前沿

共识、道德和策略

格蕾塔困境和监禁博弈

考希克·巴苏

1. 导论

我们比了解无私更了解自私。这事出有因。亚当·斯密（1776）在他的鸿篇巨制中指出，很多时候，覆盖整个经济与社会的秩序，都是可以从社会成员自利的角度加以解释的。此议一出，世人皆惊，舆论哗然。某些理论家利用斯密假说为他们的一己之利举行合理化的庆典，鼓吹无须为社会秩序与正义付出任何努力——亚当·斯密之“看不见的手”自会代人行事。更有大批道德哲学家、政治经济学家甚至数学家将斯密假说视为思维挑战，竞相形式化地描述和分析这一假说在多大程度上可以成立，并可以作为制定政策的基础。对当代经济学家来说，彼时学界对斯密假说的反响可谓遗惠诚深。所遗惠者，并非对自利的颂扬，而是势如泉涌的相关理论探究。从瓦尔拉斯（Walras，1874）到阿罗和德布鲁（Arrow and Debreu，1954），通过现代一般均衡理论的方法，对这些古老的哲学思想加以分析，使我们在这一领域获得了更加坚实的基础。

以上研究开启了一个探索过程，即个人行为是如何超越自利而通向正义、公平与道德的。这样说的理由是，虽然这方面的有些想法可追溯至托马斯·霍

---

* Kaushik Basu，康奈尔大学经济学教授和 C. Marks 国际研究讲席教授。原文“Conventions，Morals and Strategy：Greta's Dilemma and the Incarceration Game”发表于 *Synthesis*（2022）。

布斯与大卫·休谟，甚至还可以在柏拉图与亚里士多德的著作中找到踪迹，但直到近期，我们才获得了分析这一问题的形式化工具。我指的是博弈论，如由冯·诺依曼和摩根斯特恩（von Neumann and Morgenstern，1944）、纳什（Nash，1951）阐述的理论。对一个成功社会的基本社会构件的博弈论解释，还要到更晚才会完成。自20世纪中期的一些学术贡献开始（如Braithwaite，1955；Runciman and Sen，1965），对那些古代议题的重新研究才缓慢积累起来，使人们有望看到在一个形式化的框架上最终建立起既可进行实验室检验，又可进行田野实验的正义与秩序新理论。

这就是彼得·范德施拉夫（Peter Vanderschraaf）的新著《策略正义：共识与利益分歧的平衡问题》（*Strategic Justice*：*Convention and Problems of Balancing Divergent Interests*，2019）的重要性之所在。它是这一领域的集大成之作，用现代博弈论的工具对那些早期思想建立了形式化的分析框架。那些早期思想固然重要，但因轮廓不清，批评不易，付诸实践也难。范德施拉夫著作的重要性在于它是对这一领域探索的继往开来，而非它的终场谢幕。本文也是本着同样的精神参与讨论。只不过我的目的，并不是简略浏览范德施拉夫著作涉及的全部领域，而是提请读者以博弈论的观念、成果以及相关的经济学启示，聚焦某些特别的主题。我在下文中也无意掩饰对休谟理论的赞同，以及对霍布斯的更具批判性的评论。

在展开这一评论的过程中，本文将展示两个新的研究成果：其一是关于群体道德责任的格蕾塔困境，其二是监禁博弈，对领导人赋权过程中知识的作用和知识层级（knowledge hierarchies）提出一些看似违背常理却理应受到重视的问题。这两个成果都与范德施拉夫著作中提出的议题相关。格蕾塔困境的主题是道德激励，也就是在策略环境下我们采取的很多行动背后的驱动因素；而监禁博弈则表明就限制领导人的权力而言，事前达成共识（prior conventions）与制度的重要性。虽然本文深入讨论了规范问题，并开出了一些“药方”，但它最终仍是一篇本体论意义上的学术讨论，旨在证明范德施拉夫分析的那些问题绝非已获得完美解答；不仅如此，新的问题，新的疑难又已出现。对不同社会将在实践上如何解决这些问题，我们需要开出“药方”。本文认为，解决问题的“药方”必须立足于共识及宪法的设计。但本文的规范性讨论，主要目的还是要激发更多的分析，而不是为讨论画上句号。

读者不应感到惊讶，本文分析引出的一个结论是，归根结底，世界上并没

有实现正义与构建公正社会的灵丹妙药；不是人类尚未找到这种灵丹妙药，而是它根本就不存在。尽管如此，相关的研究和探索却非常重要。它加深了我们对问题的理解，当这些问题发生的时候，也使我们能够更好地加以解决。

2. 协商道德与道德意图

社会可实现的均衡结果不计其数，共识能够帮助我们无须借助超越个人利益的外力，而实现一个较好的结果。这一点已被经典的囚徒困境以及农夫困境重复博弈（Vanderschraaf，2019）所证明。但这种结果，只有在这些博弈并无人所共知的终止日期时才会实现。

在这些博弈中，通常有多个均衡，博弈者很容易产生不一致的预期，导致灾难性结果。范德施拉夫引用谢林（Schelling，1960）和刘易斯（Lewis，1969）的研究，正确地指出，在此类情况下，共识发挥着至关重要的作用。这些共识会促使博弈者围绕博弈焦点，即博弈者想象的结果，形成一致预期。在此意义上，正义就是共识；共识会帮助此博弈者就彼博弈者的信念与行为建立信念，并预判彼博弈者对此博弈者当下行为即将做出的回应。①这些信念能够帮助社会维系一种最优结果。

但有两点必须注意：第一，此处的“最优”是指帕累托最优，也就是说，这个结果不可能使一个人受益的同时而其他人不受损。众所周知，帕累托最优结果也有可能是令人难以忍受的。当一个压迫型社会对一个群体的残酷剥削达到了无以复加的程度时，这个社会也可能是帕累托最优的。为了防止出现这种压迫性的“最优”结果，我们就需要做出超越自利的选择。

第二，即使通过重复博弈中自我实施的共识有可能实现真正好的结果，但是因为众所周知的逆向归纳论，只要博弈有一个结束日期，如在囚徒困境或农夫困境博弈中一位或两位博弈者采取最后的行动，这一解决方案就会失灵。

鉴于以上情况，我们不可避免地再一次认识到，人们的自利行为确实需要受到约束（Vanderschraaf，2021）。而约束的实现，或是通过社会规范在人脑中形成程序化反应，或是通过有意识地践行某些公平或正义原则。后者本质上

① 可沿用这一逻辑为法治定义。归根结底，法律也只不过是一个焦点，一种在公民中建设一组一致信念的机制（Basu，2018）。当然，其间，惩罚也自有功效。我遵守限速法规是因为害怕受罚，交通警察处罚我超速也是出于畏惧上级对他的不作为施加处罚。最终，之所以众人尊法，社会守法，也是因为有了这样一个相互交织的信念系统。

是一种“道德意图”（moral intention），可能但未必基于共情，它也是人的一种内在特质（诚然因个人与社会的不同而又表现各异）。

现在我要做的是，证明即使这样的解决方案包含了道德考量，也有把社会导向真正好的结果的意愿，却还是有可能以耐人寻味和常人难料的方式失灵。这就为我们留下了一个悬而未决的哲学问题。何为德行？孰为义举？围绕这一主题的思辨浩如烟海。我对“有道德”的定义是一种结果主义的定义，它要求我们必须重视他人的福祉。这里使用的形式化方法就是范德施拉夫在谈及行为人“从自己的利益和伙伴的利益中获得某种程度的满足”时建议的方法。②

另有一个重要提醒就是下文描述的悖论，它被称为“格蕾塔困境”，这一悖论只有在范德施拉夫关心的策略性博弈环境下才会产生，而不会在传统的理性选择背景下产生。这也是我们在传统背景下建立的直觉一时难以接受格蕾塔困境这一结果的原因。所以，我在开始分析新的博弈之前，先简要说明传统背景下什么是“有道德”。

假设某人，姑且称他为 P1 君，必须从一组可供选择的行动 X 中选择一项行动。每一项行动 $x \in X$ 都以一个数对（x_1，x_2）表示，其中 x_1 代表 P1 君，即选择者所得的收益，而 x_2 则是局外人（bystanders）的收益。局外人有可能是子孙后代，因而对 P1 君的选择无法施加影响，而 P1 君的选择却必然影响局外人。假设 P1 君完全自利，让我们把他的选择记为 $x^* \in X$。接下来，同理推衍，对所有 $x \in X$，都有 $x_1{}^* \geqslant x_1$。

假设 P1 君现在变得有道德了，表现在他视 P2 君的每一单位收益相当于 δ 单位他自己的收益，其中 $1 \geqslant \delta > 0$。让 $z^* \in X$ 表示 P1 君变得有道德后的选。显然，在这种非策略环境下，对于局外人来说，博弈者变得有道德从来就不是一件坏事。也就是说，$z_2^* \geqslant x_2^*$。

为了说明这一点，假设上述不等式并不成立。也就是说，

$$x_2^* > z_2^* \tag{1}$$

自利时的 P1 君选择 x^* 这一事实，意味着 $x_1^* \geqslant z_1^*$。但是这一选择，一旦

---

② 在本文中，每当论及道德，我知道在行动的道德来源这一问题上自难避免争议（Markovits，2010）。如果一个人出于自我愉悦而帮助一个穷人，有人会质疑此举并不道德。但这种争论与本文的关注点无关。我对道德行为的定义是博弈者不但寻求自我收益最大化而且也顾及穷人和边缘人士的收益。共情这个词，我的确常用，但只不过是为了行文方便。读者对此等行为尽可给出自己的称谓，但都不会改变我的结论。

与式（1）结合，则意味着：

$$x_1^* + \delta x_2^* > z_1^* + \delta z_2^* \tag{2}$$

但这是不可能的，因为当 P1 是有道德的，他选择的是 z^*。这一矛盾证明，选择者有道德对局外人或子孙后代来说通常是合意的（或者更确切地说，是更合意或同样合意的）。正是出于这一理由，所有心地善良和有环境意识的人，比如格蕾塔·滕伯格③，都一致劝诫我们对子孙后代能够将心比心。

在将此作为背景的同时，还应注意到，关于群体道德责任以及不同群体道德义务的讨论已是一个广阔的学术领域（Feinberg，1968；Sartorio，2004；List and Pettit，2011；Schwenkenbecher，2019）。我个人在其他地方也有关于这方面的论述（Basu，2021a），因而有关这一领域的各种争议在这里就不再赘述。

我在这里的讨论与另一类文献有关，此类文献认为，大多数社会在道德上是多元的。事实上，道德多元化是现代社会的一个定义性特征（Moehler，2018；Thrasher and Vallier，2015）。④这种观点与主流经济学形成鲜明对比，在主流经济学里，不同的个人被视为近乎同质的，基本上每个人都不受道德的羁绊。博弈者尽管有不同的偏好，但都不会顾及他人。我（2021a）探讨过一个完全道德的社会与一个完全无道德社会的案例。但如果同一个社会里也有差异，有些人恪守道德，有些人不顾道德，又会发生什么呢？格蕾塔困境博弈显示了在某些人变成道德生物后社会变得道德多元化所带来的悖论。

如何解决上述争论，对我们制定法律、选择行为，都有重要影响。在自利的问题上，得益于博弈论，我们可以非常清晰地说明个人选择对群体行为的影响。当看到村民在公共牧场上过度放牧，最终损害其个人福祉时，我们也不会痛惜他们不按自利行事。我们认识到，问题的根源恰恰就是自利行为。囚徒困境、游客困境以及休谟（1740）的农夫困境都说明了这个问题，也都能帮助我们为现实世界设计政策以解决公共牧场问题，以及缔结国家间协议以应对全球气候变化。我们今天已经对个人理性为集体福祉埋下的诸多陷阱有所警觉，

③ Greta Thunberg，瑞典青年活动人士、政治活动家和激进环保分子。

④ 这里需要指出的是，我曾（2021a）得出的结论，即每个人都变得有道德却导致群体道德变差的结果，确实依赖于道德多元化。如果一个社会的道德矛盾根深蒂固，或者它是一个根深蒂固的道德多元化社会，上述结论就不足为奇（Moehler，2018）。但在我的论文里，亦如本文，人们的道德观却更具共性。公民中没有“基本分歧”（Quong，2005）。道德的含义是关注穷人或者子孙后代这样的局外人，他们不能直接影响我们，却深受我们行为的影响。个人固然有其私利，但共有的“德道关怀”可以弱化个人的自利倾向。

但对道德行为并非如此。

我在已发表的著作中（Basu，2021a）曾经设计了一场博弈，叫作“撒玛利亚人诅咒”，用以说明整个社会的道德化也能使社会行为趋向反道德化。在某种意义上，这场博弈就是囚徒困境在道德领域的反面。然而，在“撒玛利亚人诅咒”的博弈里，个人的道德醒悟无法扭转群体道德的指向。在本文中，格蕾塔困境博弈则说明，哪怕在社会中的某个人身上注入一点点的道德，也能引发看似矛盾的逆转。这就耐人寻味地扭转了前人研究过的一个主题。萨托里奥（Sartorio，2004）讨论过，一个人如何有可能对并非由其自身行为导致的某个结果负责。而我在这里要说明的是，在某些情况下，并非因某人的行为，而主要是因其身份，就能造成群体行为的反道德化。而且更糟糕的是，某一个人的“道德化”［Vanderschraaf（2019）意义上的道德］甚至还会导致整个群体的反道德行为。

让我们设想两个人之间的互动关系，为方便起见，姑且称他们为甲和乙。他们是“博弈者”，在可供选择的行动集中选择，博弈者甲在行动 A 和行动 B 中选择，而博弈者乙则在行动 C 和行动 D 中选择（见表 1）。所以，博弈的“结果”包括（A，C），（A，D），（B，C）和（B，D），在以上每一对结果中，第一个符号代表甲的选择，第二个符号代表乙的选择。博弈者得到的收益，列出在表 1 左边标为“基本博弈”的收益矩阵中，博弈者甲从行中选择，博弈者乙从列中选择。每一格中的收益分别是博弈者甲和博弈者乙的收益。如此这般，如果博弈者甲和博弈者乙分别选择 A 与 C，他们即会分别得到 100 美元与 101 美元的收益。按照博弈论的说法，如果他们选择了（B，C），则收益为（101，100）。为简便起见，我们可以把他们的收益想象为以美元计算。

表 1　两人社会的博弈及其结果

	基本博弈			未来一代收益	
	C	D		C	D
A	100，101	100，100	A	2	8
B	101，100	101，101	B	0	4

现在让我们为这些符号赋予具体的含义，把 A 视为环境友好型的“务农”活动，把 B 视为破坏环境的“烧砖”活动，把 C 视为会对环境造成非常大破

坏的“采煤”活动，把D视为更绿色的“养奶牛”活动。对于局外人，或本案例中的子孙后代来说，这些选择的意义显而易见。现在我们暂时假设，B是占优策略，也就是说无论博弈者乙怎么选择，从博弈者甲的自利角度看，“烧砖”肯定是比“务农”更好的生意。然而博弈者乙的理想选择，却取决于博弈者甲的行为。如果博弈者甲选择“务农”，则假设他将主要从事作为乳品替代的大豆生产。所以对博弈者乙来说，从事“采煤”将会获得更高的收益。但如果博弈者甲选择的是“烧砖”，人们对乳品的需求就会高涨，博弈者乙就会选择从事奶牛饲养业。

鉴于人们都有个人收入最大化的意愿，“纳什均衡”（以下简称“均衡”）就是博弈者中任何一方都无意于单方面偏移的一对行动组合。显而易见，在这场博弈中，唯一的均衡就是使两个博弈者得到（101，101）的（B，D）选项。如果博弈者甲偏移到A，他只会得到100美元，而如果博弈者乙偏移到C，他也只会得到100美元。所以两人都不愿意偏移。

现在假定这个社会还有可怜的“未来一代”，承受着当前一代破坏环境的恶果。未来一代的福祉完全取决于当前一代的所作所为。如果博弈者甲和博弈者乙主动选择环境友好型行动（A，D），未来一代就会获得8美元。如果他们选择具有最坏环境影响的行动（B，C），未来一代得到的就只有0美元。而如果他们的选择是兼具利弊的（B，D）或（A，C），未来一代的收益将是4美元或2美元。这样就概括了表1右边矩阵描述的未来一代的收益情况。在下文中，两个博弈者的收益（如左边矩阵所示）与未来一代或无助的局外人的收益（如右边矩阵所示）的集合被称为一个“社会”。

让我先从标准的博弈论和新古典主义经济学开始讨论，根据这些理论，两位博弈者在决定如何行动时绝不会考虑未来一代的收益，而只关心他们自己的收益，因此方能达到均衡（B，D），使两位博弈者各自的收益都达到101美元，而局外人只得到4美元。

在大多数观察者看来，这一组博弈者的行为当属不义。虽然当前一代改善了自己的福祉，其中每人的收益达到100美元甚至更多，况且他们每人如何选择，顶多只会对自己的收益造成1美元的差别。然而未来一代却有可能被置于安危叵测的境地，所得收益也大幅降低。请注意如果博弈者放弃（B，D）选项而转向（A，D）选项，未来一代或局外人就将获得8美元而不是4美元的收益。连如此微小的牺牲都拒绝做出，的确显得不够道义。

现在再来假设，博弈者甲有幸遇到了格蕾塔·滕伯格，学习并领会了我们在上文中讨论的道义，具有对无法参与我们当前辩论的未来一代理应有的关怀。博弈者甲学到了做事要有道德底线，要小心自己的行为对作为局外人的未来一代造成的影响。正如我们所有人都应做到的那样，他将格蕾塔的规劝牢记于心，变成了会将子孙后代的福祉纳入个人决策过程的人。为了让我们的讨论更加简明，让我们假设博弈者甲正是以范德施拉夫在其著作中论及的共情之心，对个人福祉和未来世代的福祉都给予同等的重视。

这就使博弈发生了改变，形成了一位自利的博弈者（博弈者乙）与另一位有道德的博弈者（博弈者甲）之间的对弈。博弈的变局如表 2 所示。我称它为“格蕾塔困境”博弈。在这一新的博弈中，未来一代的收益被加到了博弈者甲的收益上，因为博弈者甲现在也要为未来一代的收益最大化而奋斗。

表 2　格蕾塔困境博弈

	C	D
A	102，101	108，100
B	101，100	105，101

那么格蕾塔困境的结果是什么样的呢？显而易见，只存在（A，C）这一种均衡的可能性。过去的结果（B，D）不再是均衡结果，这是因为，如果博弈者乙选择 D，博弈者甲将出于对未来一代的考虑而选择 A。而一旦知道了博弈者甲的选择是 A，博弈者乙就会转而选择 C。除此之外没有其他的策略组合是稳定的。因为博弈者甲从不顾道德的行为人转变成有道德的行为人，所以博弈的最终结果必然是（A，C）。⑤

那么，这会对未来一代产生什么样的影响呢？在博弈者甲见到格蕾塔之前，未来一代的收益是 4 美元。当他在格蕾塔的影响下，变成了一个具有环保意识的人，下决心要帮助未来一代之后，未来一代的福祉反而降到了令人吃惊

⑤　虽然这一博弈寓意深刻，但对我们制定实际政策有何意义仍是一个值得讨论的问题（Basu，2021c）。我们也许应该对格蕾塔困境做一些有趣的实验检验，以确定预料中的道德退步在现实中是否有可能出现。实验检验可以让博弈者参与表 1 所示的基本博弈，但并不让他们知道博弈对局外人的影响。然后，让博弈者甲知道局外人的存在，并向他提示道德责任，再由甲乙二者重新博弈，看博弈者甲是否弃 B 选 A。有学者对操纵信念与焦点做过实验研究（Dasgupta and Radoniqui，2021），这一研究对如何在博弈中引入道德与价值观提出了建议。

的 2 美元。

当外界的观察者看到博弈中的这一组富人竟然以自己的行动使未来一代的收益下降到了 2 美元，会觉得难以理解，因为两人中的一人已经实现了道德升华，下决心要帮助未来一代。不难看出，在策略性环境下，某一博弈者的道德升华并不必然引致一个更符合道德的结果。

为了理解格蕾塔困境，让我们把任何少于 3 美元的收益都称为“坏结果”。因此（A，C）和（B，C）都是坏的。当坏结果发生时，显然博弈者甲并不负有责任，因为无论他如何选择，坏结果都会产生。而博弈者乙则负有道德责任。如果博弈者乙选择 D，就一定会使坏结果得到扭转。⑥现在的问题是，如何对这样的道德责任与博弈者甲见到格蕾塔后经历道德升华反倒造成坏结果的情况做出一致解释。

解决上述问题的一个可能方法就是论证博弈者甲肯定知道他的道德升华会造成有害影响，因此他最好表现出没有道德顾虑且自利。这样一来，博弈就会以（B，D）为结果，而局外人将获得更多的收益。但这就相当于论证，与囚徒困境相反，博弈者有意识地采取不自利的行为并带来一个好的结果。但这一论证会遭到博弈论专家的反对，其理由是，在你让另一个博弈者相信你不是他们想象的那种人以后再偏离原有的行动总是有利可图的。

在格蕾塔困境中，一个类似的建议就是由博弈者甲假装没有道德顾虑，使博弈趋向于（B，D）的结果，然后，博弈者甲在最后一刻转向 A，使博弈产生（A，D）的结果，使局外人获得 8 美元的收益。对此，博弈论专家会反驳（而且哲学家肯定也会同意）：博弈者乙也会预料到这一情况发生，因而选择 C，这样使结果又回到了（A，C）。

但是，在自利的博弈者甲与有道德的博弈者甲之间，却存在着一种传统博弈论无法体现的差别。⑦这种差别在更大程度上属于哲学范畴。在博弈论里，我们把自利（即追求个人收益最大化的冲动）当作人与生俱来的天性。但道德升华是否必然涉及对意志力的考验，却大可争论。因此，我们也可以认为，

---

⑥ 在博弈论和道德环境下的类似讨论，参见 Braham and van Hees（2012）。

⑦ 应当说明的是，虽然我讨论上述博弈是以环境损害和未来一代的福祉为背景的，但它可以拓展到任何外部性以及我们面对外部性时的道德立场。它可以适用于军队士兵，起初他们对敌方的生命损失无动于衷，而后认识到暴力并无正义可言。它亦可适用于以微小牺牲帮助穷人的行为。抽象分析的优点在于它是可应用于诸多情形的一种工具。

个人在充分审视环境后，也就是在本文讨论的博弈中，可以选择成为（或者不成为）有道德的人。如果他们有足够的远见，就可能决定在某些情况下避免做出有道德的行为。⑧

格蕾塔困境甚至可以带领我们超越道德哲学，进而触及认知科学里的联结主义（connectionism）。联结主义认识到，人类的认知由神经元的巨大网络组成，而其间每个神经元并不知道自己的角色。现在，有些学者力图将联结主义引入社会科学，指出或许也存在人类行为的大型网络，而置身其中的个人察觉不到这个网络的存在（Clark and Chalmers，1998；Fioretti and Policarpi，2020）。研究这种有机组织的整体效应非常关键。但这也意味着，在很多情况下，想要对组成集体的个人成员做道德评判是徒劳无益的，因为他们并不具有通常意义上的意志力。

我的论点是，目前的希望，只能存在于帕特南（Putnam，2005）提出的建议里。他从伊曼努尔·列维纳斯（Emmanuel Levinas）的著作里得到的启示是："对列维纳斯来说，道德的不可简约的基础，就是一旦我遇到一个蒙受苦难的人类同类，便立刻认识到我有义务为此人做些什么（即使实际上我无能为力）。如果根本没有这种帮助受难者的义务感，没有我若有能力就必定帮助受难者的认识……就是不义。"请注意，帕特南的建议与"理应意味着能够"（ought implies can）的格言并不矛盾。他并不是断言你应该帮助你无法帮助的某人，而是说你内心至少应该感到你有义务予以帮助。

在此基础上，可以做出这样的论断：即使我们尚无法达到一个好的结果，也必须培育和维护那种可被称作"道德意图"的东西，也就是最终要实现好结果的一种意图。如果这种道德意图微弱，那么当我们面对道德上的坏结果时，只能视其为既成事实而无所作为。在道德意图的驱使下，我们就会有意愿超越眼前的博弈，思考怎样改变自己的行为，比如通过采纳义务论意义上的道德，或者通过对博弈者征税和处罚来改变博弈规则，又或者在类似格蕾塔困境的策略性情形中，一个人即便本不自利也要做出自利的行为。

结论就是：想要避免格蕾塔困境绝非易事。当我们看到某个群体表现恶劣，千万不要猜想他们的行为后果就必然反映群体成员的意愿。多数人不愿意

---

⑧ 一个相关问题的讨论可以参见关于联盟如何建立的文献（Aumann and Myerson，1988；Genicot and Ray，2003；Ray and Vohra，2015）。但这其中的道德选择关乎观念上截然不同的问题。

承认，一群领导人可能不想做他们作为一个集体通常要做的事情。这些领导人可能已落入陷阱，正如哈维尔（Havel，1986）料想的，领导人可能根本没有退路。这的确是有着良好道德意图的格蕾塔必须认识到的一个困境。关怀子孙后代的人不但有可能无法对子孙后代有任何帮助，甚至事实上还有可能使他们的利益受损。策略性环境是博弈论中的核心问题，往往也是现实中的核心问题；不同的策略性环境会带给我们意想不到的挑战。

在结束讨论之前，让我提示两点，这两点太重要了，因此不宜作为脚注放在页底。第一，我讨论了需要有事前共识来保护我们避免跌入一个坏均衡的陷阱。关于博弈论与正义的很多文献关注博弈在某一点上的结果，最具代表性的就是纳什均衡，这是一个显著的结果，其显著性由共识实现，因为共识通常以某一特定的（很可能是合意的）结果为目标。

但是，还需要超越以上分析来考虑集值均衡（set-valued equilibria），也就是以“我同意绝不超出特定的行为集，你也承诺绝不超出特定的行为集”为形式的共识。对宪法的最好理解就是把它视为笛卡尔式的产物，包含着可供每人选择的策略组合。归根结底，这就是宪法与共识的现实功能。它严禁某些行动，但通常也并非把人锁死在某项具体行动中。⑨就本文的目的而言，无须阐述这一观点，但在学术文献中，这仍是一个重要的空白，值得将来继续研究。

第二，另有一点要说明的是，需要区分不道德与贴不道德标签。上文的分析使我们更加怀疑揪住某些人并让他们为集体无道德负责的做法。但是，根据帕特南（2005）的建议，相信人们负有道德责任与直言（甚至也许是感到）某些人负有道德责任，是有所不同的（Basu，2021a）。人们都不愿被指责为不道德。不道德的指责就像贴标签，是一种惩罚。

换言之，如果人们因为做出某项选择而被视为不道德，他们就有可能改变行为，使群体做出更好的选择。这就意味着，或许可以出于结果主义的理由，对做出某些行为的人贴上不道德的标签，即使我们并非真的要让他们为此承担

⑨ 与此比较接近的一个概念是“限制”（block），即每一个博弈者的非空纯策略子集。Myerson and Weibull（2015）认为，共识最好被视为自我实施的限制，也就是说，没有人想要单方面地突破这一限制。在谈到宪法时，这最多是第一步。一部宪法应当被视为不突破某些策略的规则，这些规则甚至在我们知道博弈将会是什么之前就已被制定。宪法观念以及对宪法的需要正是来自这种未来不确定性。但对此做形式化分析并不容易，因为甚至在我们知晓博弈的全部走势之前，就不得不设法制定行为规则。

道德责任。

3. 休谟式的领导观

如果社会恰似格蕾塔困境中发生的那样，倾向于滑入无政府状态或出现道德上令人遗憾的结果，那么有无破解之法呢？我们会习惯性地求助于事前共识或依靠一位领导人为社会掌舵来避免上述结果。在这方面，休谟的路线经常被视为独立于诸多早期思想，特别是霍布斯的分析（Vanderschraaf，2019，第 6 章）。调和休谟、霍布斯这两条思想路线的研究工作少之又少（Gauthier，1990；Sugden，1986）。为了调和休谟与霍布斯的观点，我在其他作品里采取了一条相当不同的路线，可以贴切地描述为休谟式解决方案。这一解决方案在某些方面与范德施拉夫（2019）的立场颇为接近。在讨论下一节中的新思想之前，我首先简要介绍我的这个解决方案。

人们经常认为，芸芸众生需要有自上而下的控制才能采取合意的行动。比如为了保证限速的法规得到遵从，就需要有交通警察来监督道路和惩治违规。一旦下一步我们询问为什么警察会像人们期待的那样完成他们的工作，我们又会得到一个同样性质的回答：那如果警察不好好惩治超速司机，就会有人惩治不尽责的警察。最后，如果社会有任何秩序可言，那么秩序必定产生于信念网络而非其他，这个网络联结着每个人对他人的预期。休谟的这一洞见，可谓精妙绝伦。以现代博弈论的语言来表达，一个稳定的社会结果只不过是人生博弈（game of life）的焦点。共识只不过是创造了这样一个焦点（Lewis，1969）；法律也不过是创造了这样一个焦点（Basu，2018）。

根据休谟的秩序观，初看起来，一个社会似乎并不需要领导人。但有一个耐人寻味的原因表明，并非如此。稍经思考，我们就可以明白，在有些情况下，博弈者事先并不知道自己会参与什么博弈。在战争或革命形势下，尤为如此。一支部队可能会突然遭遇每个战士必须决定在附近 10 座山峰中应该占领哪一座的情形。如果他们全都攀上同一座山峰，就会取胜；如若不然，就会被敌人消灭。不同的战士或许钟爱不同的山峰，但他们所有人都同意，在攀上同一座山峰还是不同的山峰这个问题上，攀上同一座山峰是有益的。

由于这样的博弈是突然出现的（正如在战争和冲突时期发生的那样），没有时间讨论并商定一个焦点或达成有限的共识（在我们的案例中，就是指全体战士应该攀上 10 座山峰中的哪一座）。针对这一困境的解决方案，就是事先

挑选一名焦点人物。换言之，把谢林关于焦点的观点拓展到焦点人物就可以解决这个问题。这里的关键是有一个人被指派为作战部队的指挥官。当博弈突然出现的时候，战士们或许会指着不同的山峰各抒己见，但每个人都知道其他人会听从作为焦点人物的指挥官的号令。⑩ 因此，军队有严明的指挥系统，也就不足为奇了。

这表明，尽管在某些情况下，我们的确需要有人领导，但领导人的权力最终植根于休谟所说的信念之网中。指挥官之所以有权力，是因为每个战士不仅知道其他战士会服从指挥官的领导，而且知道其他战士也知道其他战士会服从指挥官的领导，如此推衍，直至全体。领导人的权力并非一种外生的权力，而是支撑焦点的那些共同信念的产物。在这里，有意思的是，霍布斯（1651）在《利维坦》第十三章中不那么情愿地承认，领导人的权力并非来自他个人本身，“把所有因素集合起来考虑，人与人之间的差别并非那么大……因为即便是最弱者也有杀死最强者的体力。”

显然，一旦将这种领导观进行形式化分析，人们就会更深入地理解领导人的社会影响。我们看到，与霍布斯的理论恰恰相反，在某些重要的情景下，多个领导人相互之间的权力制衡是有利于人民的。我们还会发现，在另一些情形下，比起有利维坦式的领导，没有领导反倒能使公民生活得更好（Basu，2021b）。在这类情形中，万一有一个利维坦式的领导人，也会产生将其撤下权位的需求。

在这里，我要做的是更进一步说明有些领导人是如何制造压迫机制，阻止公民把他们逐离权位的。为此，我建立了一个监禁博弈的模型，以进行形式化分析。这场博弈的结果是稳定的，却是一种压迫性的稳定。那么，如何阻止这种压迫性的稳定结果发生呢？这个问题会再次把我们带回到范德施拉夫讨论的主题（2019）。那就是在领导人登上权位之前，需要达成共识，或订立一部宪法，能够做到远在事发之前，就消除领导人倒行逆施的可能性。

由此，亦可联系到受塔洛克（Tullock，1971）的“革命悖论”激发，由范德施拉夫（2008）参与的一场经典辩论。塔洛克讨论的主题是革命中的公共品问题；他说即便每个公民都想推翻“一个邪恶、腐败、压迫与麻木不仁

⑩ 基于 Myerson（2021）的观点，我认为需要拓展谢林关于焦点的看法，使之适用于某个焦点人物或焦点博弈者（Basu，2018）。领导人可以在行为集中创立焦点的想法就产生于此，我也给出了形式化的证明（Basu，2021b），不过基本观念可追溯至大卫·休谟。

的政府”（Tullock，1971，第 89 页），他们可能也无法实现愿望，因为每一个人是否参与抗议对结果的影响微乎其微，而必须面对的风险却太大。这样一来就没有人参与抗议，而需要每个人参与才会发生的革命，也就不会发生。范德施拉夫创新性地结合了博弈论与引爆点理论，推翻了塔洛克的论点。

在接下来的一节里，我将沿着一条新的路线进一步分析这场辩论的主题。我的做法是凸显对立面，即展现领导人如何利用博弈者的知识层级来镇压反抗。我建立这个模型的目的是为了激发更多的相关研究，旨在最终挫败利维坦式的策略。

4. 反抗与报复：监禁博弈

在其著作的第 6 章中，范德施拉夫讨论了利维坦可能造成的危害。如果全体公民皆在法律统治下生活，唯有利维坦凌驾于法律之上，是不是需要有驯服利维坦的策略？人们在范德施拉夫的著作中（第 215 页）饶有趣味地读到，在整合了休谟之论的基础上，提出有必要以民众反抗的威胁对最高权力“施加某种制约”。这也提醒我们不要忘记，在领导人尚未登上权力之巅前，也必须形成共识或者订立一部宪法，以便对其未来的所作所为定规设限。利维坦可能也需要被驯服。沿着范德施拉夫的分析进一步探讨，我们还可以获得一些深刻洞见。通常，权力仅被视为掌权人之既有，但在有些案例中，权力是在即将倾覆之际，被机敏的博弈者在应对策略不确定性过程中获得或保留下来的，故而也称为“多变权力”（protean power，Katzenstein and Seybert，2018）。我将从一个有趣的角度，展现远在博弈论尚未创立之前，独裁者就有意无意长期玩弄的一种权变策略。从博弈论的角度进行形式化分析，可以使我们对此有更好的理解。

世界上的民众反抗与国内战争，亘古于今，连绵不绝。如何从抽象的理论高度认识此类事件，的确是令人神往的探索，这就像人们着迷于了解为什么会有覆盖全社会的秩序那样，无论实现秩序的是市场的“看不见的手”，还是利维坦的铁腕。只不过，有关民众反抗的研究项目往往会力不从心，半途而废，因为从本质上说，反抗是一个与秩序、和谐相矛盾的概念，对它建模不免相对困难。但近期，学术界在对无政府状态和国家性质的理解上，已获得了长足进步（Moehler，2009；Vanderschraaf，2006），因此对反抗的理解也有望取得这样的进步。在本文接下来的一节里，我想对这个大题目略尽绵薄之力。

在历史上，有时革命与内战在表面和平的时代突然爆发，有时反抗又在即将胜利之际惨遭镇压，重陷沉寂，此等案例，比比皆是。仅在最近几十年中，两方面的情况皆有发生。人们看到了风潮迭起，也看到了在统治者的威胁之下，抗议者撤离后沉寂空旷的街道。

对信念与策略之间微妙互动的观察，使我们可以更深入地了解民众反抗的成败。欲解其详，我们可设想一个暴君，已完全被渴望变革的公民厌弃。

不难想见，官民之间如此广泛的对立，随时都会酿成公开的反叛。只需一起触发事件，就可能引爆反叛的干柴烈火，或者说，为了帮助公民们协调行动，我们也可能需要一位“焦点人物”。

数字技术使这种协调成为可能。没有人愿意贸然独自上街抗议，这样做太危险了，因为领导人会将此人逮捕、监禁或屠杀。但如果成千上万人同时抗议，其中每一个人就会相对较为安全，因为一个领导人能够逮捕和伤害的人毕竟有限。而时间与地点的协调，原本就属于经典的焦点问题（Schelling，1960；Sugden，1995）。

为什么有的反抗获得成功，而有的却归于失败？暴君们到底是采用什么样的手段，破坏了反抗运动的内部协调？人们希望破解压迫者的惯用手段，帮助我们制定法律及形成共识以挫败他们的压迫。

我将借助策略分析、知识层级和博弈论来展示暴君是如何做到压制异议的。博弈论是一种凭直觉即可掌握的东西，并不需要先期的理论积累。因此，暴君们总是会使用类似的策略镇压社会，也就不足为奇了。我们只是走运，有些暴君还没来得及弄懂博弈策略就被赶下了台；但也有其他暴君，对这样的策略烂熟于心。

现在用一个寓言故事来解释这个问题，假设在一个古老城邦有一个领导人，此人年轻时曾深得民心，但后来演变成了暴君，对敢于反对他的所有人进行监禁、酷刑与杀戮，在公民受难时独自拉琴（假想他与罗马帝国的第五个皇帝颇为相似）。假设这个古老城邦有 10 000 001 人，也就是说有 1 000 万个公民外加 1 个领导人。

所有公民都意欲把这个暴君赶下权位，我们假设如果有 100 万以上的人同时参与抗议，暴君的统治即被推翻。接下来再假设所有公民已就抗议的日期和时间达成了一致意见，也就是说焦点已被设立。现在每个公民要做的就是选择参加抗议（P 行动）或选择保持沉默（S 行动）。现在假设，当其他条件不变

时，每个公民都倾向于参加抗议而不是保持沉默。如果选择 P，每个公民的收益是100，而如果选择 S，收益将是0。这样，一个人选择 P 还是选择 S 的决定因素，就是对抗议的预期惩罚。⑪

这时的暴君心里很清楚，革命一旦爆发，他的统治就将崩溃，所以他要竭力阻止革命发生。他开始思忖如何行事。让我们假设他有能力逮捕抗议者并将他们终身监禁。但监狱的容量永远是有限的，在现实世界中几乎总是如此。当监狱的容量只有 10 人，如果超过 10 人抗议，暴君就不能逮捕所有抗议者。现在假设监狱的容量是200。所以，如果一个人被捕入狱的概率为 p，而且 $100-200p>0$，那么，此人仍将参与抗议，也就是说，如果：

$$p \geqslant 1/2 \tag{2}$$

此人将不会参与抗议。我做出的这个无伤大雅的平局假设意味着当选择 P 还是选择 S 并无差别时，人们会选择 S。

如果所有 1 000 万公民都上街抗议，对暴君来说，那就是他的末日。如果他宣布抗议者中只有 10 个会被逮捕和监禁，这样小的入狱概率将不足以威慑任何人。1 000 万人都会上街游行，暴君将失去权力。

暴君所能做的，就是制止某些人的行动。于是，他宣布 10 个人的名单，如果发现他们参与抗议就会逮捕他们，而这 10 个人中的确没人上街抗议。但那还是阻挡不了有更多的人参与抗议。反抗还是会取得胜利。暴君可以宣布更多的名字，让名单包含 20 个人，如果发现这些人参与抗议，就会把其中的 10 个人投入囹圄。很明显，根据式（2），那 20 个人中没人会参加抗议。但这还是不能有效阻止革命的发生；只要有 100 万人参加，革命照样会获得成功。

当暴君知道宣布监禁 10 个抗议者的做法并不能阻止抗议的发生，他便开始向革命妥协。在顾影自怜之余，他决定在被彻底推翻之前，必须做一件好事，表明至少天神朱庇特还没有抛弃他。

假设在这个城邦中，每一个人的年龄都是已知的（为了简单起见，假设

---

⑪ 为简单起见，我假设每个人都视革命成功与个人是否参与无关。对于较大规模的人口来说，这种想法并非没有道理，也使我们的分析更加简洁。所以，从抗议中获得的满意度 100 代表着参与国家重大历史事件的纯粹喜悦。因而塔洛克（1971）提出的公共品问题便与我现在的讨论无关，我也就无须顾忌他的问题。还必须强调的是，我的假设并非不切实际。在投票时，即便明知自己一票无关大局，人们仍会获得满足感；避免向河水里扔塑料，也会让人自我感觉良好，尽管扔一片塑料也不可能造成什么伤害。

所有公民都有一个独特的年龄，没有人同年同月同日生）。暴君宣布，他将依言监禁10名抗议者，但又要显示自己的一丝宽仁。为了让朱庇特感到高兴，他将从所有抗议者中监禁10名最年长者，以便最大限度地缩小公民被监禁的时间。请记住，这是终身监禁。因此，如果人们有相同的预期寿命，把10名最年长的抗议者投入监狱就能最大限度地缩短抗议者的集体服刑时间。暴君停止拉琴，站在房顶上宣布他的决定，以便让全城都能听到。

暴君本来以为抗议风潮一定会席卷全城，他的权位也一定不保。他只不过希望朱庇特会对他有些许宽容。但是在抗议运动预定发动的那一天，当他环顾四周，等待事变，却惊讶地发现，街上竟然没有出现任何抗议者的身影，就连根本称不上长者的青少年也没有。反抗随即挫败，暴政得以继续。

这不难看出到底发生了什么情况。我把这场公民在暴君发布告示后参与的博弈称为监禁博弈。为了论证方便，我根据年龄为公民排序：最年长者为1，仅次于他的为2，以此类推，直至最年幼者为10 000 000。

请注意：在监禁博弈中，公民中的1～10会放弃参与抗议的念头。他们知道，一旦自己出现在抗议人群里，就势必遭到逮捕和监禁，因而每个人都会选择S。接下来做选择的是公民11，12……直到20，他们也都是有判断能力的人，当他们知道了前10名公民将如何行事后，也就知道了如果自己参与抗议，就会遭到监禁。所以他们也选择S。当第1名到第20名公民选择了S，同样的逻辑驱使接下来的10名公民做出同样的选择。这一过程不断继续，直到所有1 000万人都选择沉默而不是抗议。

监禁博弈中涉及的推理，在博弈论和分析哲学里，皆属寻常现象，在突击测验悖论（O'Connor，1948；Quine，1953；Scriven，1951）、蜈蚣博弈（Binmore，1996；Rosenthal，1981），以及游客困境（Basu，1994；Halpern and Pass，2012；Rubinstein，2016）里都有相关描述。但如果有人以为这些共同知识的推理仅仅是纯粹的学术演练，那便是犯了一个愚蠢的错误。⑫ 虽然独裁者未见得通晓这种推理的门道，却时常利用，且屡屡得手。我们所要做的就是把握这种策略思维的作用和层级化知识的影响力。

---

⑫ 知识层级的应用随处可见，从各个学科，到英国侦探小说，再到现实生活场景，如Sarangi（2020）讨论过的在海滩上吃pau bhaji（印度的一种食物）的情况。值得记取的，还有源于真实生活问题的将军们如何协调对敌攻击的Rubinstein（1989）电子邮件游戏（即区块链技术经常会提到的“拜占庭将军问题”——译者注）。

在任何具有一定规模人口的国家里，把我假设的那种逮捕、监禁或处决的规则变成共同知识，实际上是不可能的。但在上述问题中，共同知识是一个充分条件，并不是必要条件。[13] 即使没有这样的共同知识，反抗也可能被镇压。在现实中，这主要取决于人们有多接近共同知识。以自上而下的镇压而论，暴君可以通过操纵新闻扩散的方法，积累关于镇压策略的知识层级，进而达到阻止反抗的效果。比如在以上博弈中，如果根据现有信息，难以把 1 000 万人在惩治过程中一一排序，却仍然有可能把社会分解为 20 人一组的 50 万个小组，并规定它们被投入监禁的次序，哪一组第一，哪一组第二，等等。

暴君可以向社会表明，如果出现大规模抗议风潮，最先遭到逮捕的不是普通公民，而是反对派领导人。而如果反对派领导人不在现场，他就会逮捕对政府发表批评意见的媒体工作者。而如果媒体工作者都噤若寒蝉，他又会逮捕其他某个类别的人士。

在另一方面，为了达到这一目的而同时采用的宣传手段却简陋至极。暴君只消说，“我只抓抗议者中的最年长者”，如果这变成了共同知识，如果每个人不仅知道所有其他人的年龄，而且知道所有其他人也知道所有其他人的年龄，如此类推，那么，想要压制所有的不同政见，仅这一条告示就已足够。余下的事，皆由社会存在的种种知识结构自动完成。一旦社会形成了众所周知的镇压顺序，包括哪些人会首当其冲，哪些人会紧随其后等，就会形成逆向归纳的想法，也就没有人再会抗议。暴君的镇压将继续存在，而且不会减弱。

如上分析的要义，就是指明异见人士遭到镇压的充分条件。基于人们规避风险的倾向，以及对不同群体先后遭到镇压的部分知识，即便没有共同知识，他们也会从反抗中退出。

当然，以上分析并未考虑现实中的各种复杂情况。首先，共同知识的建立并不能做到万无一失。它需要透视他人的思想，仅这一点，就极难办到。所以，压制异议的各种企图终会失败，多数情况下也只是依赖于领导人的直觉和手段方能得逞。其次，也总有一些个人，如甘地、哈维尔、曼德拉与马丁·路德金，因道德上的执着而进入了博弈论中所谓的“非理性”状态。这些人的行为不会因威胁而改变。变革经常随着这些人物的出现而发生。正如萧伯纳在

---

[13] 如果将这一问题应用于一个人口无限大的国家，共同知识就变成一个必要条件。在这个意义上，就像在 Rubinstein（1989）电子邮件游戏里那样，缺少了难以计数的知识层级，任何事都无法运作。

《人与超人》中所说的，“理智者让自己适应世界；非理智者却执着于让世界适应自己。于是，所有进步皆有赖于非理智者的推动。”

人有了道德罗盘，就会努力增进个人行为的社会价值。但正如格蕾塔困境表明的，万无一失的解决方案并不存在。在有些策略性环境下，善良之心也会造成无意之害。

本文的目的不是解决问题，而是要表明，以范德施拉夫的著作与持续增多的道德哲学及博弈论著作为代表的哲学辩论是多么接近现实。这实属不易。独裁者无法设计惩治办法并把它变为共同知识或准共同知识，也无法应对少数执着于道德的社会成员，因此不容易阻止反抗。同样，我们也很难通过事前形成共识和创立宪法以约束领导人的权力。监禁博弈表明，在形成种种共识约束利维坦权力的过程中，我们不得不面对这样的挑战。

5. 结语

监禁博弈旨在向世人揭示暴君有意无意沿用的一条政治压迫路径。对此做形式化分析，是为了帮助我们广大公民更加看清压迫的规律。但就如何破解暴君的压迫，说实话，我并没有具体的解决方案，除了再次指向本文的起点，那就是所有社会都需要在领导人登上权位之前，就已然完成一定的建章立制。这也说明，我们需要超越人生博弈的视野，去思考如何采取事前集体行动以保证博弈指向一个正义、公平与平等的结果。⑭ 本文的最后一节是为这样的探索抛砖引玉。

最后，我还要就博弈论与哲学的一个表面分歧发表一点看法。这一看法与“人生博弈”，也就是人类普遍置身的博弈有关（Binmore，1995）。我们经常以描述人生博弈开始分析。当结果不甚理想时，就讨论如何加以纠正。但我们是否应该将我们的思维从标准博弈论的“以我为基础”转向以集体为基础（Gauthier，1987；List and Pettit，2011；Schwenkenbecher，2019）？我们是否应该下定决心抛弃行动的结果主义，转而拥抱某些义务论的承诺？麻烦在于，这些问题会引导我们思考人生博弈之外的各种策略，而根据定义，这些策略是一个并不存在的领域。这实际上就形成了一个矛盾：一方面，这些问题的解决方

⑭ 作为一个规范性问题而提出，那就是我们“应该”做什么？同样的问题也经常面临制度变迁的分析，特别是制度消亡先导过程的研究；这种过程，以标准的博弈论观点看来，即是自我实现的过程（Greif and Laitin，2004）。

案必须被视为取得未来进步的优先选项；但与此同时，我们又不得不承认我们的探索有时会把我们带向根本无解的问题。

（张晓刚　译）

参考文献

Arrow, K. , & Debreu, G. (1954). "Existence of an equilibrium for a competitive economy." *Econometrica*, 22, 265 – 290.

Aumann, R. , & Myerson, R. (1988). "Endogenous formation of links between players and of coalitions." In A. Roth (Ed.), *The shapley value: Essays in Honor of Lloyd Shapley*. Cambridge: Cambridge University Press.

Basu, K. (1994). "The Traveler's dilemma: Paradoxes of rationality in game theory." *American Economic Review, Papers and Proceedings*, 84, 391 – 395.

Basu, K. (2018). "*The republic of beliefs: A new approach to law and economics.*" Princeton University Press.

Basu, K. (2021a). "The Samaritan's curse: Moral individuals and immoral groups." *Economics and Philosophy*, 37, 132 – 151.

Basu, K. (2021b). "Why have leaders at all? Hume and Hobbes, with a dash of Nash." *Homo Oeconomicus* (forthcoming).

Basu, K. (2021c). "The ground beneath our feet." *Oxford Review of Economic Policy* (forthcoming).

Binmore, K. (1995). "The game of life." *Journal of Institutional and Theoretical Economics*, 151 (1), 132 – 156.

Binmore, K. (1996). "A note on backward induction." *Games and Economic Behavior*, 17, 305.

Boumlik, H. , & Schwartz, J. (2016). "Conscientization and third space: A case study of Tunisian activism." *Adult Education Quarterly*, 66, 319 – 335.

Braham, M. , & van Hees, M. (2012). "An anatomy of moral responsibility." *Mind*, 121, 601 – 634.

Braithwaite, R. (1955). *Theory of games as a tool for the moral philosopher*. Cambridge University Press.

Clark, A. , & Chalmers, D. (1998). "The extended mind." *Analysis*, 58, 7 – 19.

Dasgupta, U. , & Radoniqui, F. (2021). *The republic of beliefs: An experimental investigation*. IZA Discussion Paper, No. 14130.

Fearon, J. , & Laitin, D. (2014). "Civil war non-onsets: The case of Japan." *Journal of Civilizational Studies*, 1, 71 – 94.

Feinberg, J. (1968). "Collective responsibility." *Journal of Philosophy*, 65, 674 – 688.

Fioretti, G. , & Policarpi, A. (2020). *The Less Intelligent the Elements, the More Intelligent the Whole. Or Possibly Not?* mimeo: Universityof Bologna.

Gauthier, D. (1987). *Morals by Agreement*. Oxford University Press.

Gauthier, D. (1990). *Moral dealing: Contract*. Cornell University Press.

Genicot, G. , & Ray, D. (2003). "Group formation in risk-sharing arrangements." *Review of Economic Studies*, 70, 87 – 113.

Greif, A. , & Laitin, D. (2004). "A theory of endogenous institutional change." *American Political Sci-*

ence Review, 98, 633 – 652.

Halpern, J., & Pass, R. (2012). "Iterated regret minimization: A new solution concept." *Games and Economic Behavior*, 74, 184 – 207.

Havel, V. (1986). "The power of the powerless." In J. Vladislav (Ed.), *Living in truth*. Faber & Faber.

Hobbes, T. (1651). *Leviathan* (Cambridge University Press, ed. R. Tuck, 1991).

Hume, D. (1740). *A treatise of human nature* (Oxford University Press, eds. D. F. Norton and M. J. Norton, 2000).

Katzenstein, P., & Seybert, L. (2018). "Protean power and uncertainty: Exploring the unexpected in world politics." *International Studies Quarterly*, 62, 80 – 93.

Lewis, D. (1969). *Convention*. Harvard University Press.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and the status of corporate agents*. Oxford University Press.

Markovits, J. (2010). "Acting for theright reasons." *Philosophical Studies*, 119, 201 – 243.

Moehler, M. (2009). "Why Hobbes' state of nature is best modeled as an assurance game." *Utilitas*, 21, 297 – 326.

Moehler, M. (2018). "Diversity, stability and social contract theory." *Philosophical Studies*, 176, 3285 – 3301.

Myerson, R. (2021). "Village communities and global development." *Horizons: Journal of International Relations and Sustainable Development*, 18, 228 – 241.

Myerson, R., & Weibull, J. (2015). "Tenable strategy blocks and settled equilibria." *Econometrica*, 83, 943 – 976.

Nash, J. (1951). "Non-cooperative games." *Annals of Mathematics*, 54, 286 – 295.

O'Connor, D. J. (1948). "Pragmatic paradoxes." *Mind*, 57, 358 – 359.

Putnam, H. (2005). *Ethics without ontology*. Harvard University Press.

Quine, W. V. (1953). "On a so-called paradox." *Mind*, 62, 65 – 67.

Quong, J. (2005). "Disagreement, asymmetry, and liberal legitimacy." *Philosophy & Economics*, 4, 301 – 330.

Ray, D., & Vohra, R. (2015). "The far-sighted stable set." *Econometrica*, 83, 977 – 1011.

Rosenthal, R. (1981). "Games of perfect information, predatory pricing, and the chain store." *Journal of Economic Theory*, 25, 92 – 100.

Rubinstein, A. (1989). "The electronic mail game: Strategic behavior undercomplete uncertainty." *American EconomicReview*, 79, 385 – 391.

Rubinstein, A. (2016). "A typologyof players: Between instinctive and contemplative." *Quarterly Journal of Economics*, 131, 859 – 890.

Runciman, W., & Sen, A. (1965). "Games, justice andthe general will." *Mind*, 74, 554 – 562.

Sarangi, S. (2020). *The economics of small things*. Penguin.

Sartorio, C. (2004). "How to be responsible for something without causingit." *Philosophical Perspectives*, 18, 315 – 336.

Schelling, T. (1960). *Strategy of conflict*. Harvard University Press.

Scriven, M. (1951). "Paradoxical announcements." *Mind*, 60, 403 – 407.

Smith, A. (1776). *An inquiry into the nature and causes of the wealth of nations*. Clarendon Press edition, 1978.

Sugden, R. (1986). *The economics of rights*. Palgrave Macmillan, Houndmills: Cooperation and Welfare.

Sugden, R. (1995). "A theory of focal points." *Economic Journal*, 105, 533 – 550.

Schwenkenbecher, A. (2019). "Collective moral obligations: 'We-reasoning' and the perspective of the deliberating agent." *The Monist*, 102, 151 – 171.

Thrasher, J., & Vallier, K. (2015). "The fragility of consensus: Public reason, diversity and stability." *European Journalof Philosophy*, 23, 933 – 954.

Tullock, G. (1971). "The paradoxof revolution." *Public Choice*, 11, 89 – 99.

Vanderschraaf, P. (2006). "War or peace? A dynamical analysis of anarchy." *Economics and Philosophy*, 22, 243 – 279.

Vanderschraaf, P. (2008). "Game theory meets threshold analysis: Reappraising the paradoxes of anarchy and revolution." *BritishJournal forthe Philosophyof Science*, 59, 579 – 617.

Vanderschraaf, P. (2019). *Strategic justice: Conventions and problems of balancing divergent interests.* Oxford University Press.

Vanderschraaf, P. (2021). "Contractarianisms and markets." *Journal of Economic Behavior and Organization*, 181, 270 – 287.

Von Neumann, V. J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton University Press.

Walras, L. (1874). Elements d'economie politique pure [4th edition in 1900. English translation by W. Jaffe, in 1954].