有理由可以选择不合作，这使得情况回到了没有合作的僵局（例如，OP曲线中代表完全平等但效率很低的O点）。这种僵局显然是更有利群体和更不利群体都想竭力避免的情况，否则双方各自的收入和财富都会被减损。一般而言，在给予更有利者的报酬既定的情况下，给予更不利者的报酬更多，则合作体制更有公平。例如，图2中的一系列平行于横轴的直线（例如，J-J）是平等—正义线（equal-justice line）反映了“互惠性”的理念。较高的平等—正义线，意味着更不利群体的基本善指标更大。相反地，从较高的平等—正义线向较低方向的移动，则表示更不利群体的基本善被减损。虽然此时更有利群体的基本善指标得到增加，但无法将这一现象正当化，因为此时“相互交易”（trade-off）[48]不被允许，互惠性遭到破坏。不难发现，差别原则在OP曲线上所对应的D点引导社会达到OP曲线上的最高点，达到最高的平等—正义线。

除了OP曲线上代表差异原则的D点以外，还有多个点值得注意。特别是OP曲线中反映纳什均衡的纳什点（N点）。纳什点表示的是一个合作博弈中的纳什均衡解。所谓合作博弈下的纳什解是指这样的一种合作协议，即如果一方提出与该协议不同的合作方案，并声称说，“如果你不同意，我就会采取一些举措，从而按照一定概率导致我们无法达成合作。”另一方仍可以回应说，“即便你采取措施，造成无法达成合作的危险，对我而言坚持之前的协议仍是更好的选择。”[49]数学上可以证明，更不利群体的基本善指标与更有利群体的基本善指标之乘积在N点达到最大值（即几何平均值最大）。如图3（右）所示，LAG与MAG的基本善指标之乘积即N点与O点所确定的矩形面积，大于曲线上其他点与O点所构成的矩阵面积。[50]

此外，边沁点（B点）关于更有利群体和更不利群体的基本善指标之总和达到最大，[51]意味着算数平均最大。所以，B点代表着一种关于社会最大多数人最大幸福的累加式功利计算。[52]在数学上可以证明，OP曲线在B点的切线斜率为1，任何对B点的偏离都会导致基本善总和的减少。比如说，从B点向右下方往F点方向移动，则每1个单位的MAG的增量伴随着大于1单位的LAG减损，基本善总和小于边沁点。另一方面，如果从B点向左上方往D点移动，则每1单位的LAG的增量伴随着大于1单位的MAG减损，基本善总和也小于B点。因此，更不利群体和更不利群体的基本善之和在B点达到最大值。此外，OP曲线上还有一个封建制度点（F点）。在该处，OP曲线变成几乎垂直，这意味着社会分配以牺牲更不利群体的基本善为代价，极度偏向更有利群体。



图3：几何平均和Nash均衡

**2. 几个关于差别原则的批评**

 借助于OP曲线的分析框架，我们可以对关于差别原则的几个经典批评进行讨论。第一种批评认为，最有效的OP曲线上升到最高点之后向右侧变化的斜率会非常缓慢，[53]形如下图中O—D1—P1曲线。在曲线从O向D1上升的过程中，MAG的收益远大于LAG，因此D1点对更不利群体来说是不公平的。不难发现，这一批评没有对更不利群体实际获得的基本善指标进行比较，而是聚焦于对更有利群体和更不利群体所获收益的比例进行讨论，因此如果要在合理性上进行分析，有可能涉及到关于嫉妒的概念，这让我们联想到德沃金以嫉妒测试作为资源平等的判断标准的契[54]。但是罗尔斯采用了一种更加简单直接的回应方式，即O—D1—P1曲线在一个秩序良好的社会中是不可能存在的。[55]差别原则和在先的正义原则共同作用之下，背景制度拥有机会的公平平等和有效竞争，将缩小分享份额的比例差距，更有利者无法作为一个群体来联合行动以借助市场力量强行增加收入。因此，从O到D1那种分配额度极度不平等的情况将不会发生。

第二个批评的意见与前一个批评的OP曲线刚好相反，认为OP曲线到达最高点后下落的速度将会非常缓慢，形如图中O—D2—P2的曲线。批评者认为，如果曲线能从D2点下落，则以LAG微小的损失为代价能够带来很大的LAG份额的增长（从而也大大增加了社会的基本善总和），因此停留在差别原则所规定的D点似乎对更有利群体不公平。[56]罗尔斯对此的回应也与这一假想的OP曲线是否可能存在有关。[57]在满足正义两原则的情况下，一定存在某种制度性设置来对财富进行再分配（例如税收制度），因此更有利群体的巨大报酬之中一定有相当部分转移给了更不利群体。所以，从D2到P点，LAG的下降非常缓慢的同时MAG急速增加这种情况是不可能发生的。



图 4：关于差别原则的两个批判

概括来说，罗尔斯认为，坚持了他揭示的正义原则（平等的基本自由+机会的公平平等），以上两种批评中所假设的OP曲线形状不可能实际出现，因此也就没有进一步讨论的必要性。不过，需要指出的是，对于LAG与MAG之间分享额度的比例，差别原则不规定具体的限度，而是把这个比例作为纯粹背景程序正义的结果，除非发现实际比例非常不公正才需要进行调整。[58]

最后，我们看一下诺齐克关于更有利群体与更不利群体之间的合作问题的一个批评意见。罗尔斯强调，普遍社会合作能够给各方更大的利益，尽管各方所得将出现差异（对应OP曲线中O点即平均分配点）。但是，除了普遍的社会合作之外，诺齐克指出也可以有范围较狭的局部社会合作，比如MAG群体内部相互之间的合作。因此，把O点作为最差情况显然有些脱离现实。更进一步，诺齐克甚至认为，更不利群体无疑能从普遍合作体系中得到比更有利群体的得利更大的利益，因为更有利群体内部的局部合作本身就是非常有利的。所以，差别原则试图使更不利群体更加得益的做法（例如对少数族群的优待措施）是不公平的。差别原则提出要最大限度地促进那些比方说因才智较低而处境最差者的利益，但这种安排会损害更有利群体的利益从而可能使他们宁愿退回到该群体之内的局部合作中去。诺齐克因此断言，差别原则实际上会降低社会成员间的普遍合作。[59]

因此，仅仅使用社会合作这一概念，尚无法论证更有利群体需要向更不利群体让利。如果更不利者有权利提出以给他们尽可能多的份额作为合作条件，更有利者不也有权利提出以给他们尽可能多份额作为合作条件吗？OP曲线这样一个基于合作博弈的框架，对于LAG和MAG来说，双方都能够对等地以拒绝合作为“底线”，从而迫使对方作出一定妥协，但这种妥协的限度在哪里呢？如果认为更有利者提出要获得尽可能多的份额这类条件是任意专横的（确实如此），那么更不利者提出类似的条件为什么就不是任意专横的呢？罗尔斯认为更不利者没有理由抱怨，因为每个人的幸福都依赖于社会合作，没有合作，任何人都难以过上一种满意的生活，因为生产寥寥，分配就更加捉襟见肘了。[60]但是，这并不构成更有利者就应当让利的理由，因为正如诺齐克试图论证的，同样的逻辑也同样可以用于让更不利群体向更有利群体让利。所以，基于OP曲线的论证框架最多只能为社会合作提供理由，而不能为差别原则提供理由，实际上并没有证明差别可以成为社会合作的公平条件。[61]在晚年，罗尔斯对于差别原则的正当化论证主要借助下一节我们将详细讨论的互惠性。

五、差别原则的正当化：互惠性

互惠性是《正义新论》区别于早前的《正义论》的最大特色，也是罗尔斯阐述差别原则并对之进行正当化处理的重点。[62]互惠性的思想渊源或许可以追溯到穆勒式的自由和宽容，其外延也不妨包括暂定协议，并通过重叠共识以及相互共识与社会契约精神对接。晚年的罗尔斯认为，对正义两原则进行正当化处理的根本基础是互惠性。在这一点上我们很容易联想到富勒，他认为社会的黄金律是互惠。[63]实际上，《正义新论》的根本规范就是互惠性。互惠性最鲜明地表现在差别原则中，但是在这里并没有做出很清晰的区分。本来互惠性对正义两原则都是有效的，但是它对第一原则和第二原则的第一部分的影响并没有得到充分的说明和论证。因为罗尔斯不得不借助差别原则来说明互惠性，很容易让人产生误解，似乎互惠性仅仅与差别原则有关，其实《正义新论》的目的是要论证正义两原则的正当化根据在于互惠性。

罗尔斯对互惠性的格外关注，一定意义上与他对《正义论》做出的一个重要更正相关。[64]在正义论（第一版），正义理论被视为理性选择理论的一部分（a part of the theory of rational choice），但罗尔斯在后期认为这一说法是一个错误。理性选择理论只是原初状态的缔约当事人所使用的一种推理理论，用来对合理的正义原则给出解释，并不是正义理论的全部。[65]在这里，我们可以看出关于理性和合理的区分的意义，这一区分正是罗尔斯后期思想的重要发展。在罗尔斯看来，合理优先于理性，后者对前者可以起到一定程度的解释作用。与这种更正相关联，最大最小规则对差别原则的证明不再发挥关键作用，罗尔斯转而诉诸公共性和互惠性，特别是通过互惠性来论证差别原则。[66]实际上，与带有最大最小保障的平均效用原则相比较，差别原则得以胜出的最大根据就在于互惠性的观念。

概括地说，互惠性的概念内容包括对等性、相互利益以及利他指向这样三个层面。当人们认为平均效用原则不包含互惠性时，其实是把互惠性与相互利益大致等同起来了。但在其他很多场合，互惠性可以理解为相互利益与利他指向的中间形态。[67]也许我们可以对互惠性的观念进行更具体的分类：（1）把相互利益界定为一种弱互惠性，主要通过OP分配曲线中从O到D之间的表面变化或者帕累托改进来体现；（2）把利他指向界定为一种强互惠性，主要通过在相互利益基础上的相互贡献来体现；（3）把天赋才能的公共物品化（智商税）界定为高互惠性，主要通过“能者多劳”或者处于有利位置的人们向处于不利位置的人们的单向贡献来体现。当然，更有利者群体与更不利者群体之间的力量对比关系或者互动关系会影响到互惠性的实现程度或状况，甚至造成互惠性的扭曲和偏颇。另外，互惠性还涉及道德心理学的问题，例如嫉妒、仇富的心理如何反映到差别原则中也是饶有趣味的研究课题。

毫无疑问，差别原则是具有规范含义的。差别原则也是社会稳定化的一种重要力量，可以增进公民之间的互信和协调的德性。差别原则在政治和法律领域应用的最显著的实例可以举出税收制度或者税法的设计方案，是采取平均税制还是累进税制、遗产税的轻重程度如何把握等等，都会对贫富差距产生影响。但是对罗尔斯而言，差别原则与其说是立法指针，毋宁说是说服的逻辑和修辞。换句话说，从实现社会正义的角度来看，差别原则也许可以成为每个公民的行为指针，但却未必能成为立法指针。德沃金试图弥补这个缺憾，所以通过假想的保险市场模式来解决平等分配、供求平衡、福利改进等一系列问题，并把保险制度设计与税收制度设计对应起来，为分配正义的实现提供立法的基本指针。[68]然而对于罗尔斯而言，要使差别原则成为个人的行为指针还有一个不可或缺的前提条件，即自由的公共理性。[69]也可以说，互惠因公共理性而超越暗盘交易，公共理性则因互惠而具有现实可行性——因为只有对各方都有利的时候，才能说服各方。特别是在无知之幕揭开之后，发现现实与原初状态下确立的正义原则是不一样的，在这种情况下，互惠性的逻辑就可以显示出说服力。[70]因此，罗尔斯晚年在《正义新论》中阐发的理论观点就已经与那些仅仅基于理性的概念（concept of rationality）来推导正义原则的理论观点区别开了，后者的推理主要建立在理性人的假设之上。例如宾默尔认为，相较于理性而言，合理性没有应该任何位阶上的优先性。[71]相反的，罗尔斯论证个人为什么要遵守原初状态达成的协议时，涉及到了公民的正义感，这正与两种道德能力中与合理的（reasonable）部分相对应。

六、理性选择理论的重新定位

前面我们已经指出，罗尔斯从合理的与理性的两个角度来论证正义原则，因此理性选择理论不再是唯一基石。尽管如此，罗尔斯在《正义论》中引入OP曲线、博弈规则等工具性分析框架，使得正义理论这类抽象的哲学思考有可能借用经济学等的数理方法进行更具体、更实证的分析，其结果势必加强理性决策理论的作用。为此还有必要再深入探讨一下公共理性观念，并考察罗尔斯遇到的理论障碍以及解决问题的思路。

如果沿着基于理性选择理论对正义原则进行论证这条脉络来看，可以说豪尔绍尼（Harsanyi）和罗尔斯（在正义论中阐述的早期观点）恰好构成两极，而其他学者的主张则分布在其间的不同位置上。[72]双方的争议主要在于无知之幕的厚度以及人的偏好是什么。[73]宾默尔的立场介于两者之间，并在两人研究的基础上进一步从理性选择的角度来推动正义理论发展。一方面，宾默尔认为豪尔绍尼所采用的分析框架（基于贝叶斯决策理论）比罗尔斯的最大最小规则更加严谨。罗尔斯式立约人那种对不确定性的极端厌恶（uncertainty aversion）是非理性的（irrational），[74]所以宾默尔试图抛弃所谓的要遵守原初状态所达成协议的自然责任（natural duty）。[75]或者说，抛弃合理性这一逻辑进路，完全依赖理性选择理论来论证平等意义上的公平。另一方面，他非常赞赏罗尔斯有关分配应尽可能平等的直觉，[76]并被差别原则所吸引而加以弘扬，因为它最大化了最不利群体的福利。[77]从理性选择理论的视角来看，豪尔绍尼和罗尔斯在预设前提假设时都采用了无知之幕这一装置来模拟原初状态下立约人的情况，但是两人对于立约人会如何进行决策所采用的理论是非常不同的。罗尔斯使用最大最小规则来模拟立约人的决策过程，并基于此对差别原则的优越性进行论证；豪尔绍尼的推理则基于贝叶斯决策理论。一般而言，最大最小原则被认为是求解博弈论问题时的一种特殊解法，它在二人零和博弈（two-person zero-sum）这一情形中碰巧是纳什均衡解，但其适用场景具有局限性，因此无法取代一般性的、基于期待效用最大化的求解方式。[78]宾默尔认为，罗尔斯舍弃贝叶斯决策理论的原因在于他不能接受豪尔绍尼基于该理论做出的关于功利原则的论证，因此选择了基于最大最小规则的论证路线。[79]但是，罗尔斯关于原创状态所设定的条件，在宾默尔看来是非常契合贝叶斯决策理论所适用的应用场景的，因此挑战正统的理性选择理论并不是一个好的选择。[80]

在罗尔斯设想的正义原则选择之际，最大最小规则的作用究竟有哪些？首先，罗尔斯设定了假想的原初状态，罗尔斯试图通过演绎式的论证，推导出正义原则。在这里其实存在一个默示的前提条件，即进行康德式的解释，把某种道德规范作为所有公民都可以接受并服从的命令或义务。我们知道社会契约式的论证基本上都是这么做的，从确信推论出正义。但是，在合理的多元化社会，在一个后形而上的时代，正义理论的论证需要非康德式的公共理性以及政治的正义观念，需要实证的分析。因此，罗尔斯不得不考虑存在一份正义原则的清单，并假设原初状态的当事人接到这份清单会进行怎样的选择，然后再对正义观念进行公共论证。首先要考虑人们如何获得正义感，其次还要考虑如何实现社会合作体系的政治构想。为此，罗尔斯着重分析了达成重叠共识的理由（包括原则与妥协、暂定协议的稳定化和恒久化）以及过程。在罗尔斯看来，重叠共识的目标是实现善治，因而需要根据正义构想形成一种秩序良好的社会。对秩序良好社会的判断标准主要有三条：（1）所有公民都接受统一的正义原则，即存在基本共识；（2）作为社会合作体系的制度能实现正义原则；（3）公民还都具有正义感。在这个意义上，善治意味着对每个人都是善，同时对社会也是善。[81]

由此可见，所谓重叠共识是以公共理性为前提的。对应于合理的多元化社会现实，公共理性应该包括“正确或正义的界说（conception of right or justice）”与“善的界说（conception of the good）”这两个不同侧面，这也恰好对应于公民的两种道德能力，即具有正义感的能力以及具有向善心的能力。从政治和法律的角度来看，这两个侧面分别对应于“公域”与“私域”、“公法”与“私法”。在分析重叠共识时，罗尔斯的关注点是暂定协议、宪法共识、全体一致同意以及重叠共识等不同类型共识之间的逻辑关系，而不是某种实际的历史演进过程。关于正义原则的理性选择以及重叠共识的说明，在《正义论》中罗尔斯是把正义两原则与古典的功利主义以及平均效用最大化原则进行比较，论证正义两原则的正当性。但在《正义新论》中，罗尔斯做了一个方法论上的改变，主要是导入了另外两种比较的方法。第一种比较是让正义两原则与平均效用最大化原则进行对峙，看谁更有说服力。第二种比较是用带有最大最小保障的平均效用原则（有限的平均效用原则）与正义两原则来比较。

在原初状态下，通过最大最小规则的应用可以发现支持正义两原则的论据或者选择正义原则时应该考虑的论点，但却未必能使正义两原则正当化。也就是说，最大最小规则构成在原初状态下缔约当事人发现有利论点的装置，但并非正义原则正当化的决定性因素。[82]这个最大最小规则与博弈论有一些相似之处，实际上罗尔斯的论证框架也有一点像博弈论的分析框架，当然也存在一些实质性差异。在论证过程中，罗尔斯的叙述好像有些重复和循环。但是如果把罗尔斯的主张从相反的方向来理解，考察一下罗尔斯究竟在排除什么、否定什么，就会获得启迪并更加透彻地理解他的主张。把适用最大最小规则的条件反过来看，[83]就是不能适用最大最小规则情况，大致分为如下三种。第一、各种状况的概率分布基本上是知晓的，也就是说在不知道概率的情况下才能适用最大最小规则；第二，保证水平——最坏结果在采取最佳选项时结果——是无法满足的，也就是说如果保证水平已经得到满足，也没有必要采取最大最小规则；第三、未选择的选项导致的最坏结果并不会低于保证水平太多，这时也没有必要用最大最小规则。这三种情况已经暗示，在一些情景设定下，坚持最大最小规则有可能是非理性的。

此外，即便承认最大最小规则在无知之幕下并不是一个非理性的决策，也还是存在其他广泛接受的、类似贝叶斯决策理论那样的理性选择模式，例如对效用期待的最大化。按照这种模式，处在无知之幕后面的理性选择者会最大限度追求平均（average）的好处而不是追求最低限度好处的最大化。豪尔绍尼曾以此为根据批评罗尔斯并且赞成平均效用最大化的原则。然而这样做的结果很可能造成一个非常不平等的社会，至少与罗尔斯构想相比较会不平等很多。罗尔斯理解这种论点并力求驳倒它。[84]他提出了若干反对的观点，这里只考虑其中的两个。一个是罗尔斯所说的“承诺的强度”（strains of commitment），其内容要点在于原初状态下的缔约当事人不仅进行选择，而且也在明确一种承诺，对此他们必须保持一定的信心。[85]这也正是社会契约的性质所决定的。如果他们不能使自己服从这样的原则，假如运气不好的话就会产生他们无法接受的灾难性后果。另一个是在拒绝功利主义的前提下，对整体良好的社会进行理性选择时应该聚焦处于最不利位置的群体。从罗尔斯的立场来看，在原初状态下选择功利主义是不理性的，甚至可以说是一种鲁莽行为，即使有可能因而产生巨大的财富。在中国社会背景下，这个观点特别值得重视和推敲。另外，在一定意义上，使最不利者的利益最大化是把偶然的机遇或者制约个人选择的遗传、出身等状况性条件这个与运气相关的因素也纳入效用概念的范围之内，存在着相当的合理性。而与差别原则的正当化密切相关的互惠性的说服力也有利于增强具有不同价值观的人们讨论社会正义问题时的公共理性。

七、结束语

如果我们从差别原则这个特别的视角来考察罗尔斯关于社会正义的理论构成，特别是分析《正义新论》对《正义论》所做的一些重要修改补充，就会发现在正义原则的价值序列上排在第三位的差别原则其实是理解罗尔斯正义理论的关键。在二十一世纪的世界结构大转型的过程中，面对欧美各国的族群裂变和发展中国家的贫富悬殊，差别原则似乎变得越来越重要。不言而喻，正义原则的适用对象是社会的基本结构，即分配社会合作的利益以及明确权利义务的主要制度；如果把作为罗尔斯正义理论的核心价值和生命线的差别原则作为立法指针，宪法共识应该进行什么样的重构，这是一个饶有兴味的研究课题。

本文留意到差异原则限制的只是社会基本善的基层分配，即收入和财富的分配，使政府对结果不平等的矫正尽量向处于最不利位置的阶层或群体倾斜，在一定条件下使他们的利益最大化。为此需要对社会基本善进行指数测量、对阶层的收入状况进行分析和排序、形成作为机会收入的人生期待收入的阶梯，并在这些作业的基础上进行OP分配曲线分析。虽然我们希望在OP曲线中找到立法指针，但实际上却做不到；罗尔斯描述的OP曲线提供的只是对处于有利位置的阶层或群体进行说服的逻辑或修辞，充其量只是提供了个人行为的规范性指针。如果要把OP曲线与立法指针联系在一起，就必须深入分析税收和财政制度对分配的影响，这是另一个富于现实意义的研究课题。在这里，把罗尔斯正义理论与德沃金正义理论进行对勘和深入的比较分析也是饶有兴味的。

从中国的语境出发，特别值得重视的是罗尔斯在《正义新论》中对互惠性的强调，这是对《正义论》的一个最重要的修正。晚期罗尔斯的正义理论把互惠性作为说明差别原则并使之正当化的主要抓手，并推进了互惠性的类型化——从单纯的相互利益到互利加贡献，再到利他指向的贡献。尽管罗尔斯认为互惠性是正义两原则得以正当化的根本基石并试图论证这一点，但在实际上互惠性在差别原则上的体现最为鲜明。互惠性对正义第一原则以及第二原则中的机会平等原则的影响机制究竟如何，罗尔斯本人语焉不详，因此也就在此留下了一个非常复杂的研究课题。我以为，对于中国法学界而言，互惠性的正义理念正是与各国进行学术的深度对话、用世界可以理解的语言讲述固有秩序原理和制度设计思路的一个极其重要的切入点，或许构成我们一直在上下求索的社会最大公约数以及重构人类普遍共识的基础。

（载《现代法学》2021年第1期）

[1] John Rawls, *A Theory of Justice*, Cambridge: The Belknap Press of Harvard University Press, 1971, pp. 12-13, 18-19.

[2] John Rawls, *Political Liberalism*, Cambridge: Cambridge University Press, 1993, pp. 47-54.

[3] W. M. Sibley, “The Rational versus the Reasonable”, *Philosophical Review*, vol. 62, no. 4 (October 1953), pp. 554-560. 参见John Rawls, *Justice as Fairness: A Restatement*, Erin Kelly ed., Cambridge: The Belknap Press of Harvard University Press, 2001, p. 7.

[4] *Justice as Fairness: A Restatement*, op.cit., p. 196.

[5] Ibid., p. 7.

[6] 见*Justice as Fairness: A Restatement*, op.cit., p. 19. 值得注意的是，那种纯粹为了自身利益而遵守在原初状态所达成的公平条款的行为，不宜被解释为合理的。因为，对于在合理的人（reasonable person），即便在环境迫使他们不得不以牺牲自身利益为代价时，也会承诺履行那些关于公平合作条款的原则。*Justice as Fairness: A Restatement*, op.cit., p. 7. 关于“正的界说”与“善的界说”的区别以及在法学研究方面的意义，参阅季卫东：《互惠的正义——法理学的视角转换及其实践意义》，载《中国法律评论》2018年第3期1-5页；季卫东：《中国式法律议论与相互承认的原理》，载《法学家》2018年第6期1-15页。

[7] 在《正义论》中，罗尔斯将具有这两种道德能力的人定义为道德人，一种具有自己的目标并且具备正义感的理性存在（rational beings），但是理性的与合理的之间没有一个明确的区分。参阅*A Theory of Justice*, op.cit., p. 12. 到了《正义新论》阶段，罗尔斯的表述更为清晰——作为政治意义上的人，公民的道德能力可以由善（good）和正（justice）两方面构成。或者说，公民具有理性的和合理的能力。*Justice as Fairness: A Restatement*, op.cit., p. 196.

[8] *Justice as Fairness: A Restatement*, op.cit., pp. 81-82.

[9] Ibid., pp. 29-31.

[10] Ibid., p. 82.

[11] 在《正义论》中，罗尔斯非常推崇一种从前提到结论的具有严格形式化特征的道德几何学（moral geometry）。尽管在论证过程中不得不借助高度直觉性的推理，罗尔斯仍认为严格的演绎推理是应当追求的理想目标。参阅*A Theory of Justice*, op.cit., p. 121, 126. 对于这种严格形式化的推理，罗尔斯的态度在晚年所持有的态度明显更加谨慎。在《正义新论》中，他专门列举了两条理由来解释为什么理想状态的演绎推理式是难以达到的。详见*Justice as Fairness: A Restatement*, op.cit., pp. 133-134.

[12] *A Theory of Justice*, op.cit., p. 14.

[13] *Justice as Fairness: A Restatement*, op.cit., p. 8.

[14] *A Theory of Justice*, op.cit., p. 7.

[15] Ibid., p. 120.

[16] Ibid.

[17] Ibid., p. 579.

[18] John Rawls, “Kantian Constructivism in Moral Theory”, *The Journal of Philosophy*, vol. 77, no. 9 (1980) p. 523.

[19] *A Theory of Justice*, op.cit., pp. 85-86; *Political Liberalism*, op.cit., p. 73.

[20] *Political Liberalism*, op.cit., p. 72.

[21] *Justice as Fairness: A Restatement*, op.cit., p. 8.

[22] 罗尔斯描述了三种可以使用最大最小规则的情况，如果从反面的角度来看待这些情况，便是最大最小规则不一定适用（或者说，使用该规则有可能是非理性的）的情况。参见*Justice as Fairness: A Restatement*, op.cit., pp. 98-99.

[23] Ibid., p. 98.

[24] 参见何怀宏：《公平的正义——解读罗尔斯<正义论>》，济南：山东人民出版社，2002年，第215—216页。

[25] 同上，第215页。

[26] Ronald Dworkin, *Taking Rights Seriously*, Harvard University Press, 1978, pp. 181-183.

[27] 参阅罗纳德•德沃金：《至上的美德——平等的理论与实践》，冯克利译，南京：江苏人民出版社，2003年，特别是67-129页。

[28] Ken Binmore, *Playing Fair: Game Theory and Social Contract Volume 2*, MIT Press, 1998, p. 319.

[29] Ibid., p. 319.

[30] *A Theory of Justice*, op.cit., p. 20.

[31] *Playing Fair: Game Theory and Social Contract Volume 2*, op.cit., pp. 319-320.

[32] *Justice as Fairness: A Restatement*, op.cit., p. 19.

[33] Ibid., p. 29.

[34] Ibid., p. 31.

[35] 本图的绘制受到一种经典机器学习算法“支持向量机”（Support Vector Machine, SVM）的启发，该算法中的支持向量，即位于决策边界上的正类与负类样本，与反思平衡中的不动点有异曲同工之处。关于SVM算法的更多介绍，可参阅张学工：《关于统计学习理论与支持向量机》，载《自动化学报》2000年第1期，第36—第46页；丁世飞等：《支持向量机理论与算法研究综述》，载《电子科技大学学报》2011年第1期，第2—第10页。

[36] *Justice as Fairness: A Restatement*, op.cit., pp. 429-434.

[37] 罗尔斯承认，涉及平等的基本权利的第一原则也有可能让位于公民的基本需求，当这些需求对于公民理解以及充分行使其基本权利是必要的。此外，罗尔斯甚至明确指出，有时候牺牲一定的自由来换取长期的收益是合理的。不过，从长远来看，权利优先于善，仍是是一个公正系统内在的长期均衡（long-run equilibrium）。换言之，按照罗尔斯的正义理念，如果一个社会的情况最终无法带来权利的平等，那么这是不可被接受的。*A Theory of Justice*, op.cit., p.247-248; *Political Liberalism*, op.cit., p. 7.

[38] 罗尔斯关于两条原则作为四部分的表述受到了Peffer的启发，参见Rodney Peffer, *Marxism, Morality, and Social Justice*, Princeton University Press, 1990, p.14. 尽管罗尔斯不认同其中的3(b)，这一条要求一种经济组织的社会主义形式。*Political Liberalism*, op.cit., p. 7.

[39] 参见[英] 迈克尔·H. 莱斯诺夫：《二十世纪政治哲学家》，冯克利译，北京：商务印书馆，2002年，第314页。

[40] *A Theory of Justice*, op.cit., p. 275.

[41] 参见《二十世纪政治哲学家》，（前引）第319页。

[42] *A Theory of Justice*, op.cit., pp. 78-79.

[43] *Justice as Fairness: A Restatement*, op.cit., p. 62.

[44] 参见[美] 约翰·罗尔斯：《作为公民平的正义：正义新论》，姚大志译，北京：中国社会科学出版社，2011年，第79页。

[45] 同上，第77页。

[46] 关于联合博弈与非合作博弈的辨析，参见Martin J. Osborne and Ariel Rubinstein, *A Course in Game Theory*, Cambridge: The MIT Press, 1994, p. 255. 此外，联合博弈模型中有一个假设，博弈的结果必须得到双方的共同认可。如果任何一方不合作，就会导致无法达成合意。无法合作（disagreement）的结果被假定为最差的结果，是双方尽力避免的，同样参阅前Osborne书, 第300页。

[47] *A Theory of Justice*, op.cit., p.103.

[48] *Justice as Fairness: A Restatement*, op.cit., p. 63.

[49] *A Course in Game Theory*, op.cit., p. 302.

[50] 对于基本善指标乘积取得最大值时得到联合博弈的纳什解，相关证明参见*A Course in Game Theory*, op.cit., pp. 300-304. 简单来说，当LAG和MAG的基本善的乘积达到最大值的时候，如果任何一方如果提出N点以外的分配结果，并且暗示出于己方利益最大化的考量，将有一定概率p（0到1之间的一个数值）停止合作，对另一方而言，在考虑了合作失败的概率p之后，坚持N点的方案仍然是最优选择。举一个具体的例子，假设MAG和LAG的基本善在纳什点分别是6和4，如果MAG对LAG提议，“我想分配7，你分配3，如果你不愿意的话，我会采取一些措施，使得最终合作失败的几率是10%”。对MAG而言，这一提议显然是理性的（rational），因为在考虑失败的风险之后，MAG的预期收益为(1-0.1) \* 7 = 6.3，比纳什点的收益6更高。但是，LAG不会受这种风险的影响，考虑10%的失败概率后，如果LAG坚持此前的方案，其预期收益为4 - 0.1\*4 = 3.6，仍比新方案中的3更高。因此，LAG仍会坚持纳什点，MAG想要用新方案取代纳什点将无法奏效。以下图片摘自*A Course in Game Theory*, op.cit., p.304；图中两条曲线的交点即纳什解，此时博弈双方的基本善乘积（v1\*v2）最大。对于联合博弈的纳什解有更复杂而详尽的论述，可参阅Roger B. Myerson, *Game Theory: Analysis of Conflict*, Cambridge: Harvard University Press, 1991, p. 379。



[51] 参见《作为公民平的正义：正义新论》，（前引）第78页。

[52] 博弈论经典文献将边沁点视为一种基于最大善原则（the greatest good principle）之下，通过磋商所得到的一个功利解（utilitarian solution），见*Game Theory: Analysis of Conflict*, op.cit., p. 381。

[53] *Justice as Fairness: A Restatement*, op.cit., p. 66.

[54]《至上的美德——平等的理论与实践》，（前引）69-73页。

[55] *Justice as Fairness: A Restatement*, op.cit., pp. 66-67.

[56] Ibid., p. 66-67.

[57] Ibid., pp. 67-68.

[58] 参见《作为公民平的正义：正义新论》，（前引）第85页。

[59] Robert Nozick, *Anarchy, State, and Utopia*, New York: Basic Books, 1974, pp. 192-197.

[60] *Justice as Fairness: A Restatement*, op.cit., p. 61.

[61] 参见《公平的正义——解读罗尔斯<正义论>》，（前引）第194—195页。

[62] 在《正义论》中，互惠性概念只是作为差别原则的一个优点被提及，并未被置于特别重要的地位。罗尔斯假定了社会合作中有A和B两个人，并且假定A是处于更有利位置的人，B是处于更不利位置的人。说服B接受差别原则似乎并不困难，因为如果不允许A处于更有利的位置，B的情况就会比现在更糟。但是，如何说服A接受差别原则呢？罗尔斯的论证由两条组成：第一，两个人的生活都依赖社会合作，否则没有人能过上更好的生活；第二，只有在合作条款合情合理的时候，我们才能要求每个人进行合作。详见*A Theory of Justice*, op.cit., p.103.

[63] 参见[美] 富勒：《法律的道德性》，北京：商务印书馆，2005年版。

[64] *Justice as Fairness: A Restatement*, op.cit., p. 82.

[65] *Political Liberalism*, op.cit., p. 53.

[66] *Justice as Fairness: A Restatement*, op.cit., p. xvii.

[67] Ibid., p. 77. 详细的解释见*Political Liberalism*, op.cit., pp. 16-17.

[68] 参阅《至上的美德——平等的理论与实践》，（前引）6-115页、400-405页。

[69] John Rawls, “The Idea of Public Reason Revisited”, *Collected Papers*, Cambridge: Harvard University Press, 1999, pp. 537-615.

[70] *Justice as Fairness: A Restatement*, op.cit., pp. 122-124.

[71] 基于理性的立场，宾默尔对于康德式的定言命令这一具有绝对性（unconditional）的概念持有强烈的批评态度。Ken Binmore, *Rational Decisions*, Princeton University Press, 2009, p. 4; Ken Binmore, *Just Playing: Game Theory and Social Contract Volume 2*, Cambridge: MIT Press, 1998, p. 157.

[72] 参见丁利：《社会正义理论：豪尔绍尼与罗尔斯的比较》，载《思想战线》2006年第2期，第14—16页； *Rational Decisions*, op.cit., pp. 73-74.

[73] 无知之幕的厚薄之别，实际上体现了原初状态与功利主义传统上不偏不倚的旁观者之间的实质区别。后者可以说从休谟和斯密等古典功利主义者一直传承至像森和豪尔绍尼。豪尔绍尼为不偏不倚的旁观者提供了两种不同的模型：其一，个人具有内在化的道德偏好（internalized moral preferences），这种偏好可能通过第三方的视角表达出来，然而，这些偏好也可能是因人而异的；其二，不偏不倚的旁观者相当于参与到一个思想实验中，他将试着考虑每个人的主客观条件，并且想象自己有同等的概率来成为任何一个他人，同时也能够忽略自身的实际境况。由于第二个模型包含从一种假想情境中得出判断的过程，因此和原初状态有近似之处。观察者模型一见John C. Harsanyi, “Bayesian Decision Theory and Utilitarian Ethics”, *American Economic Review*, vol. 68, no. 2, 1978, pp. 223-228; 模型二见John C. Harsanyi, “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility”, *Journal of Political Economy*, vol. 63, 1955, pp. 309-321.

[74] *Just Playing: Game Theory and Social Contract Volume 2*, op.cit., p. 319

[75] Ibid., p. 431

[76] *Rational Decisions*, op.cit., pp. 73-74.

[77] *Just Playing: Game Theory and Social Contract Volume 2*, op.cit., p. 316.

[78] Ibid., p. 319.

[79] Ibid., p. 316.

[80] *Just Playing: Game Theory and Social Contract Volume 2*, op.cit., p. 316.

[81] *Justice as Fairness: A Restatement*, op.cit., pp. 8-9.

[82] 罗尔斯在《正义新论》中提到，很多学者用“最大最小原则”（the maximin principle）或简单的“最大最小正义”（maximin justice）来指代差别原则。但是，罗尔斯强调，差别原则与在不确定情况下进行决策的最大最小规则（the maximin rule）是非常不同的，并在后者使用了“规则”（rule）一词，以与差别原则中的“原则”（principle）相区分。在论证差别原则比其他分配原则的优越性时，并不诉诸最大最小规则。罗尔斯坦言，对最大最小原则存在一个普遍的误解，即认为该原则是基于对不确定性的极端厌恶的假设，尽管正义论在描述方面的缺陷是形成这一误解的根源之一。罗尔斯在《正义新论》中承认了最大最小规则的局限性——它无法作为一个针对所有风险和不确定性情况的一般性的理性决策原则，特别是豪尔绍尼已经对此有了充分的论证。但是，罗尔斯对于最大最小原则在其理论中的地位仍然在摇摆。一方面，他认为考虑到原初状态非常特别、极端特殊的状况，最大最小规则是一个非常有价值的启发式经验法。但是，罗尔斯又提到最大最小规则被用于将差别原则与平均功利主义的比较中，这似乎意味着最大最小规则是论证差别原则的必要条件。参见*Justice as Fairness: A Restatement*, op.cit., p. 43, 97。

[83] 关于可以适用最大最小规则的条件，参见*Justice as Fairness: A Restatement*, op.cit., pp. 97-98.

[84] *A Theory of Justice*, op.cit.*,* pp. 164-183.

[85] *Justice as Fairness: A Restatement*, op.cit., pp. 103-104.