· 外国哲学 ·

道德建模：伦理问题的形式化研究方法

李　晓　冬

[摘　要] 道德建模是理性选择理论应用于伦理学研究的产物，它利用数理工具，将“道德”模型化为包含结构特征的决策程序，以此来研究伦理问题。道德建模有两种类型：典型道德场景的理性重构和道德价值的剧本化表达。前者是指用形式化的、能体现互动结构特征的博弈模型“复述”道德场景，后者是指将抽象的道德价值构想为一套直观具体的决策与互动程序。理性重构主要用于解释性目的，剧本化表达则可同时承担解释与证成两种功能。“使用囚徒困境描述合作难题”和“罗尔斯的原初状态”分别是这两种类型的典型代表。道德建模虽然在表现形式上与建构主义和思想实验相似，但在场景构想的灵活性和用途等方面有所不同，是一种有独立地位的伦理学研究方法，同时可被视为罗尔斯“道德几何学”理想的当代继承与发展。

[关键词] 理性选择理论　博弈论　囚徒困境　道德建模　道德几何学
[中图分类号] B82 - 0

何谓建模？建模就是为复杂的研究对象建立一个与其中元素存在“映射”关系的表征系统，这个表征系统一般是复杂对象在某一核心特征上的抽象与简化。建模方法主要应用于自然科学，如今已被广泛引入人文社会科学领域，比如经济学中的理性人假定、政治学中的选民投票模型以及历史学中的“满天星斗”和“漩涡模式”。① 模型化思维在哲学传统中也不令人陌生，契约论即是一例，它借助对现实道德实践的假想模拟来理解制度起源和证成道德义务，霍布斯的“自然状态”（state of nature）和罗尔斯的“原初状态”（original position）是其中典型代表。如今，随着理性选择理论（rational choice theory）的发展和引入，哲学研究中的建模

① “满天星斗”是苏秉琦提出的用以说明新石器时期中华多文明分布情况的概念，“漩涡模式”是赵汀阳提出的用于解释商周至清朝的古代中国的生长方式的概念。

方法表现出了更为形式化的特征，并逐渐应用于伦理学，笔者将这种使用形式化工具研究伦理问题的做法称为“道德建模”（moral modeling，亦可称“伦理建模”）。道德建模是对契约论建模思路的一种延续、推广和深化，其建模对象由先前的“初始状态”扩展到了一般的道德情境与实践。然而相比于其他学科领域，伦理学对建模方法的运用还远不够成熟完善，一个可能原因是“道德建模”迄今没有得到系统总结提炼。本文提出“道德建模”的理念，并对其定义、框架、类型和特点进行考察说明，为将其纳入伦理学研究方法论体系奠定基础。

一、道德建模与理性选择理论

道德建模是理性选择理论应用于伦理学研究的产物。理性选择理论是近几十年来在社会科学领域发展起来的一种以个体收益成本计算为基本方法的形式化分析框架，它最初应用于经济学，后来逐渐被引入政治学、社会学、法学和心理学领域。实际上，该理论也能够应用于伦理学。一方面，道德与审慎理性（prudential rationality）的正反关系是伦理学的传统议题[①]；另一方面，它可以为伦理学研究提供精密的分析工具。理性选择理论在研究社会互动的过程中发展出了一套能够涵盖不同决策情境与互动结构的数理模型（比如囚徒困境），道德行为是社会互动的子集，那些以后者为研究对象的工具自然可以应用于前者。这些数理模型构成了“建模”道德的“数据库”，也是本文主题“道德建模”的立足点。

理性选择理论有决策论（decision theory）和博弈论（game theory）两个分支，它们在伦理学中的应用都始于上世纪 50 年代。在 1953 和 1955 年的两篇文章中，约翰·海萨尼（John Harsanyi）——作为罗尔斯《正义论》中的论敌——首先将决策论应用于伦理学，提供了一个对平均功利主义的证明。在 1954 年的就职演说中，理查德·布雷思韦特（Richard Braithwaite）首次将博弈论应用于伦理学，主要使用议价理论（bargaining theory）讨论了分配正义问题。如今，选择理论与伦理学的结合可分为三种进路：功能主义（functionalist）、契约主义和演化博弈论。功能主义主要探讨信任、利他等道德要素在克服囚徒困境中的作用，埃德纳·乌尔曼－玛格丽特（Edna Ullmann-Margalit）的《规范的产生》（*The Emergence of Norms*）是最重要代表；契约主义将形式化的选择理论与传统契约论相结合，比如罗尔斯就将决策论用于对两个正义原则的证明，大卫·高蒂尔（David Gauthier）则使用议价理论建构自己的“协议道德”（Morals by Agreement）；演化博弈论是一种新近形式，它致力于复原许多传统道德规范与实践，代表人物主要有萨格登（cf. Sugden）、宾

① 一方面，基于长远利益的理性自利被用于证明做道德之事的合理性；另一方面，对短视的理性搭便车行为的克服又被用于论证道德存在的必要性。

默尔（cf. Binmore，1994，1998，2005）、斯科姆斯（cf. Skyrms，1996，2014）和范德施拉夫（cf. Vanderschraaf，1995，1998，2019）等人。

道德建模的基本框架来自选择理论建模社会互动的方式。一个社会互动（比如合作问题）可以从不同角度得到描述，选择理论从决策的角度切入，将社会互动理解成特定结构中的决策问题。其建模相应地包含两部分——对“决策者”的模型化和对“决策结构”的模型化，前者指行动者，后者指所处情境。首先，行动者一般被抽象和简化成一个理性质点，它有完全理性（perfect rationality）和有限理性（bounded rationality）两种类型。前者就是新古典经济学的“理性人”模型，此时行动者有充分的推理与计算能力来最大化其偏好。完全理性是一个非常理想化的设定，有限理性作为更现实的替代品出现了，此时理性被模型化为一种受限的学习与模仿能力，一般与演化动态（evolutionary dynamics）相结合。根据该进路，道德先是被理解成一种自发秩序或隐性契约（implicit contract），后来又被视为一种启发法（heuristics），此时情感、直觉和规范是一种帮助当事人在复杂的道德场景中迅速作出判断的启发式原则。（cf. Sunstein；Bicchieri）

建模的第二个要素是决策结构，这也是更重要和复杂的部分。理性选择理论试图涵盖所有类型的日常互动场景，它将这些场景归类总结，发展出了一系列决策与互动模型。最简单的决策结构是个体决策模式，它是决策论分支的研究主题。该模式的特点是决策主体与他人之间的选择不存在相互依赖关系，他不必参照任何人的行为来安排自己的行动。[①] 个体决策结构有三种类型：确定性（certainty）、风险（risk）和不确定性（uncertainty），分别描述了不同行动及其（预期）后果之间的关系。[②] 个体决策结构是最早应用于伦理学的，主要用来建模一个不偏不倚的道德立场，即借助选择的风险和不确定性来避免对某些对象的偏倚。（cf. Harsanyi，1953，1955）罗尔斯的“无知之幕”亦是此例。无知之幕遮蔽了当事人的身份禀赋等信息，使其无法制定偏爱自身的特殊原则。这即是制造了一个完全不确定的选择状态，如罗尔斯所言：“无知之幕直接引出完全不确定条件下的选择问题。”（罗尔斯，第 133 页）

更常见的决策结构是互动决策，这是博弈论分支的研究主题。该决策场景包含相互关联的多方，如甲的决定要依赖于他对乙的选择的判断，而乙的选择取决于他对甲将如何行动的预期，等等。此时博弈中的行动也被称为策略（strategy），体现了互动争胜的特点（如田忌赛马）。多人互动的决策结构包含对局方式与分布方式

① 一个典型的独立决策环境是自由市场。在市场中，消费者根据当下不变的商品价格作出是否购买以及购买多少的决定。

② 确定性环境中的后果与行动是一一对应的；风险环境中一个行动对应着不同可能后果，当事人了解这些可能后果及其发生的概率；不确定性则进一步放宽了要求，当事人不掌握不同选择的可能后果，甚至也不清楚其发生概率。

两方面，前者可被理解为双方“打交道的方式”，它由所要处理的问题决定，比如双方要对合作利润的分配进行谈判，他们的对局方式就是一个讨价还价博弈；如果两国正在进行军备竞赛，他们的对局方式就可以用囚徒困境来表示。对于“对局方式”的研究是博弈论中最为充分的部分，集中了丰富的模型。然而单纯的对局方式并不能全面描述现有互动，越来越多的理论家开始强调“分布位置”对博弈结果的影响。（cf. Axelrod；Skyrms，2004；Alexander）贾森·亚历山大（Jason Alexander）在《道德的结构演化》（*The Structural Evolution of Morality*）中系统研究了博弈参与者的分布情况对规范产生的影响，他考察了随机分布、晶格网络（lattice models）、小世界网络（small-world networks）等在影响合作、信任、公平和报复等规范形成类型、形成速率和出现概率中的作用。

一些其他的建模要素还包括人数规模和迭代次数等，它们在伦理学中的含义通常对应着“熟人或陌生人社会”和“是否存在长远利益”。比如，当社会规模过大（意味着与陌生人打交道的几率大大增加）和不存在反复互动所带来的潜在利益时，人们更不倾向于合作。由上述基本要素所构成的建模框架可由**图 1** 表示。任何社会互动都必然对上述结构中的每一部分有所规定，即使一些条件或规定是未言明的。道德交往是一种典型的社会互动，道德建模自然也遵循上述做法。结合已有研究，笔者将其描述为利用选择理论模型，将道德还原[①]或阐发为一套包含结构特征的决策程序，通过对这些结构与程序的考察来研究道德现象，其中被建模的“道德”主要有典型的道德场景与社会规范、重要的道德价值和实践特征。在本文接下来的部分，笔者将分别阐述。

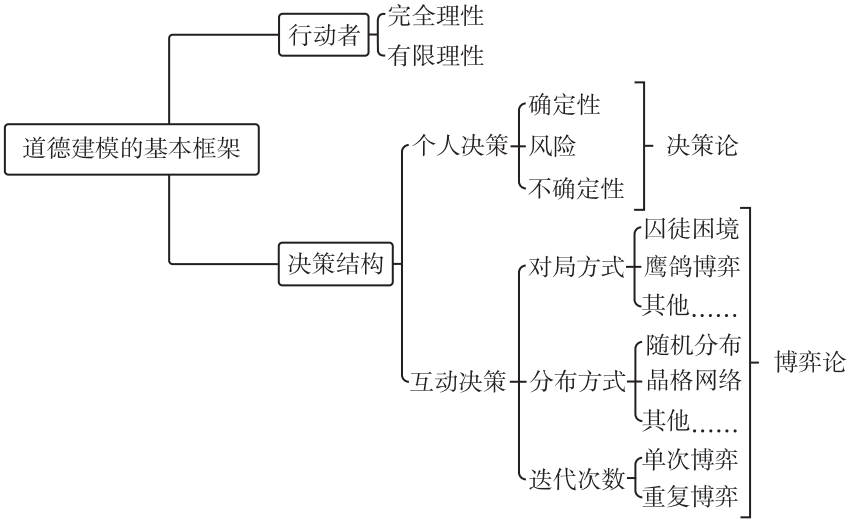

图 1　道德建模的基本框架

① 在日常用法中，“还原”意指“将事物恢复到原来的状况”，但笔者是在“抽象与简化”的意义上使用“还原”这一概念的，它用来表达“将复杂事物化简为基本要素，以显现其基本结构”的意思。

二、道德场景与规范的理性重构

在介绍完道德建模的基本方式后，我们转向它的主要类型。根据选择理论在伦理学中的应用，道德建模可分为两类：理性重构（rational reconstruction）和剧本化表达（scripted expression），根据已有研究，前者主要针对典型道德场景，后者则针对道德价值与特征。本部分介绍理性重构。上文指出道德被理解成特定结构中的决策问题，许多道德场景在时间、地点、参与者身份等方面不同，但都蕴含着相同的互动结构，理性重构就是用形式化的、能体现互动结构特征的博弈模型“复述”这些场景。借助理性重构，复杂场景背后的结构特征得以显现。另外，由于社会规范与互动结构之间存在对应关系，某类规范可能是特定结构的衍生物，因此理性重构也能模拟这些规范的形成过程。“理性重构”的说法最初来自布雷思韦特，他在《博弈论作为道德哲学家的工具》（*Theory of Games as a Tool for the Moral Philosopher*）中指出自己的工作是对明智、审慎和公平概念的理性重构。（cf. Braithwaite, p. 52）理性重构做法的最系统表述来自玛格丽特，在《规范的起源》中玛格丽特指出自己“要对规范赖以产生的社会互动状态的形式特征提供一个理性重构”（玛格丽特，第 1 页），并将其与“历史的方法”作了区分，后者是指“重构事件赖以发生的具体历史环境”，前者则是“描述规范可能产生的社会状态的基本特征”。（参见同上）

早期的理性重构是一种静态模拟。静态模拟的特点是用单次博弈直接说明道德情境的结构特征，“使用囚徒困境描述合作难题”和“辨析霍布斯自然状态最佳模型”是该类型的两个典型代表。前者中，“囚徒困境”与“合作难题”的对应关系是显而易见的，它们都体现了基于个人自利的搭便车行为破坏整体合作这一逻辑。后者中，霍布斯的自然状态是道德建模的理想对象，然而不同模型表达了不同逻辑，突出了不同特点，这就导致了关于谁是最佳解释模型的争论。目前存在三种主要模型，分别是囚徒困境、猎鹿博弈（stag hunt game）和鹰鸽博弈（hawk-dove game）。其中，囚徒困境利用“背叛”行为来说明自然状态中的先发制人策略和履约困难问题，利用双方同时背叛的占优策略均衡（dominant strategy equilibrium）来定义战争状态。（cf. Gauthier, 1969; Margalit; Kavka; Rawls, 2007）然而这一模型没有考虑人性中的激情部分，如果考虑到情感的作用就该将自然状态刻画为猎鹿博弈，该博弈体现了风险要素，能够解释“猜疑”和“求荣誉”是导致冲突的原因。（cf. Hampton; Alexandra; Palumbo; Moehler）鹰鸽博弈是激情解释的第二种类型，它将自然状态理解为一部分人是激情做主的“鹰”，另一部分人是安分守己的“鸽”，利用“鹰 – 鹰”和“鹰 – 鸽”之间的竞争说明“爱慕虚荣者”与“安分守己者”之间的冲突。[①]（cf. Sugden; Slomp and LaManna）

① 如今随着博弈论的发展，对自然状态的理性重构开始使用更为精致的工具描绘更多要素，比如不确定性（cf. Chung）、人数与对称性等（cf. Crettez）。

	坦白	不坦白
坦白	-5，-5	0，-8
不坦白	-8，0	-1，-1

图 2　囚徒困境

	猎鹿	猎兔
猎鹿	2，2	0，1
猎兔	1，0	1，1

图 3　猎鹿博弈

	鸽	鹰
鸽	0.5，0.5	0，1
鹰	1，0	-1，-1

图 4　鹰鸽博弈

如今，理性重构已经远远超出对单一情境的静态描述，而开始使用重复博弈来解释社会规范，特别是“复原”这些规范的形成过程。日常生活中人们的交往通常不是一次性的，长期利益的存在会对当事人的行为产生重要影响，因此一些促成合作和协调的规范可能会通过长时间的博弈而出现。这种情况需要重复博弈来刻画，构成了理性重构的第二种应用。一个典型例子来自罗伯特·阿克塞尔罗德（Robert Axelrod）对“合作的进化”的研究。（cf. Axelrod）阿克塞尔罗德的问题是，既然社会合作总是面临搭便车问题，那么人类合作是如何实现的？他以囚徒困境为原型组织了两次“电脑程序锦标赛”，通过计算机模拟仿真的方式将不同策略两两对局数百次，发现“首步合作，然后每一步都重复对方上一步行动”的“一报还一报”策略（tit-for-tat）总会脱颖而出。一报还一报策略又被称为“针锋相对”策略，它要求己方在对方合作或背叛之后都要立刻予以“回报”，既不姑息，也不过度惩罚，被认为具有友好、公正和宽容的特点。阿克塞尔罗德的研究在当代伦理学、经济学以及政治学中产生了深远影响，相当多的社会规范都可以通过一报还一报策略得到说明，比如“以牙还牙、以眼还眼”的同态复仇法、基督教的“黄金法则”以及孔子的“己所不欲，勿施于人”等主张。该策略的胜出很好地解释了人类社会中普遍存在的以“对等互报”为核心的朴素正义观。

道德建模的进步部分地体现于模型的进步。相比于静态模拟，阿克塞尔罗德的方法能够刻画日常交往的历时性，但不允许行动者随时调整策略，只是通过比较两两对局的最终得分来定胜负。① 然而，真实世界并不存在分数，成功的策略应该是更多人趋向的策略，失败的策略则是逐渐被放弃和淘汰的策略。一个更完善的动态模拟来自演化博弈论。演化博弈是博弈论的新兴分支，它是博弈论与生物演化思想相结合的产物，最重要的分析概念是演化稳定策略（evolutionarily stable strategy）。目前该进路最完善的研究是为休谟“正义起源于人类惯例（human convention）”的观点提供的一套演化阐释。休谟在《人性论》中认为正义是借一种“共同利益感”而逐渐建立起来的，即通过反复互动所形成的彼此有利的行为模式，这恰恰是演化稳定策略的思想。伦理学家们以鹰鸽博弈为原型，使用复制者动态（replicator dynamics）、相关均衡（correlated equilibrium）和聚点（focal point）等

① 此处通过得分来定胜负的方式仅指阿克塞尔罗德在两次正式锦标赛中所使用的方法，在一系列的后续研究中他已经使用策略动态调整方法了。

概念详细展示了正义规则是如何从克服协调冲突的过程中一步步建立的，相位图（phase diagram）则帮我们“看到”了它的成长轨迹。值得注意的是，基于演化博弈论的理性重构放宽了传统契约论理解道德形成所预设的理性条件，主要指使用有限理性代替了完全理性，最后将道德看成一套自发秩序或隐性契约。（参见李晓冬）如今，使用演化博弈论分析道德现象是一个非常有前景的研究进路。

三、道德价值的剧本化论证

剧本化表达是道德建模的第二种类型，属于更复杂的建模方式，它要求的不是模仿力而是想象力。“想象”在此体现为——这也是剧本化表达的定义——将一些非直观的抽象概念构思为一套直观具体的决策与互动程序，借助对这些程序的考察来解释和论证。此时的建模者类似于一位试图传达某种理念的剧作家，为了“深入人心”，他创作了一部以该理念为主题的“剧本”，让观众在这些展开的“剧情”中体会作者的“深意”。比如，罗尔斯在《正义论》开篇中指出“正义是社会制度的首要德性”，但他认为这种首要性目前只是一种直觉上的确信，他准备建立一套合理的决策程序，借助于它，这些直觉上的论断可以得到解释、论证和评价。（参见罗尔斯，第 3 – 4 页）传统契约论也是剧本化表达的典范，它将政治权威、政治义务与政治正当性等概念放在了有关自然状态、自然法、一致同意与理性选择等要素构成的“剧本”中来阐述。

从选择理论在伦理学中的应用历史看，剧本化建模主要有两种类型和功能，一是针对道德价值的阐释与论证；二是对道德实践特征的解释说明。本节考察第一种类型，目前该类型的建模集中于分配正义问题，主要是使用决策论——代表人物是罗尔斯——和博弈论——代表人物是布雷思韦特和斯科姆斯——将“正义”扩展成一套特定情境中的选择程序，阐发其平等主义内涵。

首先来看罗尔斯的“剧本”。罗尔斯的“剧本化”最典型地体现在反思均衡这一概念中。在反思均衡过程中，个人关于正义的深思熟虑的判断与来自原初状态的正义原则被不断调整至吻合，最终直觉判断被具象化为集中了合理限制条件的选择程序。借助该程序，罗尔斯建立了正义与平等之间的联系，按照他的说法就是“假如我们想获得一种平等学说，我们必须以另一种方式解释它，即（把它）当作一个纯粹程序性的原则”。（罗尔斯，第 400 页）

罗尔斯的论证思路是广为人知的，笔者此处重点论述该剧本与选择理论之间的联系。在罗尔斯的剧本中，各方的选择由一个代表人来执行，不存在他者立场，因此其剧本是按照决策论的逻辑展开的。首先，罗尔斯拒绝了博弈“互动”的思路。布雷思韦特是最先使用博弈论来讨论分配正义的哲学家，但罗尔斯在《作为公平的正义》（*Justice as Fairness*）中批评布雷思韦特所使用的博弈论方法认可了谈判优

势（或者说个人禀赋）对分配结果的影响，是不公平的，于是转向了决策论思路。（cf. Rawls, 1958）其次，罗尔斯使用决策论中的风险和不确定情境来刻画原初状态。这一点可能与其论敌海萨尼有某种联系，后者最先将该情境应用于道德选择问题。海萨尼是平均功利主义的代表，他在 1953 和 1955 年的论文中指出为了满足道德判断的不偏不倚要求，人们应该被置于一个决策的风险和不确定情境，由于无法偏袒自身而会采取一个公正立场。最后，原初状态中所使用的最大最小值规则（maximin rule）也来自选择理论，其原型是冯·诺依曼（von Neumann）二人零和博弈（zero-sum game）的最大最小解。按照标准的选择理论，风险与不确定情境中的决策规则应该是预期效用原则（expected utility principle），罗尔斯实际上了解这一点①，但他执意选择了最大最小规则，因为他看到了这个规则与平等之间的联系。② 冯·诺依曼 1928 年创立了二人零和博弈理论并提出了最小最大定理（minimax theorem）③，认为在二人零和博弈中理性参与者将按照最大最小准则进行选择，即选择一个不低于自己保障支付水平的策略。保障支付水平是指当事人能够单纯通过自己的选择而确保的最高收益，它是所有策略的最小支付中的那个最大值。④ 上述说明揭示了罗尔斯正义论中部分概念的原型，表明了选择理论对其剧本的影响。

博弈论是选择理论的另一分支，研究的是多方互动的逻辑。由于它能涵盖更广泛的场景，因此基于博弈论的道德建模是一种更悠久、也是如今更流行的方式。首先来看布雷思韦特基于议价理论研究分配正义的剧本。卢克和马修是毗邻而居的音乐爱好者，在每晚的相同时间，卢克会弹钢琴来演奏古典乐，马修会吹小号来演奏爵士乐。二人演奏会相互打扰，但每一方既不可能搬走也不能用合法手段阻止对方演奏。现在的问题是，如何分配演奏时间是公平的？布雷思韦特规定了二人对不同结果的偏好程度，最终能够使他们的偏好以一种基数效用（cardinal utility）的方式呈现，如**图 5** 所示。

---

① 罗尔斯指出，“最大最小值规则一般来说并不是不确定性选择的可靠指导”。（罗尔斯，第 119 页）

② 罗尔斯指出，“（正义的）两个原则与用于不确定条件下选择的最大最小值规则之间有某种联系”。（同上，第 118 页）

③ 前文提到“最大最小解”，此处又提到“最小最大定理”，这种写法并非笔误。“最小最大定理”是一个关于二人零和博弈均衡点存在的证明定理。在二人零和博弈中，由于一方所得即为另一方所失，所以从己方立场看，“我”要最大化自己的最小收益；但从对方立场看，他要努力最小化自己的最大损失，最终“最大最小值”与“最小最大值”是相同的，这是零和博弈的特点。

④ 人们首先考察自己所有可选策略的最低收益，然后比较这些最低收益，其中最大的那个收益就是自己的保障支付水平。这意味着不管对方选择什么策略，这个收益值是当事人可以通过自己的单独行动而得到保证的。这自然意味着对偶然性的排除。

		马修	
		演奏	不演奏
卢克	演奏	0，1/9	1，2/9
	不演奏	1/2，1	1/6，0

图 5　卢克和马修的基数效用

布雷思韦特提出了一个平等增益的方案，要求达成协议的好处对双方是相等的。但由于马修一开始就具有谈判优势——二人的偏好排序决定了他对“噪声”有更高的忍受程度——因此最终的分配结果给予马修更多的演奏时间，即马修平均演奏 43 天中的 26 天，而卢克演奏另外的 17 天。（cf. Braithwaite）另一个有代表性的剧本则来自约翰·纳什（John Nash）和斯科姆斯。纳什在 1951 和 1953 年分别提出并实现了后来被称为“纳什纲领”（Nash program）的方案，他认为一个关于分配的合作博弈可以被转换为一个扩展了的非合作博弈来求解和验证，这实际上已经蕴含着过程化、剧本化的要素了，称得上是选择理论中剧本化建模的第一个实例。斯科姆斯则是使用演化博弈论进一步扩充了“剧情”。（cf. Skyrms，1996，2014）他们关于分配问题的剧本是，有两个人要分一块蛋糕，二人都不是蛋糕的生产者，无人有事先的权利，如何分配完全取决于二人自己的决定。但是如果他们不能达成一致，蛋糕就会变质，双方将一无所获。现在的问题是他们会达成何种分配方案？纳什首先提出了解决这个问题的公理化方法，然后又将其转换成了一个双方各自“要价”的过程。他规定两名参与者不能交流，而是分别将自己所要求的份额呈交给一位仲裁者，如果两人所提交的份额之和不大于 100%，那么二人将得到各自份额，否则仲裁者将没收这块蛋糕。很明显，他们会谨慎地“报价”，因为过多和过少对自己都是不利的。斯科姆斯设计了一个基于演化的解决程序，在一个成员固定的群体中，每次随机挑选两人进行分蛋糕博弈，并不断重复这一过程，每个人都可以根据之前的经验调整自己接下来的报价。随着这个博弈的不断进行，斯科姆斯发现，只有要求 50% 的报价才是这个群体的演化稳定策略，即所有人如果采用了这个策略，他们将不再试图改变自己的报价；即使有人由于各种原因而采取了一个不同的报价——这个报价要么使其得到更少，要么使其一无所获——在新的一轮博弈中他也会及时更正。[①] 借助这个模型，斯科姆斯论证了平等分配是演化过程中的稳定策略。类似地，宾默尔认为罗尔斯平等分配的主张可以通过博弈论得到证明，即使他本人拒斥了这个方法。（cf. Binmore，1998，2005）

① 设想如果某刻绝大部分人采取的是要求 40% 的策略，那么有的人开始要求 60% 就会获得更大收益；同理，如果大部分人要求 60% 的份额，那么接下来转而要求 40% 会更有利。

四、道德特点的剧本化解释

上文说明了道德建模的论证功能，本部分通过对道德“敏于类别”（sensitivity to categories）特点的建模来呈现剧本化表达的解释性功能。实际上，罗尔斯的剧本也承担了解释功能，借助反思均衡，他认为原初状态的观念解释了我们的道德判断并帮助说明了我们所拥有的正义感。（参见罗尔斯，第 93 页）不过本节我们考察一个更典型的事例。首先看下述问题：（1）为什么用化学武器杀死 100 个人在人们的道德直觉中要比用常规武器杀死 10000 个人更严重？（2）为什么人权是根据人类身份而非个体所具有的不同理性、知觉或其他感受能力来划分？（3）在慈善中为什么人们更强调捐赠行为而不是善款的去向和效果？（4）在谋杀的案例中为什么人们更关心的是一个人的生命是否被剥夺而非失去了多少有效生命年（useful life years）？上述问题都体现了道德实践的一个典型特点，即道德要求或道德评价更关注的是事件的性质（或类别）而非数量。针对这一现象，莫舍·霍夫曼（Moshe Hoffman）和埃瑞兹·约利（Erez Yoeli）等人尝试使用博弈论来回答。（cf. Hoffman et al.；Hoffman and Yoeli）

霍夫曼等人将上述问题转换为一种“模型语言”，即相对于连续变化，道德规范为什么对类别上的区分有一种过度的敏感性？他们首先区分了两种类型的规范：类别规范（categorical norms）和阈值规范（threshold norms）。类别规范是用 1 和 0（分别表示“A”和“非 A”）这两个离散变量来表示的规范。“制裁使用生化武器的国家”是一个类别规范，因为它根据对象的性质或类别制定制裁的条件——相比于常规武器，使用生化武器更会受到制裁。阈值规范是用连续变量中的某一门槛值作为行动条件的规范，比如“制裁造成一万平民伤亡的国家”是一个典型的阈值规范，因为它是依据一个连续性变量（即伤亡人数）的某一门槛值（即一万）来作为制裁条件的。霍夫曼的结论是日常情形中类别规范是一个纳什均衡（Nash equilibrium），但阈值规范不是纳什均衡，其存续不稳定，不会普及开来。

上文指出，剧本化建模的实质就是将一个概念扩展为一套包含特定情境的选择程序，此处霍夫曼构造了一个协调惩罚的情节来解释道德的类别敏感特点。简述如下：设想存在甲乙丙丁四个国家①，其中丙准备入侵丁，而甲乙则准备对丙实施制裁。制裁有三个约束条件，第一，制裁是一个协调博弈。相对于无动于衷，甲乙都想制裁，但都不想单独行动。制裁是有潜在成本的（比如丙的报复），共同制裁则能分担这种成本，因此只有当一方确信对方也会行动的时候，己方才会行动。第

① 此处对原文故事有改动，原文为三个国家，为了论述方便，本文设定四个国家，且具体国家名称由甲乙丙丁来表示。

二，触发制裁的信号是一个阈值规范。甲乙之间由于各种原因缺乏直接交流机制，而商定将一个连续变量的门槛——观察到超过一万平民的伤亡——作为采取行动的条件。第三，信号存在噪声，即各方对伤亡人数的评估存在出错的可能。霍夫曼指出，当面临上述三个条件时，基于连续性变量的阈值规范不能构成一个纳什均衡，原因是任何一个门槛值都是不稳定的。假设现在甲对伤亡人数的评估正好是一万，由于存在信号噪声，他无法确定乙是否得出了相同结论，因此为了保险起见他决定等伤亡人数稍高出门槛值，比如观察到一万零一百时，再实施制裁。乙也存在相同担忧并会采取类似做法，于是最初的门槛开始向后“挪动”。当甲乙后续需要不断协调时，门槛值就会进一步提高，是不稳定的。霍夫曼进一步论证，如果甲乙采用的是一个类别规范——比如，如果丙使用了生化武器，进行制裁；否则不制裁——即使存在一定程度的信号噪声，该规范也是稳定的。上述结论证明了在协调惩罚的情形中，基于离散值的类别规范要比基于连续变量的阈值规范更可能成为一个纳什均衡，也因此存在更多存续和普及的机会。霍夫曼最后指出，如果放松上述条件，比如所处理的问题不需要很高的协调性，那么连续性标准就会出现，比如个人对于约会对象的选择就会基于年龄、身高、体重以及财富数量等连续性变量。但是如果一个规则要具有广泛适用性，那么它对协调性的要求就会上升。萨格登借助博弈论表达过与霍夫曼一致的观点，如果人们需要在大规模情形中形成对“美”的共识，那么一些客观指标比如双眼皮、白皮肤和高鼻梁这些类别标准就会脱颖而出，它们会逐渐成为“美”的代名词，替代人们对“美”的直观感受。(cf. Sugden)

霍夫曼的论证对于我们理解道德的绝对性有重要意义。人们的道德原则很少以一种权变的方式发布，比如人们会说“不允许偷盗”而不会说“不允许偷盗，除非被盗窃物品的价值微不足道”，会说“欠债还钱”而不会说“欠债还钱，除非这笔钱有更重要的用途”，会说“禁止闯红灯”而不会说“禁止闯红灯，除非路口只有你一个人”。严格来讲，后者是一种更有效率的方式，它能够在某种程度上取得“两全”，但这种方式也以一种允许例外情况的做法引入了权衡和讨价还价，开启了个人判断的“口子”。当这种例外情况或中间状态足够多时，原来的道德规则就会变成一串连续规范，根据上述博弈论证明，它在大规模的社会层面上是不稳定的。这一逻辑能够帮助我们理解义务论与功利主义之争。义务论可以说是一种典型的类别规范，它依据行为类型判断对错；功利主义则是一种典型的阈值规范，它要求根据后果——这些后果实际上构成了一串连续性变量——来作判断。功利主义能够比较好地解决效率问题，但在协调性上不如义务论，这导致——除了一些显而易见的情形——功利主义较少成为日常层面的决策标准①，而且即使在功利主义内

① 日常层面是相对于伯纳德·威廉斯（Bernard Williams）所说的“总督府功利主义”（government house utilitarianism）而言的。

部，也需要借助某些类别标准来解决预期和协调问题，比如规则功利主义的引入。相比之下，义务论则需要增加细分规则来解决不够灵活的问题。

五、建构主义、思想实验与道德建模

在利用假想情境方面，建构主义、思想实验与道德建模有相似之处，本部分通过对比来说明后者的特点，为其确立一个有价值的独立地位。罗尔斯的“康德式建构主义”（Kantian constructivism）展现了与道德建模非常类似的做法，它们都可用于对某个根本理念进行剧本化阐释。按照罗尔斯的说法，“作为公平的正义”有三个模型观念（model-conceptions），分别是道德人、原初状态和良序社会，康德式学说是对这三个模型观念的特殊诠释，最根本的一点是将“人”理解成合乎情理（reasonable）与理性、自由而平等的道德人，其他要素——良序社会、公平的社会合作、首要的正义原则和原初状态——在某种意义上都是自由而平等的道德人的“展现”。[①] 这实际上就是对自由与平等这些理念的剧本化阐释。罗尔斯指出：“通过呈现一个良序社会里作为自由和平等公民的人，建构主义程序产生了推动每一个人最高阶利益的原则，并定义了如此理解的人之间社会合作的公平条款。”（Rawls，1980，p. 570）

但道德建模并不等同于建构主义，实际上道德建模可以在下述方面被视为对建构主义的补充发展。首先，从构成上讲道德建模的内涵更丰富，理性重构是其中可单独使用的部分，这是建构主义所不具备的。研究者通过对复杂场景进行理性重构而获悉背后的互动结构，也可对社会规范进行理性重构而了解其形成过程。建构主义很明显是一个与理性重构相反的过程，因为它强调“有意义的构造”，而理性重构则强调仿真与还原。但不可否认的是，在这种构造中建构主义也可能被误认为存在与理性重构相似的做法，比如上述道德人的观念就是罗尔斯从现代民主社会的公共政治文化中抽象出来的，这种抽象表面上与理性重构有相似之处，但实质并不相同。一方面，理性重构的结果是互动结构等要素，而道德人的观念与这些要素完全不同。对于这些实践理性观念，与其说是抽象，不如说是一种从特殊角度对常识性信念的提炼。另一方面，这种“抽象”也不具备独立使用地位，它只是建构程序获取“原料”的方式。其次，道德建模的用途更广泛。建构主义主要用于证成，这一点与剧本化表达相似，但剧本化表达与理性重构还可用于解释性目的，比如囚徒困境能够揭示公地悲剧、军备竞赛等现象背后的互动逻辑，霍夫曼的剧本可以解释道德要求的绝对性特征等。

① 比如原初状态对于社会和自然偶然因素的排除就是将人“仅仅”理解成自由而平等的道德人的结果。

再次，道德建模可以刻画“开放的剧本”，这是建构主义还未实现的。罗尔斯的建构过程的结束标志是反思均衡的达成，通过不断调整建构程序与深思熟虑的判断，正义原则成为剧情发展的必然结论。这实际上是一种“封闭的剧本”。但道德建模完全可以通过引入随机变量来使结果不确定，最终的走向是开放的，这也意味着道德建模可以刻画更真实的世界。该特点也引出了二者的另一差异：道德建模相比于建构主义使用了更成熟和更广泛的形式化工具。实际上，这些形式化工具也是道德建模能够存在和实行的前提，如果这些工具不被发展出来，“建模”就无从谈起。但很明显，建构主义与这些形式化工具的联系是偶然的，它可以用也可以不用这些数理工具，只是对这些工具的利用可以丰富建构程序或剧本的内涵、扩展其长度。实际上，我们完全可以认为充分利用了这些形式化工具的道德建模是对罗尔斯所主张的“道德几何学”（moral geometry）理想的实现。最后，理性重构预设了一个更强的道德客观性立场。罗尔斯认为建构主义能够提供一个相比于它的对手——直觉主义——更好的关于道德客观性的说明，理性重构实际上可以在该问题上更进一步，因为它认为道德场景本身就存在一个客观的互动结构，而无须认为客观性仅相对于对合理限制条件的广泛接受才有效。

思想实验也是一种与道德建模类似的方法，它通过假想的情境和逻辑推理来引出悖论或冲突。首先，思想实验的一个初衷是检验和反思某一理论、原则或观点的局限性，它通过构造一些特殊的情境来“引发”被试对象的“问题”，因此在场景的设置上会更灵活和“离奇”。道德建模并没有此类目的和倾向，实际上道德建模为了解释与论证的力度——即使剧本化表达涉及对情境的虚构——它也要保证一定程度上的真实性。其次，思想实验非常依赖想象力，这同时也是剧本化表达的特点，不过道德建模还强调图形、符号等形式化工具的使用。这也使后者拥有更高的精确性与逻辑性。实际上道德建模中的公理化方法可以帮助研究者明确了解不同结论所依赖的公理条件，这是思想实验做不到的。最后，从传统上看思想实验的应用范围更广，除伦理学外，它还应用于哲学的其他领域。根据上述分析，建构主义、思想实验与道德建模——根据不同标准——处于光谱的不同位置。在场景和模型的灵活性与丰富度方面，思想实验明显高于其他二者；在目标与功能方面，道德建模既可用于证成也可用于解释，建构主义主要用于证成，而思想实验用于反思与检验；在论证道德客观性和使用形式化工具方面，道德建模强于建构主义，而思想实验则较少涉及这些。

六、结论

本文试图为当代伦理学研究展示并确立一种建模方法，就像在其他学科领域中已经运用的那样。模型化思维在哲学传统中由来已久，但形式化的建模方法却是随着选择理论的诞生才兴起的，虽然这一运用目前已经卓有成效，但缺乏系统总结。

本文将二者的结合提炼为道德建模的概念，意图为这一应用确立一个基本理念和框架。笔者将道德建模区分为理性重构和剧本化表达两种类型。理性重构的最本质工作是揭示道德互动背后的结构特征，囚徒困境是其最基本和最常见的模型，剧本化表达则将来自理性重构的互动结构视为可以再组装的“模块”，通过对这些模块的组合拼接，它可以对道德场景、规范、价值以及特点等进行解释和论证。

作为一种伦理学研究方法，道德建模有其长处和不足。关于前者，首先，由于使用了形式化分析工具，建模方法相对传统思辨方法具有直观与明确、细致与深入、严密与可信的特点。借助这些数理工具，道德建模使哲学论证变得强有力，称得上是对罗尔斯“道德几何学”理想的继承与发展。另一方面，方法论上的进展也推动了讨论问题的扩大与深化，比如正义与惯例（convention）的关系正引起学术界越来越多的关注，这得益于协调博弈、多重均衡、聚点以及相关均衡等概念的引入。此外，博弈模型的使用还促进了伦理学与其他学科，尤其是经济学、生物学以及进化论等的跨学科研究，实验哲学也从博弈论中获益良多。但道德建模仍有其不足，从已有研究看，一方面存在着一些不容易被建模的对象，包括道德动机、品德与义务等；另一方面，道德建模的规范性意义有待进一步发展，毕竟建模注重的是“原理”而非“原则”。未来道德建模可在以下几方面得到深化，一是尝试使用更多样和复杂的模型，这会使建模的广度和深度得到提升。二是对重要道德模型进行系统总结，建立“模型数据库”。道德模型是道德建模的成果，它指的是带有鲜明伦理含义、专门针对某类伦理问题的分析模型，比如无知之幕和囚徒困境。如果将重要伦理问题的道德模型总结成“集”，这将极大方便伦理学家们取用。三是研究道德与决策结构之间的关系。道德建模预设了道德要素是相关决策结构的衍生物，该观点构成了道德建模——尤其是剧本化表达——的实质性预设，有必要进一步探讨和论证。四是借助博弈论等工具回答道德要求的规范性来源问题，并尝试确立各种规范性原则。最后笔者断言，道德建模是一种有价值的研究方法，值得被吸收进当代伦理学方法论体系。

参考文献

李晓冬，2019 年：《规范起源的社会惯例论——从演化博弈论的方法看》，载《哲学研究》第 12 期。

罗尔斯，2009 年：《正义论》（修订版），何怀宏、何包钢、廖申白译，北京：中国社会科学出版社。

玛格丽特，2020 年：《规范的产生》，秦传安、秦忆译，上海：上海财经大学出版社。

Alexander, J., 2007, *The Structural Evolution of Morality*, Cambridge: Cambridge University Press.

Alexandra, A., 1992, “Should Hobbes’s State of Nature be Represented as a Prisoners Dilemma?”, in *The Southern Journal of Philosophy* 30 (2).

Axelrod, R., 1984, *The Evolution of Cooperation*, New York: Basic Books.

Bicchieri, C., 2006, *The Grammar of Society*, Cambridge: Cambridge University Press.

Binmore, K., 1994, *Game Theory and the Social Contract Vol. 1: Playing Fair*, Cambridge, MA.: MIT Press.

1998, *Game Theory and the Social Contract Vol. 2: Just Playing*, Cambridge, MA.: MIT Press.

2005, *Natural Justice*, Oxford: Oxford University Press.

Braithwaite, R., 1955, *Theory of Games as a Tool for the Moral Philosopher*, Cambridge: Cambridge University Press.

Chung, H., 2015, "Hobbes's State of Nature: A Modern Bayesian Game-Theoretic Analysis", in *Journal of the American Philosophical Association* 1 (3).

Crettez, B., 2017, "On Hobbes's State of Nature and Game Theory", in *Theory and Decision* 83 (4).

Gauthier, D., 1969, *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*, Oxford: Clarendon Press.

1986, *Morals by Agreement*, Oxford: Clarendon Press.

Hampton, J., 1986, *Hobbes and the Social Contract Tradition*, Cambridge: Cambridge University Press.

Harsanyi, J. C., 1953, "Cardinal Utility in Welfare Economics and in the Theory of Risk-taking", in *Journal of Political Economy* 61 (5).

1955, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility", in *Journal of Political Economy* 63 (4).

Hoffman, M. and Yoeli, E., 2022, *Hidden Games*, New York: Basic Books.

Hoffman, M., Yoeli, E., and Navarrete, C. D., 2016, "Game Theory and Morality", in T. K. Shackelford, R. D. Hansen (eds.), *The Evolution of Morality*, Cham: Springer International Publishing.

Kavka, S., 1983, "Hobbes's War of All against All", in *Ethics* 93 (2).

Margalit, E., 1977, *The Emergence of Norms*, Oxford: Clarendon Press.

Moehler, M., 2009, "Why Hobbes's State of Nature is Best Modeled by an Assurance Game", in *Utilitas* 21 (3).

Nash, J. F., 1951, "Non-cooperative Games", in *Annals of Mathematics* 54 (2).

1953, "Two-Person Cooperative Games", in *Econometrica* 21 (1).

Palumbo, A., 1996, "Playing Hobbes. The Theory of Games and Hobbesian Political Theory", in *UEA Papers in Philosophy* 8.

Rawls, J., 1958, "Justice as Fairness", in *The Philosophical Review* 67 (2).

1980, "Kantian Constructivism in Moral Theory", in *Journal of Philosophy* 77 (9).

2007, *Lectures on the History of Political Philosophy*, S. Freeman (ed.), Cambridge, MA.: Belknap Press of Harvard University Press.

Skyrms, B., 1996, *Evolution of the Social Contract*, Cambridge: Cambridge University Press.

2004, *The Stag Hunt and the Evolution of Social Structure*, Cambridge: Cambridge University Press.

2014, *The Evolution of the Social Contract* (2nd ed.), Cambridge: Cambridge University Press.

Slomp, G. and La Manna, M. M., 1996, "Hobbes, Harsanyi and the Edge of the Abyss", in *Canadian Journal of Political Science* 29 (1).

Sugden, R., 1986, *The Economics of Rights, Co-operation, and Welfare*, Oxford: Blackwell Press.

Sunstein, C. R., 2005, "Moral Heuristics", in *Behavioral and Brain Sciences* 28 (4).

Vanderschraaf, P., 1995, "Convention as Correlated Equilibrium", in *Erkenntnis* 42 (1).

1998, "The Informal Game Theory in Hume's Account of Convention", in *Economics and Philosophy* 14 (2).

2019, *Strategic Justice: Convention and Problems of Balancing Divergent Interests*, New York: Oxford University Press.

(作者单位:山东大学哲学与社会发展学院)

责任编辑:韩 骁

Chan Buddhism, they accurately reflected the prevailing shortcomings of the learning of Wang School in the middle and late Ming period. Due to a lack of theoretical development in the profound aspects of the School of Mind, the philosophical principles and practical methods of the Two Xi's teachings inevitably became ambiguous, while the practice of "shattering mental phenomena" often devolved into emotional and intellectual indulgence.

Moral Modeling: A Formalized Analytical Method for Ethical Issues

Li Xiaodong

Moral modeling is a product of applying rational choice theory to the study of ethics. By employing mathematical tools, it models "morality" as decision-making procedures with structural features, thereby providing a means to study ethical issues. Moral modeling can be divided into two types: rational reconstruction of typical moral scenarios and scripted expression of moral values. The former refers to the use of formalized models of game theory that employ interaction structures to "restate" moral scenarios; the latter refers to conceiving of abstract moral values as a set of intuitive and concrete decision-interaction procedures. Rational reconstruction is explanatory in the first place, while scripted expression can serve both explanatory and justificatory functions. The use of the Prisoner's Dilemma to describe the problem of cooperation and Rawls' original position are the paradigmatic examples of these two types, respectively. Although moral modeling resembles constructivism and thought experiments in its form, it differs in its flexibility of scenario construction and functions. It thus constitutes an independent methodological approach in ethics and may be regarded as a contemporary continuation and development of Rawls' ideal of "moral geometry".

How does Jin Yuelin's Theory of Natural Law "Tolerates Exceptions"

Zhang Liying

In his *Theory of Knowledge* and other works, Jin Yuelin explicitly distinguished between natural laws and their realization. He argued that there is no conflict among natural laws in themselves, but conflicts arise when all these laws become realized at a particular time and place, and the patterned set of ideas can greatly enhance the reliability of inductive conclusions. Jin Yuelin's theory of natural law can provide new perspectives for default reasoning and *Ceteris Paribus* which both focus on how to "tolerate exceptions". Different from many research approaches which put the question "how to express the exception tolerance of law-like propositions" to the first place, Jin Yuelin's theory of natural law shifts its focus from expressing exceptions to comparing and selecting natural laws when conflicts arise in their actualization. Following Jin Yuelin's theory, we can raise further questions, such as: when conflicts occur in the actualization of different natural laws, are there some rules for priority? In what specific way does the patterned set of ideas affect acquisitions of inductive conclusions and discoveries of nature laws?