从解释到证成

——最优解释方法是否可以充分证成正义理论？

叶金州 *

摘要：基础主义和以“反思平衡”为代表的融贯论代表了证成规范性正义理论的主要进路，但它们都存在重要缺陷。“最优解释方法”作为第三种进路被提出，意在为正义理论提供更好的证成。然而，这一方法的效力建立在将事实性真理与规范性理论（主张）未加严格区分的基础之上，并预设了普遍道德真理的存在，但这是缺乏依据的。此外，在该方法中规范性理论的规范性权威何以建立这一问题完全缺席，是一个需要弥补的重大缺陷。

关键词：最优解释方法；证成；规范性权威

一、作为第三种进路的最优解释方法

在一种规范性理论（例如正义理论）的建立过程中，人们的深刻道德直觉如应得、自由、平等、人道、公平等往往起着举足轻重的作用。然而，这些道德直觉自身是否合理？它们在证成规范性理论的过程中起到何种作用？以及，更为根本的是，这些直觉是否应当起任何作用？对于任何完备的规范性理论来说，这些都是不可回避的元（meta）问题，它们必然以或显或隐的方式在这些问题上持有某种立场。最为关键的是，道德直觉是否可以为规范性理论提供终极证成？抑或，直觉为规范性理论只能提供

* 作者简介：叶金州，华中科技大学哲学学院讲师，研究方向为政治哲学与道德哲学。电子邮箱：yejz@hust.edu.cn。本文系教育部人文社会科学研究一般项目（青年基金）“公共理性的内在逻辑与适用范围研究”（20YJC720027）阶段性研究成果。

初步证成？又或者，直觉本身并不能对规范性理论形成真正的约束，反而应由规范性理论提供检验标准，从而可以剔除含混错乱的直觉，纯化并提升合理的直觉？一种规范性理论如何在这些进路中进行选择，在很大程度上决定了该理论能否获得充分的证成。

这便触及了如何证成一个规范性理论这一基础性问题。一个完备的规范性理论首先必须满足最低的形式化要求，即其内部各要素要实现融贯一致。这是因为，如果一个规范性理论内部各元素之间存在紧张甚至冲突，它便无法对行动者提出一致的行动要求。但除此内部形式化要求之外，更为重要的是它需要在目标行动者面前真正建立规范性权威，否则它还不算是一个真正的“规范性”理论。而为了建立这种规范性权威，候选理论必须严肃对待人们多样而深刻的道德直觉。在如何处理这些道德直觉的问题上，道德哲学（政治哲学）中形成了不同的进路：有将规范理论之证成完全建基于某种特定道德直觉之上的基础主义，也有以道德直觉和规范理论之间经过相互调整而建立融贯关系的融贯论，个中代表便是罗尔斯倡导的“反思平衡”。长期以来，这两种理念代表了规范性理论证成的主流进路。近来有些学者提出了第三种进路[1]，倡导“最优解释方法（inference to the best explanation）”。根据这种方法，道德直觉并不为规范性理论提供实质性证成，而只是充当一个“解释”约束，即一个胜任的规范性理论应当能够充分说明（尽管并不一定直接接受）人们为何具有这些根深蒂固的道德直觉。如果这一进路可以取得成功，那它将为道德哲学（政治哲学）带来方法论变革，势必为新理论的建构提供助力。然而，最优解释方法真的具备这样的证成作用吗？其倡导者只是预设了它的效力但对其基础却鲜有说明。本文试图填补这一空白，检验最优解释方法作为一种证成规范性理论的方法是否具备预期的证成效力。

由于最优解释方法的倡导者是在建构一种新的分配正义理论过程中系统地使用了这种方法，本文亦将依托这一具体案例来对其效力进行检验。

---

［1］这一进路的集中表述见葛四友：《分配正义新论》，北京：中国人民大学出版社，2019。其概要介绍见葛四友：《论道德思想试验中的直觉错位与后果主义的证成》，《社会科学》，2019（11）：116—127。

具体而言，葛四友教授在《分配正义新论》中对几种主流的（义务论）分配正义理论做出了系统批判和检讨之后，提出了自己的替代性正义理论。不惟如此，他还为该正义方案的阐明和证成设计了新的方法论设施。这与罗尔斯当初的做法颇为相似：为了取代功利主义正义观，罗尔斯不仅创造性地提出了公平正义观，他还引入了原初状态、反思平衡和现实的乌托邦等全新方法论设施来更好地阐明和证成这种新的正义观。遗憾的是，研究者对这一方法论层面的创新关注不甚充分。葛著引入了最优解释方法，这既是对其实质性正义理论进行证成的必要支撑，也可视为对罗尔斯的反思平衡方法的继承和超越。

这种替代性正义理论的一个鲜明特点是它对现实人性的充分重视。在当代主流的规范性理论建构中，人性是被边缘化甚至被直接拒绝的考量，这不能不说是某种形式的进步主义的巨大胜利。但这种主流的做法忽视了人性现实对规范性原则的构成性约束：人性现实并非仅仅对理想规范理论的应用范围形成制约，它在更深的层次上对该理论的内容 / 原则也施加了影响[1]。葛著强调了这种构成性约束，意在表明：规范性原则自身的基础来源于人性本身，即来源于现实人性所支撑的基本动机结构以及由此而生成的基本道德直觉。在他看来，主流的分配正义理论事实上都依赖于人道、公平这两种基本直觉，而它们都可由休谟的“有限利他心”这一现实人性刻画所概括。

在构建其基于公平和人道的分配正义理论时，葛著选取了后果主义的思路。一般说来，正义理论往往依赖基本权利或资格等考量，它们具有强烈的义务论特征。这些权利或资格往往被赋予某种绝对性或优先性，而后果主义却无法实现这种绝对性；恰恰相反，后果主义原则可以允许各种权利因总体福利（或类似的后果性指标）而被牺牲。也正因为此，后果主义（功利主义）在当代有关正义的讨论中饱受批评。在葛著中，后果主义的思路却可以和由公平与人道所表达的分配正义相兼容，不仅不会引发冲突，反而可以克服义务论正义理论的一个根本困难，即义务论正义理论无

[1] 例如见 Miller, David: “Political Philosophy for Earthlings.” In *Justice for Earthlings*, Cambridge & New York: Cambridge University Press, 2013: 16—39.

法协调不同的根本道德直觉之间的不一致。而葛著之所以可以克服这一困难，在根本上是因为它使用了 最优解释方法从而实现了理论与直觉的分离，而不必像义务论那样将道德直觉作为正义理论的奠基。

不同于规范性理论中常见的基础主义或融贯论等进路，最优解释方法对人们所持有的各种道德直觉并不直接接受，也不将规范性（正义）理论的证成建基于这些直觉之上，而是认为它们只具有初步（prima facie）的证成作用，原则上可被更宏观更一贯的理论所否决（而非仅仅压倒）。如此一来，它们之间的紧张和冲突便不会造成理论内部的不一致。在这个意义上，最优解释方法应当具有与实质性正义理论同等重要甚至更为重要的地位。离开了这一方法自身的合理性，这种替代性正义理论也便难以真正建立。接下来本文将考察最优解释方法是否可以成为一种对规范性理论进行证成的合理方法。

二、最优解释方法的构成

所谓最优解释方法，其传统的应用领域是科学理论的检验与证成，属于知识理论中的一种思路。与其他知识论思路相似，这一方法处理的是直觉可以对理论提供何种证成的问题。“这种方法认为，一种理论，如果它能对人们的直觉给出最好的解释，说明人们为什么会有这样的直觉，那么它就是最好的理论，是人们应该接受的理论”。有了这样一种判定标准，“如何看待直觉与真理之关系的问题，也就是直觉在理论证成中能起到什么作用的问题”[1]。便有了解答。当然，该方法并非对此问题的唯一解答。关于直觉和真理之间的关系，尤其是直觉可以为理论提供怎样的证成，不同的知识论思路提供的解答各不相同[2]。类似地，在规范性理论中，道德直觉对于理论可以起到何种证成作用，不同的理论家也有不同的看法。最优

[1] 葛四友：《分配正义新论》, p.12。
[2] 极端的观点甚至会认为，直觉并不能提供任何证成，至少哲学家尚未为这种直觉的证成作用提供证成，但却一直在使用，这是哲学的耻辱，例如见曹剑波：《哲学直觉方法的合理性之争》,《世界哲学》, 2017（06）：52—60。

解释方法的倡导者认为，直觉可为规范性理论提供三种不同类型的证成[1]：

（1）结论性证成：“直觉到的内容如何，那么情况就如何”。

（2）初定（prima facie）证成：“概率意义上的证成，指的是直觉到的内容有一定的概率为真。一般来说，直觉越强，且没有相冲突的直觉，则其内容为真的概率就越高”。

（3）初步（pro tanto）证成：权重或分量意义上的证成，“一般来说，直觉越强，且没有相反的直觉，那么支持命题的权重就越大”。

（1）所表达的便是直觉主义，与理论证成中的基础主义方法最为契合，其基本理据在于作为基础的直觉本身是自明的或不可错的，从而可以推出由其所生发的理论的可靠性 / 真理性。但不论在事实领域还是价值领域，这样不可错的直觉在当今时代难以再被广泛承认；与之相应的是，基础主义方法饱受攻击。直觉的初定证成与初步证成作用表面看起来很容易混淆，但它们之间存在一个关键的区分，即相关直觉的合理性是否可以被（彻底）取消。提供初定证成的直觉自身具备的是初定合理性，也即其为真的可能性是一个概率，范围从 0 到 1。而提供初步证成的直觉自身具备的是初步合理性，即其内容具有一定的权重是正确的。若一个直觉具备的是初定合理性，那么当相反证据足够强时，其合理性在原则上可以被彻底取消（即降至 0）；例如，如果一个直觉的内容是“张三杀了李四”，但人们确认李四并没有死，那么该直觉便被完全推翻。而如果一个直觉具备的是初步合理性，那么即便它在更强的竞争性直觉面前被压倒了，但其合理性并不会因此被取消；例如，如果一个直觉的内容是“应当投资项目 A，因为它有 10% 的利润”，但当决策者发现项目 B 有 20% 的利润之后，他便会认为项目 B 而非 A 更应当获得投资，即“应当投资项目 A”这一直觉已被“应当投资项目 B”所压倒，但即便如此“应当投资项目 A”这一直觉所具有的权重（由 10% 的可能利润提供理据）并不会因此就被取消[2]。

---

[1] 对直觉这些证成类型的总结，见葛四友：《论道德思想试验中的直觉错位与后果主义的证成》, pp. 119—20。

[2] 葛四友：《分配正义新论》, pp.13—4。

需要说明的是，具备初定合理性的直觉“特别明显地显现于事实性问题”，其可提供的是“概率意义上的证成”，强调的是“直觉内容为真的概率有多大”；以此为基础而为理论提供证成的方法便是“科学理论中采用的最优解释推理”。与之相对的是，具备初步合理性的直觉“一般显现于规范性信念，主要是针对行动的合理性或正当性”，其强调的是“支持某个规范性命题（或行动）的分量有多大。”“初步证成作用大概与罗尔斯在《正义论》中采取的那种融贯论相对应”。[1] 现在我们要处理的主题是分配正义，它属于规范性领域而非有关事实的科学领域，似乎应当采取第（3）种进路，而非第（2）种。但葛著明确指出其采用的方法正是最优解释方法[2]，这明显需要一个解释。

之所以要选择最优解释方法，主要是因为这里要证成的是一种后果主义的正义理论。相较于其他正义理论来说，后果主义正义理论是一种“根本性的道德理论”，在对其的证成中“道德直觉的直接作用是有限的，一般来说只能提供初定证成的作用。”[3] 道德直觉起到的作用之所以是有限的，是因为道德直觉本身并非自明或不可改变的；恰恰相反，它会因为背景环境的变化而相应地改变，即为后者所塑造。而一种根本性的道德理论不应当被某种特定背景环境塑造的道德直觉所约束，因为后者是偶然的、可变的，即不具备根本性。后果主义之所以经常受到各种道德直觉的挑战，相当程度上是因为这些直觉来源于一个特定的背景环境，即现实人性（现代社会）；而若将这个特定的背景环境按照后果主义的要求进行调整（理想化），那么人们也会自然产生与后果主义相符的道德直觉，而不会对其造成挑战。在理想的背景环境下，人是全善的人，是“具有充足的利他主义精神的人，是有（形式意义上的）善良意志的人，他有足够的动机去履行自己的道德义务，总是能做自己该做的事情”[4]。例如，在与后果主义相关的理想情境中，人们会自愿地牺牲自己，从而成全更大的总体福利，

[1] 葛四友：《论道德思想试验中的直觉错位与后果主义的证成》, p.119—120。
[2] 葛四友：《分配正义新论》, p.13。
[3] 葛四友：《论道德思想试验中的直觉错位与后果主义的证成》, p.127。
[4] 葛四友：《论道德思想试验中的直觉错位与后果主义的证成》, p.120。

甚至可以形成自杀（以促进共同善）的义务感。

这里特别需要强调的是“（形式意义上的）善良意志”这一刻画，它的提出是为了解决现实中的人经常会在做道德所要求的事情上出现动机不足这一问题。在对后果主义的经典挑战“电车难题”中，一个无辜的生命要为救 5 个人而被牺牲掉，这样做尽管从后果主义的计算上是正当的（甚至是应当的），但它明显与我们的直觉相悖。然而，这里的直觉并非后果主义理想化背景环境下的直觉，故而不应当拿来驳斥后果主义理论本身。若是将背景环境理想化，即每个人都会自觉地做道德上正确的事情（这由相关的理论来定义），那么上述直觉与理论相悖的现象便不会出现了[1]。从形式上讲，善良意志保证了理论所确定的正确之事会被人们真心地接受并自愿地执行，也即成为他们直觉的一部分。

如果这一思路是正确的话，那么对于任一种根本性道德理论（不论是后果主义还是义务论）来说，道德直觉都不会造成真正的挑战。因此，直觉在对规范性理论的证成中便无法起到实质性的作用；充其量，道德直觉可以提供一个附加的“解释约束”，即，道德理论尽管不需要接受这些道德直觉或在其基础上得到证成，但一种完善的道德理论最好能够为这些既定的深刻道德直觉提供一个解释，说明它们为何会根深蒂固地存在。在这样一种进路中，道德直觉与道德理论之间仅有一个（被）解释关系，而全无证成性关联。道德直觉所起的作用，仅仅是锦上添花：“对于一种根本的道德理论来说，它最终的证成，除了内部的融贯性外，还源于它能够最好地解释我们的各种道德直觉，至少要比其他理论做出的解释更好”[2]。

三、理想理论与现实情境

以上的分析针对的是理想背景环境下的道德直觉，但我们所处的具体背景环境却是现实的。那么在现实世界里，这些根本性的规范性理论（例

[1] 更细致的分析见葛四友：《论道德思想试验中的直觉错位与后果主义的证成》, p.122。
[2] 葛四友：《论道德思想试验中的直觉错位与后果主义的证成》, p.127。

如后果主义）应当如何理解呢？是否应当将它们的无限要求直接施加给现实中的人们，还是要对这些要求作出调整，以适应现实的人性条件？以后果主义为例，它的根本要求是促进行动后果的最大化。这是否意味着每个人每时每刻每个行动都应当不计个人得失地遵循这一指导，即将后果主义作为具体的行动指南？抑或，应当将后果主义视为一种间接的评价标准，转而评判现实人性条件下的人类行动网络作为一个总体是否可以产生更大的总效用？事实上，“在不少条件下，后果主义作为决策程序反而会带来更坏的后果 …… 后果主义作为一种评判标准会要求后果主义不能直接作为决策程序。人们的动机不够显然就是这样的一个条件。”相比理想世界，现实世界的人们并不具备全善的特征，即不能自愿地满足道德所提出的无限要求，若强行将道德要求施加于他们，结果反而会适得其反。依据后果主义的基本思路，最重要的考量是总体后果是否最佳，故而，道德需要降低自身的要求，使得履行者所付出的代价不必太高，从而人们可以有足够的动机去履行。于是，“动机不足会改变道德原则本身”[1]。

现实的人性并非全善的，即人们的（形式意义上的）善良意志是有限的，这就造成了现实中人们的动机结构与理想设定下有所不同。如果休谟对人性的描述是准确的话，人们具有且只具有“有限的同情心”。这表明他们在基本动机结构上既不是全然自利的、但也不是彻底利他的。为了在这样的现实条件下实现后果主义的基本宗旨，道德的要求必须与人们的动机结构相适应，即在敦促引导人们发扬其利他心的同时，也认可并尊重其自利心的诉求。为了充分实现这一点，后果主义纲领会允许甚至要求设立一些基本的权利，如自我所有权、一定形式的财产权和自由交易权等。尽管初看起来这些权利具有显著的义务论特征，但它们是可以得到后果主义（即后果之最优化计算）的证成的[2]。尽管从根本上讲，后果主义并不承认这些权利具有内在价值，但它可以认可它们的工具性价值。至少在现实的人性以及由现实人性所组成的复杂社会面前，这些具备工具性价值的权利是不可或缺的。

---

［1］葛四友：《论道德思想试验中的直觉错位与后果主义的证成》, p.125。
［2］葛四友：《论道德思想试验中的直觉错位与后果主义的证成》, p.125—126。

这里需要强调的是，在现实人性条件下，人们在与之相适应的制度、道德背景环境下会形成相应的道德直觉，例如强调按照贡献进行分配的公平直觉。如此一来，这些直觉在相当程度上来源于现实人性本身，同时也在一定程度上来源于与现实人性相适应的社会文化和制度环境的塑造。此时若还是以直觉本身作为基点而试图为规范性（正义）理论提供结论性或初步证成，那便是本末倒置了。除开不可变更的基本人性结构[1]所造成的影响之外，这些直觉的主要塑造者便是奠基了社会制度和文化（或罗尔斯所说的社会基本结构）的主导性规范理论，故而，这些理论自然可以说明 / 解释人们为何会拥有这些直觉。由于葛著倡导的正是这种与现实人性（尤其是休谟所描绘的现实人性）相适应的正义理论，那么它采用最优解释方法来为其提供佐证便也顺理成章了。

除了上述理由之外，还存在一个更为显白的理由将诸道德直觉的证成作用视为初定的而非初步的（遑论结论性的），即基本道德直觉的多元性。同样一个理论（包含对其的论证）在不同的人看来，会有截然不同的判断，这是由于他们并非拥有相同的道德直觉。而且，即使我们聚焦同一个人，他在不同的情境下所持有的直觉也可能是不同甚至冲突的[2]。若要以直觉作为立基支点而对候选理论进行证成，那么人们很难取得一致的、令所有人信服的结论。于是，我们需要处理一个关键的问题：“面对众多道德直觉，特别是有可能相冲突的道德直觉，如何加以取舍？”[3]不论是基础主义还是融贯论的方法，都无法充分解决这一问题，因为在它们的思路中已经预设了作为起点的某种道德直觉的（至少是初步）合理性。即使在融贯论方法中经过“反思平衡”的调整，这些直觉的初步合理性也不会被取消。因此，在面对与之相反的直觉时，该理论便只能要么放弃自身的立场，要么忽视这种相反的直觉。相较之下，最优解释方法却可以有效地避免这一难题，因为它只接受这些直觉的初定（即原则上可被取消）合

[1] 关于人性这一说法是否有意义，或者即使有意义，人性是否是固定而一成不变，主张激进方案的理论家通常持否定态度。更深一层的问题是，即使人性不可改变，其是否应当成为构建规范性理论时的必要参考？

[2] 例如，在电车难题的不同表述中，同一个人也会形成不同的判断。

[3] 葛四友：《分配正义新论》, p.11。

理性，而将重点放在寻找一种对所有这些直觉的解释。提供解释意味着并不将（初看起来）相反的直觉直接忽视，同时也意味着并不直接接受该直觉，而是用一个理论来说明人们为何会持有如此的直觉。此处一个典范性的例子是筷子入水给人们造成的筷子弯曲的直觉，事实上并不是筷子真的变弯了，也并非观察者在撒谎，而是光线经过水的折射之后给人们造成了这样的感觉[1]。在这个例子中，正是“光的折射定律”这一理论为我们判定这一直觉的性质提供了可靠的依据。也正因它可以最好地解释相互冲突的直觉（水中筷子变弯、但拿出后依然是直的，以及筷子水中看着是弯的但伸手去摸却是直的），所以它是一个得到充分证成的科学理论。在这里，该理论并不是与直觉 / 现象直接相关或直接建立在它们之上的规律总结，而是一种外在于它们的解释机制。与此相似，在对道德直觉进行解释的时候，我们往往也需要引入外部的资源。“有时，这种关于道德直觉的根本分歧在哲学内部无法解决，需要引入其他学科，达成更大、更融贯的理论体系”；一种理论若是“与其他学科理论产生严重冲突，就会大大削弱自己的合理性”[2]。

由于这里所要处理的并非事实性的科学理论而是规范性的正义理论，故而它在使用最优解释方法时将与之相关的一些关于科学理论的预设也带了进来。“最优解释方法预设我们的认知对象具有客观规律性，本书则预设存在普遍有效的分配正义。至于这种分配正义是帕菲特认为的那种道德事实，还是罗尔斯所说的建构，抑或是德沃金所强调的是对伦理实践的最好诠释，这里可以保持中立”。尽管有关分配正义的理论“还没有发展到物理学或几何学那样成熟，能够找到定理或公理”，但它依然以近似的思路为一种新的正义观念提供证成，即说明其“能够得到更多的证据，能够具有更好的理由”[3]。这里作为“证据”而出现的主要是人们的道德直觉。“尽管最佳解释方法并不预设道德直觉拥有任何绝对的特权地位，但道德直觉依然非常重要，因为它有筛选作用。正义观念并不一定要符合道德直

[1] 葛四友：《分配正义新论》, p.13。
[2] 葛四友：《分配正义新论》, p.14—15。
[3] 葛四友：《分配正义新论》, p.15。

觉，但它应该有能力解释各种道德直觉”[1]。相较之下，那些不能充分解释人们道德直觉的正义理论，或只能解释其中一部分直觉的理论，便没有得到足够的证成。

四、理论与直觉的分离

如上所示，最优解释方法被用于证成一种着眼于现实人性的、后果主义的分配正义理论。它体现了一种论证思路上的创新，但也暴露了若干重大缺陷。将最优解释方法方法从知识理论领域拓展到规范性理论领域是个根本性的范式创新，它有力地解决了由基本道德直觉的多元性所带来的两难抉择。以往的正义理论往往立基于某种特定的道德直觉，但忽视其他道德直觉的作用，从而导致其难以得到广泛的承认，甚至会遭到严重的抵制（例如柯亨的平等主义、诺齐克的自由至上主义等）。虽然有一些正义理论也兼顾了主要的道德直觉（如德沃金的资源平等主义和罗尔斯的公平正义理论），但由于它们未能在基本方法的层面上与这些直觉保持应有的距离反而是直接或间接地建基于其上，这些理论容易遇到内部动机不一致（即其所奠基的直觉之间相互冲突）的问题[2]。在使用了最优解释方法之后，正义理论的合理性便不依赖这些直觉自身的合理性，而只需要为它们提供解释 / 说明即可，即使这一说明完全取消了原有直觉的合理性也不会造成问题。一言蔽之，最优解释方法实现了正义理论与道德直觉间的分离。

在传统的规范性理论建构中，基本的道德直觉往往都占据着相当核心的地位，即使不被视为理论的终极证成基础，但也绝不能够轻易地被取消，至少其初步合理性要得到承认。最优解释方法在根本上否定了这一传统思维范式，转而将道德直觉降格为类似科学实践中的（未加校准的）观察数据，其合理性只是初定的，有待校准以及系统理论的解释。在这样一种的新的范式下，道德直觉在（规范性）理论证成中的地位大为下降，在

[1] 葛四友：《分配正义新论》, p.49。
[2] 此处依然同情性地按照葛著的思路来定性主流的正义理论，但这并不代表笔者完全同意这些定性，详见后文对“反思平衡”的分析。

事实上近乎被取消。尽管人道、公平两种道德直觉依然被作为基本元素来构建正义理论，但在最优解释方法这一基本范式之下，这两种直觉所能提供的证成也只是初定的，而非初步的遑论终极的。原则上它们可以被取消，被替换为与规范性理论更为契合的其他道德直觉，例如不计回报的自我牺牲。之所以没有这样做，乃是因为一方面现实人性（休谟刻画）具有一定的合理性，另一方面要证成后果主义的正义理论。后果主义的总目标是提升总体福利（或相似的其他后果），在人性现实无法改变的前提[1]下，它便寻求在现实人性这一约束之下如何实现后果的最大化。如此一来，现实人性和最大化考量所奠基的社会制度框架便共同塑造了人们的道德直觉，因此这种直觉自然可以得到后果主义正义理论的解释。

在最优解释方法的范式下，有一个区分变得非常重要，即作为行动指南的规范性理论和作为评判标准的规范性理论之间的区分。只有将规范性理论的作用转为评判标准而非行动指南，其与道德直觉之间的关键分离才能得以实现；也即，正是由于这一区分的存在，一种理论才有可能既不直接接受道德直觉，但又可以对其提供解释。例如，作为评判标准的后果主义并不要求每个行动都严格按照后果最大化的计算去进行；恰恰相反，在一些限制条件（例如现实人性）下，机械地执行这一标准反而会在总体上妨碍后果的最大化。于是，在后果主义的总体纲领下，一些表面上反对后果主义的权利可以（应当）被设立。这些权利可以支持人们的道德直觉，但它们并不完全接受这些直觉中蕴含的（似义务论的）绝对性。它们需要在后果主义的总体思路下被设定一个限度，而限度究竟设定在何处（例如人道 vs 公平的占比）取决于具体条件下何种安排可以实现后果最大化。在这样的规范性框架中，后果主义的最大化考量并不直接作用于人们的具体行动，而是作为这些行动所要遵循的规则、制度等的评判标准。如此一来，后果主义的正义理论与最优解释方法便实现了契合。这种契合的关键在于二者在基本思路上都要求一种分离，即直觉（未经校准的数据）与理

[1] 关于人性是否具有可塑性，不论科学界还是哲学界都存在争议。与之相关的问题是，即使人性可以改变，敦促某个方向的改变是否合理？葛著在此问题上默认了休谟所刻画的人性现实是合理的。

论之间要保持一个关键距离。

当然，最优解释方法并非后果主义理论的专有证成方法，只要某个规范性理论能将其作为行动指南的部分与作为（间接）评判标准的部分有效地分离开来，从原则上讲，它都可以援引最优解释方法为其辩护[1]。然而，主流的几种义务论正义理论却无法实现这一点，因为在这些理论中并不存在道德直觉与正义理论之间的分离，它们要么直接要么间接地将自身的合理性奠基于这些道德直觉自身的合理性之上。从另一个角度来看，这些正义理论主要以直接的行动指南的方式呈现出来[2]，而非作为对行动的间接评判标准。故而，它们无法采用最优解释方法实现自身的证成。

最优解释方法与规范性理论中的传统证成方法在基本结构上存在巨大差异：它仿效知识理论的思路，将有待考察的规范性理论视为一个真理的候选者，而非一个规范性（价值）主张。如此一来，它自然可以采取考察科学理论的方法来检验一个规范性理论是否（最）合理，因为二者本质上归属于相同的类别（即真理），尽管在精度上有所差异。使得最优解释方法可以在科学领域和价值领域发挥相似作用的关键原因在于，在这两个领域中都要处理“理论接受”这一关键问题。在科学领域，接受一个理论即是接受其真理性，接受其对外在世界的描述最具有准确性；而在规范性的价值领域，最优解释方法的倡导者认为，接受一个理论也是接受了其真理性，尽管这是另一重意义上的真理性。对“真理性”的检验标准便是：理论能否最好地解释相关的现象 / 直觉？对于科学理论来说，完成了“最优解释”这一任务的理论便实现了对自身的证成，因为“解释”正是科学加

[1] 例如，一种剥除了直觉影响的康德式义务论便可以满足这一要求。康德的“绝对律令”有“普遍法则”、“人性目的”、“目的王国”等表述，这些表述中无论哪一个都很难说是对人们行动的具体指导。以“普遍法则”表述为例，它是“纯粹实践理性”的一个具体表达，从而是一种评判标准，而非直接的行动指南。故而，康德义务论有时也会给出似乎违背直觉的指示，例如不许对纳粹撒谎。依据本文所述类似的操作，康德似乎也可以使用最优解释方法来为自身理论辩护。当然，这涉及到对康德主义义务论做怎样的具体定性，此处不做展开。

[2] 罗尔斯的公平正义理论略有不同，因为它的主要作用对象并非人们的具体行为，而是“社会基本结构”。而且，用以规制这一结构的正义原则是在“无知之幕”之下的“原初状态”中由理性选择所确定。因而，公平正义究竟多多大程度上是纯粹的义务论正义理论，还需进一步讨论。

于自身的任务。然而，这一情形对规范性理论是否也适用呢？是否实现了“最优解释”的规范性理论便完成了对自身的证成呢？

五、真理与规范

为了回答上述问题，我们必须首先明确：何谓一个规范性理论？一个规范性理论与一个科学理论之间有何异同？从形式上看，“道德真理”可以与科学真理有着相似的结构，都可以用命题的形式来表达。科学真理描述世界是怎样的，而道德真理描述一种行为具有怎样的道德价值 / 属性，二者都可以通过联系动词“是”将被谓述的主词与对其的谓述连接起来。一个具体的理论是否正确，就看这一连接是否可靠，或是否能够得到辩护。就范例性科学来说，验证一个理论是否正确主要看其能否经过重复实验的验证。尽管从根本上讲，科学理论并不能被彻底证成（如符合论所期待那样），充其量只是尚未被证伪而已，在原则上存在被证伪的可能性。但就一般的科学实践来说，可重复实验所提供的验证已经非常充分。那么，这种对科学理论的验证 / 证成方式是否适用于规范性理论呢？我们是否可以通过类似于实验的方法来确定一个道德理论是否正确，即是否揭示了道德真理呢？

此时，我们有必要对“道德理论”做更精细的剖析，此处引入罗尔斯对道德理论的定性颇有参考意义。在罗尔斯看来，道德理论主要包含两个部分，即各种道德观念之结构以及它们与人类感性之间的关系，一个道德理论便是关于“道德结构及它们的道德心理基础的说明”[1]。在道德理论家的角色上，我们考察的是“人类心理的一个侧面，即我们道德感性的结构”；这一角色需要与我们作为一个特定道德观念的持有者之角色区分开来[2]。按照这种理解，道德理论的主要任务是系统地提炼人们所持有的道德观念，并以一个内部容贯的“原则体系”（scheme of principles）将其表达

[1] Rawls, John. 1974. “The Independence of Moral Theory.” *Proceedings and Addresses of the American Philosophical Association* 48: 5—7.
[2] Rawls “The Independence of Moral Theory” p.7.

出来。换言之，道德理论的任务是总结人们的道德观念，并对这些观念进行比较。在一重意义上，这是一个经验性工作，即收集和整理人们所持有的道德观念，或从道德哲学的历史中提取主要的“原则体系”并与人们的判断和直觉相比较，看看二者之间是否吻合。从这个角度来看，道德理论的确可以与科学理论相比较，因为二者的正确性都奠基在理论与现实的符合之上（尽管严格的符合论无法实现）。

但问题在于，这种刻画适用于最优解释方法所支撑的道德理论吗？首先，罗尔斯在对上述道德理论的描述中事先排除了“道德真理”这一说法；“道德真理”只在某个具体的道德主张 / 原则内部才可能存在，而道德理论关注的却并非这一领域。然而最优解释方法却明确预设了“道德真理”的存在。其次，就其性质来看，其所主张的基于人道与公平的分配正义理论更像是罗尔斯所说的一个具体的“原则体系”，代表的是一种实质性的规范性 / 价值主张，而非对这些“原则体系”进行高阶研究的道德理论。至此，我们可以确定，最优解释方法所理解的道德理论事实上相当于罗尔斯术语中的“道德观念 / 结构”。此时，我们的问题转为：一个道德观念 / 结构是否与科学理论具有结构上的相似性以致可以用最优解释方法予以证成？

尽管在经验层面上，我们可以考察一个具体的分配正义原则是否能够最好地捕捉到或解释人们的道德直觉 / 判断，但这仅仅提供了它作为一个经验性理论是否准确，却不能为这一原则本身是否正当提供证成。这里关键的区分是“准确”和“正当”，二者所蕴含的证成要求是不同的。如上文所述，“准确”是一种经验性标准，可由重复试验验证来满足，但“正当”作为一种规范性标准却并非如此。在一种极端情形下，例如在强盗群中，即使所有人的直觉都与某分配正义原则相符（即这个原则准确地捕捉到了所有人的道德感性），但这一符合也不能说明该原则是正当的。“正当”需要得到其他形式的证成，例如促进全体福利最大化，或符合纯粹实践理性的要求。如果这一思路是正确的话，那么葛著所建构的分配正义理论的证成基础便不是最优解释方法，而是后果主义自身的理据。然而后果主义自身的理据为何是成立的，本身也需要证成，这便回到了义务论与后果主

义的经典争论。但无论如何，最优解释方法对解决这一争论并无贡献。

从另一个角度来看，最优解释方法所依赖的科学真理与道德真理之间的类比也并不成立。尽管两种类型的真理都可以用命题的形式表述出来，但二者所使用的主词和谓述词却截然不同。科学命题的主词是事物或事件，其谓述词是有关特征或机制的描述。“道德真理”命题的主词却是人的行动，而谓述词则是价值或道德性质的判断。科学真理总体上是与行动（者）无关的，并无行动指导功能；而“道德真理”却总是事关某些行动者（或人类行动者这一总体），其主要功能正是引导人们的行动。由此可见，所谓“道德真理”并不是真正意义上的进行客观陈述的命题，其所表达的是某种价值判断，其主要功能是传达某种价值主张，可以无损地转化为祈使句的形式，也即失去命题的形式。如果这一定性是准确的话，那么适用于陈述式命题之证成的最优解释方法便无法为本质上是祈使句的“道德真理”提供证成。

六、规范性权威

用最优解释方法来为正义理论提供证成还会在另一个方面会遭遇严重困难，即它无法为正义理论建立“规范性权威（normative authority）”。事实上，这一问题被最优解释方法的倡导者完全忽略了。值得忧虑的是，即使这个问题的重要性被注意到了，最优解释方法也不具备充分的资源来对其提供回答。这里的核心关切是，一个完备的规范性理论 / 观念不仅要将其规范性主张表述出来，它还要具备对其目标受众的规范性权威。一个规范性理论的表述无论多么完备、清晰，它若不能建立起这一权威，终究也不过是一厢情愿的主张而已。为了获得真正的规范性权威，一个规范性理论 / 观念必须（至少能够）被其预期的对象所接受；更确切地说，它需要得到他们的授权。

虽然在一种根本的层次上，科学理论也需要树立其权威，即要求目标受众相信该理论，或接受其真理性；但这种权威与一个规范性理论所应具有的规范性权威之间存在显著的不同。面对一个得到充分验证的科学理

论，一个人若是拒不接受并继续按照非科学的方式生活和行动，我们或许可以称其为不理性的[1]（unreasonable），但对此却不能做什么，尤其不能对其进行惩戒或强制其按照科学理论所指引的方式生活和行动。但对于一个得到充分证成的规范性理论（尤其是得到法律强制力保障的政治规范），一个人若是拒不接受亦不服从，他便是不合情理（unreasonable）了，我们不仅可以在道德层面对其进行谴责，亦可以借助国家力量强制其服从，否则便予以惩戒。科学理论和规范性理论要求不同类型的权威，前者是知识性 / 理论性权威，后者是规范性 / 实践性权威。简而言之，规范性理论的目标在于实际地规范、指导人们的行动和实践，而非仅仅向他们介绍某个有关世界的描述性命题。在这个意义上，规范性理论潜在地与人们的（实践性）自由形成紧张关系，如若不能得到充分的证成（授权），便会对后者造成无理的干涉。

当一个规范性要求被向某人提出时，便是向她提出了一个权威主张（authority claim），即要求建立一种凌驾于（over）后者之上的发号施令 – 服从命令的不平等关系。显然，后者不会自动地接受这一权威主张；事实上，后者对此主张可以有各种不同的反应性态度（reactive attitudes）[2]。如果这一要求不能赢得她的自愿接受，那么即使她（例如，出于现实的审慎考虑）确实遵从了，它也没有对她建立起真正的权威。唯有在她真正接受（或认可）了该要求之后，这种权威才真正建立起来。然而，她为何会接受这种从外部对她提出的要求呢？如萨特所指出的那样，除了一个人自己，没有任何他人可以真正命令她做任何事，说不的选项一直向她敞开。一个人之所以会听从他人的指令，根本上是因为她至少在最低的意义上接受了该指令，即授权该指令来指导自身的行动。现在的问题是，是何种理由促使她给予外部指令这一授权？单纯的审慎考虑（比如保全自身生命及

---

［1］此处尽管使用了 unreasonable 一词，与后文中所用者相同，但二者意涵（尤其是外延）却有显著差别，此处仅有认识意涵，而后文中则在认知意涵之外增加了道德意涵。故而，本文使用略微不同的中文翻译来显示这种差别。

［2］这种反应态度理论最初由斯特劳森所发展，见 Strawson, Peter F. “Freedom and Resentment.” In *Freedom and Resentment and Other Essays*, 1—28, London: Routledge, 2008.

根本利益）当然可以诱导这种授权，比如在枪口胁迫下一个人会自愿地将钱包交给劫匪；但这种授权显然不是也不应该是规范性理论所寻求的授权类型。规范性理论所寻求的授权应当是其目标受众在不受胁迫和压力的条件下、在充分信息的基础上进行慎思之后依然可以自愿给出的授权。在这样的条件下，目标受众所要考虑的便是其根本利益、深刻信念和根本承诺等是否可以与外来的规范性要求相容或得到后者的支持。若这一规范性要求深刻挑战了他们的根本利益和/或信念与承诺，他们是难以接受的。放弃自身的深刻信念（信仰）而将自身的行动交由外来的规范性要求来主导，对他们来说不啻放弃自身的生命；对有深刻宗教信仰的人来说尤其如此。所以，一个规范性理论/观念必须处理其目标受众的这些深刻关切。

如此一来，一个规范性理论/观念的可接受性便不仅取决于自身主张的性质，例如是否明确、一贯、公开等，它在很大程度上还取决于其目标受众的一些特征，例如取决于他们深思熟虑的道德、哲学和宗教等信念与承诺，换言之，取决于他们的道德直觉。一个根本性的困难是这些信念与承诺、直觉等可以是多元的，彼此不相容甚至冲突。上文已经提及，正是因为这一多元性的存在，许多正义理论难以得到人们的普遍接受，甚至会遭到一些人的坚决反对，如果这些正义理论所立基的道德直觉恰好与其目标受众所实际持有的道德直觉正面冲突的话。正如罗尔斯所指出，这种价值领域的多元主义是任何一个非压迫性社会的自然现象，应当得到承认和尊重。对于规范性理论的证成来说，价值多元主义所提出的挑战在于：如何让同一个规范性理论/观念可以同时得到持有不同价值立场（直觉）的人们的普遍接受，并由此真正建立起对他们的规范性权威？

诚然，人们的有些道德直觉未必根深蒂固，也未必合理（reasonable），这些直觉可以通过个人慎思、社会教育等方式加以凝练和纯化，并达到合理与深思熟虑的程度；或者，从理论建构者的角度，可以不必将人们实际持有的各种直觉都纳入考虑，而是以一定的理想化标准（例如设定合理性这一门槛）进行筛选，只处理满足这一标准的几种代表性道德直觉。但不论采取何种做法，基础道德直觉的多元性无法彻底消除。因此，任何一种完备的规范性理论/观念都必须赢得多种价值立场（直觉）的支持，至少

应当不被它们（或其中的一些）坚决反对。无法获得这种普遍支持的规范性理论 / 观念便有降格为宗派性（sectarian）主张的风险。面对人们道德直觉多元化这一事实，试图从某种特定的道德直觉出发并建立起具有普遍规范性权威的规范性理论，注定是徒劳的。更有前景的做法是与这些多元的道德直觉保持适当的距离，即不将规范性的合理性直接建立在任何一种具体的道德直觉之上，但同时仍可获得它们的普遍支持，从而可以在目标受众面前建立普遍的规范性权威。 最优解释方法与罗尔斯所使用的“重叠共识（overlapping consensus）”方法都是这一思路的具体展现。本文接下来考察，最优解释方法是否可以实现这一目的。

最优解释方法试图为后果主义正义理论提供证成，其理据在于这一理论尽管并不与人们的多元直觉相符，但可以对它们提供充分解释。正是这一解释（相较于其他理论）的充分性使得后果主义正义理论成为人们（最）应当接受的规范性观念。单纯从解释力来说，情形或许确实如此，但关键的问题是，解释是否可以等同于证成？尤其是，一个规范性理论 / 观念对诸道德直觉的解释力是否可以建立其面对这些直觉的规范性权威？例如，后果主义正义理论承认人们要求按贡献分配的“公平”直觉具有初定合理性，进而援引后果主义的思路说明与其密切相关的几种权利为何需要被设立；但与此同时，它并不接受将这种（似义务论）直觉所附带的绝对性，并以“人道”直觉所产生的道德要求与之对冲。同样，对“人道”直觉（尤其是其所附带的似义务论的绝对性）它也并非尽然接受，尽管对这一直觉也可提供充分的解释。这样一来，它便兼顾了主要的道德直觉，而且对它们都提供了充分的解释。按照最优解释方法的标准，后果主义正义理论便完成了自身的证成。然而，正如前文所述，一个规范性理论 / 观念的证成中最为关键的一步是其规范性权威的真正建立，也即其赢得目标受众的授权。现在的问题是，一个并不承认“公平”直觉（附带之绝对性）的正义理论究竟如何才能赢得深刻持有这一直觉的受众的认可？

仅仅向他们解释该直觉是如何形成的似乎并不足以让他们放弃这一直觉。而在该直觉（附带之绝对性）未被放弃之前，后果主义正义理论便与之形成根本冲突，难以取得持有该直觉的受众的授权。为了让他们放

弃（放松）这一直觉，似乎需要进一步加强自身的主张，即不仅解释这一（绝对性）直觉为何不成立，还要人们事实上接受这一解释。换句话说，后果主义正义理论需要将其所确认的“道德真理 / 真相”真正灌输到人们心中。只有这样，其所倡导的“公平 + 人道”的混合式正义理论才能得到人们的普遍认可，因为此时人们心中已经普遍接受“公平 + 人道”这一混合式的道德真理了。然而，此时的问题便转化为，为何可以要求人们接受这一“道德真理 / 真相”？由于最优解释方法借助的是科学真理的类比，我们可以推测它的倡导者认为人们理应接受真理，否则他们便是不理性（unreasonable）的，而对不理性的人也无需进行注定徒劳的论证。表面上看，这一论证似乎行得通。但这里存在一个关键的混淆，即面对理性论证无动于衷的人也许在知识的意义上是不理性的（unreasonable），但这并不意味着他们在道德上也是不合理的（unreasonable）。事实上，拒绝接受达尔文进化论的宗教信徒，完全可以在道德上是典范性的。于是，以知识范式来处理道德上的不同意见者，理由并不充分。

七、结语

最优解释方法的倡导者以现实人性为基础构建并证成了一种基于后果主义的正义理论，试图以这种方法超越各种具体道德直觉之间的冲突和不一致，在方法论层面引入了范式创新。然而，由于这一新的范式是建立在对科学真理及其证成的类比之上，未能充分关注价值主张与事实性命题之间的结构性差异，也未能充分关注规范性理论需要对其目标受众建立起规范性权威这一要求，故而未能真正实现这一方法的预期证成效力。因此，最优解释方法需要在这些方面做出重大补充，才能为其颇具创新的正义理论提供充分的证成。

From Explanation to Justification

Can inference to the best explanation sufficiently justify a theory of justice?

(Ye Jinzhou, Huazhong University of Science and Technology)

Abstract: Foundationalism and coherentism as represented by "reflective equilibrium" are the major approaches to justify normative theories of justice, but both face deep difficulties. "Inference to best explanation" is proposed as a third approach, aiming to better justify theories of justice. The efficacy of this method relies on not strictly distinguishing factual truths from normative theories (claims) and it presupposes the existence of universal moral truths, but both are groundless. Furthermore, the issue of establishing normative authority of normative theories is totally absent in this method, leaving a significant defect for remedy.

Keywords: inference to the best explanation, justification, normative authority