伦理学的计算复杂性：心智与机器的道德可处理性

Jakob Stenseke[a, b]

[a]隆德大学哲学系，*Helgonavagen 3*，隆德 *221 00*，瑞典[b]*jakob.stenseke@fil.lu.se, 0000-0001-8579-3975*

arXiv:2302.04218v1 [cs.CC] 8 Feb 2023

摘要

为什么道德哲学家、道德心理学家和机器伦理学家应当关注计算复杂性？关于人工智能（AI）能否或是否应当用于解决伦理领域问题的争论，主要围绕人工智能在人类能力方面能做什么或不能做什么而展开。本文从另一个角度探讨这一问题：根据计算系统能做什么或不能做什么，哪些道德机器才有可能实现。为此，我们从计算复杂性的视角分析规范伦理学。首先，我们向初学者介绍计算复杂性，并讨论如何在马尔的三个分析层次框架下描述伦理问题的复杂性。随后，我们考察一系列基于结果主义、义务论和美德伦理学的伦理问题，旨在阐明与问题本身（例如由组合学、不确定性、策略动态所致）、所采用的计算方法（例如概率、逻辑、学习）以及可用资源（例如时间、知识、学习）相关的复杂性。结果表明，规范框架提出的大多数问题，在我们分析的每个类别中都会引发可处理性问题。我们的研究还揭示了规范伦理学计算本质的若干特点，包括规则型与基于结果的道德策略之间的差异，以及道德资源方面的实现差异。接着，我们讨论复杂性结果对道德机器前景的影响，尤其是最优性与效率之间的权衡。最后，我们阐明如何通过推进道德可处理性论题（MTT），运用计算复杂性为有关人类道德的哲学研究和认知心理学研究提供参考。

关键词：计算复杂性、机器伦理学、人工道德主体、结果主义、义务论、美德伦理学

目录

1 引言 2

2 制作沙拉的复杂度 4

3 伦理学的计算复杂性 83.1 按照 Marr 的三层次分析探讨伦理的复杂性 . . . . . 10 3.1.1 层次一：计算问题.10

3.1.2 第 2 级：算法.13 3.1.3 第 3 层：实现.15

4 结果主义　　与因果引擎 1　　7

4.1 结果组合学.18 4.2 因果推断.21 4.3 动态与部分可观测环境中的决策 25

5 义务论与规则遵循者 30

5.1 规则的普遍性.32 5.1.1 法律与自由主义语境中的人类规则遵循 . . . . . 33 5.1.2 通用规则及其辩护.34 5.1.3 策略博弈中的道德行为.36 5.1.4 不完全信息策略博弈中的道德行为 . . . 39 5.1.5 计算道德均衡.41 5.1.6 道德系统设计.45 5.2 道德逻辑与语义学的复杂度.46 5.2.1 可判定性与描述复杂性 . . . . . 47 5.2.2 模态逻辑与道义逻辑的复杂度 . . . . . 49 5.2.3 道德语义学问题.52 5.3 结果主义—义务论混合理论.58

6 美德伦理学与道德机器学习 61

6.1 学习的复杂度.62 6.1.1 弱样本复杂度与强样本复杂度.64 6.1.2 机器学习理论与实践.68

7 心智与机器的道德可处理性 69

7.1 对道德机器前景的影响.69 7.2 对人类道德的影响.71 7.3 道德可处理性论题.72 7.4 局限性.74 7.5 未来工作.77

8 结论 77

1. 引言

硬件和软件构成的计算系统不断进入并改变越来越多的人类领域。随着自动驾驶车辆、虚拟教师和护理机器人增强甚至取代驾驶员、教育者和照护者等传统人类角色，人们越来越难以忽视这样一种需求：这些系统应当与这些角色所关联的规范和道德标准保持一致。[1] 这些关切催生了跨学科领域——机器

[1]除非另有说明，“人工智能系统”“机器”和“计算机”等术语将交替使用，统称由硬件和软件构成的计算系统。

伦理学，该领域广泛探讨将伦理学付诸机器实践的前景（Wallach 和 Allen，2008年；Anderson 和 Anderson，2011年）。机器伦理学位于计算机科学与道德哲学的交叉领域，涵盖一系列或多或少相互关联的研究目标，包括探讨现有人工智能方法的价值对齐（Gabriel，2020年）、可解释性（Gunning 等人，2019年）和安全性（Amodei 等人，2016年）所面临的挑战，开发应对各种伦理困境的系统（Cervantes 等人，2020年；Tolmeijer 等人，2020年），以及围绕人工道德主体是否可行或值得追求、以及在多大程度上可行或值得追求的理论争论（Floridi 和 Sanders，2004；Behdadi 和 Munthe，2020年）。

可行性争论主要围绕人工智能系统在人类能力方面能做什么或不能做什么而展开；例如，人工智能体能否自主行动或拥有自由意志（Hellstr¨om，2013年）、能否具备类人理性（Purves 等人，2015年），或能否拥有意识体验（Himma，2009）。然而，人工道德的讨论聚焦于那些从计算视角来看仍难以捉摸、概念上也含糊不清的能力，未能触及人工智能的技术层面，因此对伦理机器的设计与开发在实践中无甚助益（Mabaso，2021年；Behdadi 和 Munthe，2020年；Stenseke，2022b）。另一个令道德机器的可行性难以评估的问题，是缺乏系统性评估工具（Tolmeijer 等人，2020年）。在机器伦理学领域，目前既没有针对特定论域的基准，也没有可用于评估不同伦理系统性能的通用基准。因此，由于对系统的评估仅限于其特定实现所处的实验条件，解决方案的可扩展性和结果的可推广性都受到严重限制。

在本文中，我们通过探讨基于计算系统能够或无法有效解决的伦理问题，哪些类型的道德机器是可能的，来应对这些问题。为此，我们透过计算复杂性理论的视角分析规范伦理学；该理论根据计算机解决问题所需的资源（例如时间和空间）对问题进行分类。尽管先前的研究以更非正式的方式讨论了道德机器的计算限制（Brundage，2014年；Stenseke 和 Balkenius，2022），并对伦理行为（Reynolds，2005）和计划（Lindner 等，2020）的复杂度进行了初步分析，但伦理学的计算复杂性及其对机器伦理学的潜在相关性仍在很大程度上未得到探索。例如，如果人工系统要在时间至关重要的伦理领域中运行（例如自动驾驶救护车），那么确保此类系统能够高效且胜任地作出伦理决策至关重要。此外，如果人类道德认知受到可处理性的约束（Van Rooij，2008），那么这项分析还可以通过限制遵循某一规范理论的智能体能够合理预期解决的问题空间，为道德心理学和规范理论提供帮助。

在本文其余部分，我们将介绍并阐释道德哲学和计算机科学中的概念与理论，尽量让对其中一个或两个领域了解有限的读者也能轻松理解。全文结构如下。首先，我们介绍计算复杂性与可处理性，旨在向初学者说明它们为何重要（第2节）。第3节中，我们回顾机器伦理学领域以往的实现，并运用马尔的三个分析层次（3.1）讨论对伦理复杂性的各种解释。这引出了对规范理论所提出问题的分析（计算层，3.1.1）：这些问题由确定性图灵机通过多种计算方法（算法层，3.1.2）加以解决（实现层，3.1.3）。随后，我们将探讨基于结果主义（4）、

义务论（5）和美德伦理学（6）的各种伦理问题的复杂度。本文的主要目标是阐明与问题本身相关的复杂度（例如由不确定性、组合学、策略动态和一般性造成的复杂度）、可用资源（例如时间、认知和领域知识），以及用于解决这些问题的计算方法（例如概率、逻辑和学习）。研究结果表明，规范理论提出的大多数问题都会导致可处理性问题（简要总结见第3节），尤其是在应当以最优方式满足规范性理想时。具体而言，鉴于行动计划组合学（4.1）、概率因果推断（4.2）、动态与部分可观测环境（4.3）、一般规则（5.1）、策略动态（5.1.3）、逻辑（5.2）、语义学（5.2.3）和学习（6.1）所导致的不可处理性（及不可判定性），我们有充分理由得出结论：完美道德机器是不可能实现的。我们的研究还进一步揭示了规范理论在计算层面的性质，包括：（i）行动策略与基于结果的策略之间的差异；（ii）道德混合体的益处（5.3）；以及（iii）道德资源方面极大的实现差异。第7节中，我们聚焦于最优性与效率之间的权衡、规范理论的模棱两可作用，以及不同道德资源之间的密切关系，探讨这些结果对道德机器前景的影响。最后，我们通过提出道德可处理性论题，展示如何运用计算可处理性为关于人类道德的哲学与心理学研究提供参考。

2. 制作沙拉的复杂性

让我们先从一个例子开始。[2] 你很可能去过沙拉吧，在那里可以按照自己的喜好挑选食材。[3] 问题是，要做出最好吃的沙拉，你会挑选哪些食材？假设你能立即单独评估每种食材的美味程度，并在从最令人倒胃口（-10）到最美味（$+10$）的范围内，为它们赋予一个“味道值”（v）。有了这些值，你会发现，做出一份不错沙拉的一种高效方法是，只挑选味道值为正（I）的食材（ v （v（I） > 0），或味道值高于某个阈值的食材（例如， v （I） > 5）。我们将这种策略称为 Ψ。事实上，随着排在沙拉吧后面的队伍越来越长，你会欣赏到 Ψ 让你制作沙拉的速度：你只需逐一查看每种食材，并判断它是否足够美味，值得加入你的搭配中。此外，你会发现， Ψ 的性能在最坏情况下会随沙拉配料的数量线性增长。这意味着，无论食材有多少， Ψ 始终高效：对于任意输入——在这里就是食材的数量——制作沙拉所需的时间都会与输入规模密切对应（也就是说，1000 种食材就需要逐一查看 1000 种不同食材）。

但进一步思考后，你意识到 Ψ 有些古怪。它让你把日晒番茄放在菠萝上面。你想象着日晒番茄的咸味与菠萝的酸甜味在舌头的味蕾间交融。你对这个画面的本能厌恶揭示了 Ψ 的一个致命缺陷：即使这两种

[2]建议已经熟悉计算复杂性的读者直接跳到第3节。[3]这个例子受 Van Rooij 等人（2019）所作的精彩复杂性分析入门介绍启发。

食材的味道值都很高（$v > 9$），它们组合起来却会产生一种令人极其反感的味道值（$v = -10$）。你意识到 Ψ 违反了美食学的一项基本原则，即食材组合产生的味道值不一定与各自食材的美味程度相对应。我们可以将这一原则称为美食学组合原理（CPG）。

幸运的是，你拥有完美的味觉想象力，能够立即评估任意给定食材组合的味道值。如何找到能使味道值最大化且符合 CPG 的最优食材组合？我们可以将其正式描述为以下计算问题：

OPTIMAL SALAD FOLLOWING CPG

输入： 一个沙拉吧，包含一个由 $SB = \{I_1, I_2, ..., I_n\}$ 种 n 食材构成的集合，以及一个值函数 v，它为每个子集（或沙拉）$S \subseteq SB$ 赋予一个味道值。
输出： 一个沙拉 $S \subseteq SB$，使得 $v(S)$ 在沙拉吧中所有可能的沙拉里达到最大值（$S \subseteq SB$）。

你意识到有一种简单直接的策略，可以保证在满足 CPG 的同时做出最优沙拉，于是你将它称为 Φ：只需想象每一个可能子集 $S \subseteq SB$ 的味道值，并选出 $v(S)$ 最高的沙拉。但你隐约觉得 Φ 肯定有蹊跷。你做了一些基础的组合学计算：如果只有一种食材，例如 $\{cucumber\}$，就只有一种可能的沙拉（完全由黄瓜制成）；两种食材可以做出三种不同的沙拉，例如 $\{cucumber\}$、$\{onion\}$、$\{cucumber, onion\}$；三种食材可以做出七种；四种食材可以做出十五种；以此类推。你发现，可能的沙拉数量会随着食材数量呈指数级增长，因此 n 种食材可以做出 $2^n - 1$ 种可能的沙拉。你眼前的沙拉吧有 30 种食材，这意味着共有 $2^{30} - 1 = 1$、073、741、823 种不同的沙拉。由于你那非凡的味觉系统每秒只能评估一种沙拉的味道，Φ 就得想象沙拉约 34 年，才能最优地满足美食学组合原理。不幸的是，你已经浪费了太多时间，排在你身后的人们都非常不满。

这个例子旨在说明有关计算复杂性的五个重要教训：

（1）首先，我们在日常生活中遇到的许多判定问题，都可以用类似的方式来表述，例如规划行程、为旅行打包行李，或邀请一群朋友到你的小公寓参加生日派对。正如我们将在本文中看到的，伦理问题也不例外。你或许会纳闷，找到最优沙拉为什么如此重要；最坏的情况不过是做出一份味道不佳的沙拉，后果远谈不上灾难性。但如果问题关系到生死，你还会如此轻易地忽视最优结果吗？即便你不太在意美食学组合原理，也可能有一些道德原则对你的伦理生活至关重要。

(2) 第二个启示是，问题的复杂度可以用智能体或算法解决该问题所需的资源来表示。对于计算系统而言，最值得关注的两种资源是时间 和空间。后者通常指计算机内存的大小（例如，比特数），而前者指机器的

操作次数（或同义词“计算”“运算”“步骤”或“状态转移”）。为什么要用机器操作次数而非秒或分钟来衡量时间？原因在于，尽管计算机通过运行某种算法 A 解决问题的实际速度可能相差很大，但它们需要执行的机器操作数量 A 却保持不变。21世纪的计算机和20世纪60年代的计算机，如果要按照 Φ制作出味道最佳的沙拉，都必须考虑 $2^n - 1$ 沙拉，尽管现代计算机的速度可能快上百万倍。重要的是，这构成了不变性论题的基础，[4] 使我们能够分析和比较计算问题固有的最坏情况复杂度，而不受特定机器的影响。

（3）这引出了第三个启示：一个简单的观察，即有些问题比其他问题更复杂。如果一个问题是不可判定的，这意味着可以证明，不可能构造出任何算法来解决该问题。[5] 在可判定问题中，最重要的区分是易处理问题与难处理问题。粗略地说，如果一个问题可以使用“现实可行”的资源量来解决，那么它就是易处理的。然而，对大多数计算理论学家而言，易处理与“可在多项式时间内计算”同义。这意味着，一个算法的运行时间（机器操作的次数）以上界受其输入的多项式表达式约束，即属于 n^c 类型（其中 c 是某个正常数）。这包括随着输入 n 增加而呈现对数（$\log n$）、线性（n）、二次（n^2）或三次（n^3）时间增长的函数。能够由确定性图灵机在多项式时间内解决的判定问题类称为 P，它体现了具有“有效”判定程序的判定问题这一概念（Cobham，1965年）。相反，无法在多项式时间内解决的问题被称为难处理问题，因为其运行时间呈指数增长（c^n）、按阶乘增长（$n!$）或超指数增长（n^n）。这种可处理性概念可以通过判定程序 Ψ 与 Φ之间的差异来说明。对于 Ψ而言，制作沙拉所需的时间相对于食材数量的增长永远不会超过线性增长。使用大O符号来表示函数的渐近上界[6] （在此处，即输入规模 n 与时间之间的映射），Ψ 的时间复杂度为 $O(n)$。相比之下，对沙拉吧中所有可能的沙拉进行穷举搜索，会使 Φ呈指数增长 $O(2^n)$。即使你的味觉想象力能够利用大脑并行化神经计算的速度，使你每秒想象十亿种沙拉，一个包含 50 种食材的沙拉吧仍然需要你花 13 天才能穷尽，而包含 60 种食材的沙拉吧则大约需要 370 亿年。当然，没有理智的人会花那么长时间去想象不同沙拉的味道。但问题的难题依然存在：如果我们想要

[4]更正式地说，该论题指出，给定两台机器 M_1 和 M_2以及一个给定的计算问题Θ，由 M_1 和 M_2 执行 Θ 的复杂度至多相差一个多项式量。也就是说，如果 M_1 能够在时间 t内计算 Θ，那么 M_2 就能在 t^c内计算 Θ，其中 c 是一个常数。计算机科学家普遍接受这一论题，前提是 M_1 和 M_2 属于任何类型的图灵机或其他合理的计算模型（例如元胞自动机、神经网络），且输入的编码合理（例如，不包含无关信息）。参见 Garey and Johnson （1979 年）。[5] 停机问题是不可判定问题的一个例子：1936 年，艾伦·图灵证明，不存在能够判断任意程序最终是否会停机的算法。可判定性将在第5.2.1节进一步讨论。[6]“渐近”意味着，在描述一个函数时，我们可以忽略低阶多项式项和常数。例如，函数 $f(n) = 4n + n^3$ 写作 $O(n^3)$，因为随着 n 增大，$4n$ 相较于 n^3 变得微不足道。

有效地解决这些问题，我们可能需要放弃对最优性的要求。我们需要的不是“能够想象出的最佳方案”，而是在现有资源条件下“足够好”的折中方案。因此，难处理问题可能会在理想与可行之间造成令人不适的权衡。恰恰是这种令人不适的权衡如何影响计算主体的伦理决策，将成为本文的主题。

遵循 CPG 制作最优沙拉之所以没有有效方法，部分原因可以用广为人信的猜想 P $\neq$ NP 来解释。该猜想指出，有些决策问题虽然其解决方案可以被有效地检查 （或验证），却未必能被有效地求解。更准确地说，它指出复杂度类 P 不等于 NP：NP 是指可由非确定性 TM 在多项式时间内求解的决策问题类；等价地说，这类问题的解决方案可以在多项式时间内得到验证。 Φ 便是这样的一个例子。即使你能快速（在多项式时间内）检查任意一组沙拉配料的味道，也没有确定性过程能让你找到最优方案；要确保找到最优子集，你仍须检查整个组合空间。事实上，遵循 CPG 寻找最优沙拉是 NP难问题，这意味着它至少与 NP 中最难的问题一样难。更形式化地说，当 NP 中的每个 问题都能在多项式时间内归约 到问题 X 时， X 就是 NP难问题。这意味着，如果我们假设求解 X 需要一个时间单位，那么便可以利用该解决方案在多项式时间内求解 NP 中的每个问题。与之密切相关的一个性质是完备性。NP完全问题既是 NP难问题，又 属于 NP。请注意，P 和 NP 是决策问题类——这类问题可以表述为是／否问题——而 NP-hard问题并不限于决策问题；它们只需至少与同一问题最难的判定版本一样难。例如，NP难问题的判定变体可能是 NP完全问题，如布尔可满足性问题（SAT）或子集和问题（SSP）；但同一问题的其他变体，例如表述为优化问题或搜索问题时，则不是 NP完全问题（因为它们不是决策问题）。同样，我们的例子也说明了这一点：遵循 CPG 制作沙拉是 NP难问题，因为它是子集和问题（SSP）的一个优化版本，而后者是 NP完全问题。

此外还要注意，虽然 NP困难性只表示一个一般的下界，却没有说明上界；而上界可能更有助于我们准确理解一个问题究竟有多 难。[7] 在计算复杂性理论中，计算问题的类别则根据它们在最坏情况下所需资源量的上界（或限制）来定义（使用大O符号形式化）。这反过来使我们能够描述问题复杂程度的一般层级。例如，确定性图灵机能在多项式时间内解决的问题，非确定性图灵机也能解决，这意味着 P类是 NP 的子集（$\subseteq$）。同样，人们普遍认为 P $\subseteq$ NP $\subseteq$ PSPACE$\subseteq$ EXPTIME $\subseteq$ EXPSPACE（本文所用复杂性类的概述见附录（8））。最坏情况分析的依据是：算法需要考虑问题的所有可能输入，其中包括最坏情况的输入。但在算法分析中，还有其他几种研究复杂度的重要工具。例如，如果下界与上界重合，我们就得到了一个紧界。同样，

[7]例如，停机问题是 NP难但不可判定的（无法在有限次数的操作内判定）；真正的量化布尔公式语言（QBF）是 NP难的，但可在多项式空间内判定（PSPACE 完全）（Garey and Johnson，1979年）。

按照 Φ制作最优沙拉的时间复杂度也有一个紧界：为了确保最优性，我们必须设想至少且至多 $2^n - 1$ 份沙拉的味道。或者，我们可以设想沙拉吧的摆放方式有利于利用，例如将预制组合分行排列。如果是这样，我们就可以根据在若干不同沙拉吧（输入）中制作沙拉所需的操作次数来衡量算法的时间复杂度，并考察它在最佳情况、平均情况和最坏情况下的表现。[8] 简言之，计算复杂性提供了丰富多样的分析工具，帮助我们理解问题及其算法解决方案的难度。

（4）第四个启示，也是第三个启示的推论，是智能体解决问题的方式最终取决于它所拥有的资源（广义而言）。除了时间和内存大小，这些资源还包括启发式方法（高效策略）、认知（在世界中感知和行动的能力）、[9], 知识和学习。实际上，你或许会把 Ψ 和 Φ 的某些方面结合起来。你可以挑选几种关键食材作为基础，因为你已经知道这样做能做出味道相当不错的沙拉，然后设想这个基础是否能通过进一步添加食材而得到改进。凭借丰富的烹饪经验——结合以往的试错、一般规则、经验法则和有根据的推测——你能够迅速做出近乎完美的沙拉，同时仍遵循 CPG（尽管并非最优）。事实上，甚至在你看到吧台提供什么之前，你的胃可能就已经知道自己想吃什么样的沙拉；你只需挑选食材。在这种情况下，较小的输入规模（例如 10 种食材）可能非但不是福音，反而会成为负担，因为你会发现缺少一种关键食材。重点在于，尽管某些问题可能因特定资源（例如时间）而难处理，或因策略的选择（例如 Φ）而难处理，但在特定情境下，很难判断能否通过其他方式（例如采用不同策略，或获得更多某种资源）找到有效的解决方案。重要的是，这一区分引出了问题本身 （例如，做出一份美味的沙拉）与如何 解决问题（例如，遵循 Φ）之间的差别。虽然在计算语境中，问题与解决方案之间的区别或许相对清晰（例如，问题与算法解决方案之间的区别），但本文将投入大量篇幅，阐明它们在道德语境中的差异。

3. 伦理学的计算复杂性

伦理学的计算复杂性是什么？首先，我们应当指出，“伦理学”是一个多面且含义模糊的概念，贯穿不同学科的多个分析层次。古往今来，道德哲学家或多或少以系统的方式，试图解决关于何为道德上的“善”与“恶”的问题。在现代，英语分析伦理学通常分为（i）应用伦理学 （判断具体情境中的“善”与“恶”），（ii）规范伦理学 （提出关于“善”与“恶”的标准和原则），以及（iii）元伦理学 （探究道德的意义

[8]然而，需要注意的是，遵循 Φ寻找最优沙拉时，这类性能度量并不适用，因为食材以何种方式排列并不重要。[9]在本文中，“认知”一词将被广义地用于指称能够实现感知、行动、推理和学习等能力的各种信息处理。因此，它不同于元伦理学中的认知主义（即道德语言能够表达可真可假的命题这一观点），也不同于那些强调前额叶活动（例如思维、记忆、判断）的认知观念；后者将前额叶活动与“大脑后部”的感觉处理相对立（Block, 2019年）。

与本质）。然而，伦理学的版图远不止这些划分。从生物学视角来看，它包括合作的进化基础（博弈论对此已有大量研究（Axelrod and Hamilton, 1981; Nowak, 2006））；在这一视角下，道德可以被视为一种适应性解决方案，用以应对利己生物体之间的竞争问题，[10] 从个体细胞（Hummert 等人，2014）到人类（Leben，2018）皆是如此。如果我们还考虑社会、心理和认知维度，问题就会进一步复杂化，例如伦理行为如何与具身智能体的共情、情绪和推理交织在一起，又如何由高度分布式并行的认知系统执行（Newen 等人，2018; FeldmanHall 和 Mobbs，2015）。道德行为并非“固定不变”，而是在经验中发展并得到积极完善的。[11] 伦理学不仅存在于个体层面，也体现于社会和文化层面；它通过实践和制度得以维系并发生转变，通过意识形态和宗教的语言传递，其辩护依据则从神圣权威（例如上帝之言）、维持政治秩序（霍布斯，1651），到促进自由（密尔，1859年）或正义（罗尔斯，1971年）不等。

因此，为了界定我们的研究范围，我们将聚焦于机器伦理学领域中所界定的伦理问题的复杂度。机器伦理学的大多数技术研究都聚焦于规范伦理学，或更具体地说，探讨如何实现规范理论的某些原则或方面，使人工智能体能够依照该理论行事（Cervantes 等人，2020年；Tolmeijer 等人，2020年）。因此，它可以被视为一种应用规范伦理学，因为其主要关注某一特定理论的实际实现，而非讨论应当采用何种理论。在对各种实现方法进行全面调查后，Tolmeijer 等人（2020年）指出，道德机器研究方法可从三个广泛维度加以刻画：伦理理论、实现和技术。第一个维度指所采用的伦理理论，包括义务论等规范框架（Anderson 和 Anderson，2008年；Malle 等人，2017年a；Shim 等人，2017年）、结果主义（Abel 等人，2016年；Armstrong，2015年；Cloos，2005年）、美德伦理学（Stenseke，2021年；Govindarajulu 等人，2019年；Howard 和 Muntean，2017年），以及混合理论（Dehghani 等，2008b；Thornton 等人，2016年）。第二个维度依据 Allen 等人（2005年）提出的分类，考察伦理学如何在系统中实现，例如，是通过“自下而上”的学习过程，还是借助“自上而下”的原则，抑或结合自上而下与自下而上的处理方式。技术维度则考察用于实现该方案的计算技术，其中包括来自各种人工智能范式的方法，例如逻辑推理（如归纳逻辑、演绎逻辑和溯因逻辑）、机器学习（如神经网络、强化学习和进化计算）以及概率方法（如贝叶斯模型和马尔可夫模型）。

[10]或者换句话说，道德的作用是缓解理性的失效（Ullmann-Margalit，2015年）。[11]从 Kohlberg 和 Hersh（1977年）的开创性研究，到 Rest 等人（1999年）的进一步完善，道德心理学已发展成为一个成熟的范式，致力于探究道德与认知发展之间的联系。

3.1. 马尔的三层次分析视角下的伦理学复杂性

基于这些考量，我们该如何界定机器伦理学的计算复杂性？回想上一节的最后一课，我们首先需要找到某种方法，将问题本身与如何解决这些问题区分开来。Marr（1981）提出的一套颇具影响力的框架体现了这一区分。Marr认为，认知系统的信息处理可以从三个彼此不同但互为补充的分析层次加以解释：(i)计算层、(ii) 算法层和(iii)实现层。计算层描述问题本身（例如，输入输出映射）；算法层说明为解决问题而执行的算法过程（例如，策略或启发式方法）；实现层则说明该算法过程如何由系统的物理硬件（例如神经元或电路）实现。这些层次可以用沙拉吧示例来说明：(i) 计算层规定食材的数量、价值函数（例如单种食材或食材组合的美味程度）以及期望的输出（尽可能美味的沙拉）；(ii) 算法层描述问题求解过程（例如 Φ）；(iii) 实现层描述大脑或机器在物理层面上实现问题求解过程的方式。系统的每个层次都可以独立分析。例如，由于同一个计算问题可以通过多种不同的算法程序解决，我们可以独立于算法层来描述认知系统的计算层，进而建立关于该计算系统的计算层次理论。同样，由于算法可以在多种不同的系统中得到物理实现——例如硅或碳构成的系统——我们可以建立一种认知系统的算法层次理论，而无须解释其如何在物理层面上实现。尽管如此，Marr认为，通过自顶向下视角来阐明认知系统的运作更为容易；也就是说，应从系统要解决的问题着手，而不是从它用来解决问题的精确机制着手（Marr，1977年，1981年）。[12] 这是因为高层次解释会对较低层次作出承诺，而这又形成了一个欠定性层级。例如，如果我们推测某个认知系统在计算层面上解决问题 P，那么我们可能无法确定或对它用于计算 P的具体算法持不可知态度。然而，如果我们的推测要具有超越计算层面的解释价值，就必须承诺至少存在某种算法能够计算 P。如果能够证明不存在这样的算法，那么我们的问题就是不可判定的。同样，如果我们认为某个系统使用算法 A来解决 P，就等于承诺存在某种物理系统能够实现 A。

3.1.1. 层次一：计算问题

我们如何将伦理问题纳入 Marr 的框架？更准确地说，规范理论（NT）所提出的伦理问题，其算法层和计算层分别是什么？首先，我们注意到，规范伦理学模糊了 Marr 前两个层次之间的界限。具体而言，其规范性组成部分与其行动指导密切相关，也就是说，通过

[12]需要澄清的是，这里“自上而下”的特定含义——从计算（上层）到算法，再到实现（底层）——不同于《认知心理学》中的常见用法。在后者中，“自下而上”加工从感觉输入开始，而“自上而下”加工则侧重于依据知识、经验和预期来解读输入信息。

回答什么是善（例如，遵守道德义务），它告诉你如何行善（例如，只采取符合道德义务的行动）。[13] 反过来，这又引出了一系列可能的解释，我们将讨论其中三种：

(1) 规范理论作为一般化道德的算法层解决方案——从最一般的意义上说，如果将计算层问题表述为“做道德的事”，我们或许可以把规范理论解释为计算问题“如何普遍地做到道德”的算法层解决方案。这种解释体现了规范理论在人类情境（或至少在关于规范理论的哲学论述）中的普遍性追求：规范理论应当针对是非问题提供适用于一系列具体情形的一般答案或标准。致力于 NT_1 的智能体，只有在其一般行为中遵循 NT_1 ，才算是道德的。[14] 然而，很难看出如何才能以计算复杂性分析所要求的形式化体系，切实地构建如此宽泛的解释；这将需要某种形式的通用算法——例如以价值、原则或准则为基础——为所有可能的道德困境提供解决方案。[15]

(2) *NT* 作为解决具体道德问题的算法层方案 ——一种相似但范围更窄的解释是，NT 提供了算法层策略，可用于解决具体 道德问题。至少乍看之下，这种解释似乎符合“道德困境”一词的日常用法，即一种由两个或更多 NT 之间的冲突引发的决策问题（其中“NT”也可以换成价值观、义务、美德或规范）。例如，我们可以将计算层问题具体化为电车难题，以凸显行动型与结果型 NT 之间的冲突：即使拯救 5 个人需要采取本质上不道德的行动（例如谋杀），这样做在道德上是否正确？然而请注意，此类问题的道德复杂性（或不可判定性）并不在于计算层问题本身，而在于应如何解决算法层方案之间的冲突（例如，通过双重效果原则（Foot，1967年））。无论如何，这种解释仍与以下观点相容：可以根据问题的性质和可用资源，采用不同的 NT 来解决不同的问题。这似乎与实验研究相呼应；研究表明，人类会灵活地根据不同语境采用不同的道德策略（Capraro 和 Rand，2018年；Conway 和 Gawronski，2013年；Greene 等，2008年）。直观而言，面对某个伦理问题 E_1 时，你可能不愿采取某个行动，因为你认为该行为本身是不道德的（根据 NT_1）；而面对另一个伦理问题 E_2 时，没有任何行动本身显得不道德，但有些行动带来的结果似乎比其他行动更可取（根据另一种理论， NT_2）。也就是说，如果 NT_1 与NT_2 之间没有冲突，你只需选择最适合当前计算层问题的一方。按照这种解释，至少原则上，我们可以评估某个 NT 在解决同一个计算层问题时，是否比另一个 NT 在计算上更高效。

[13]较少强调行动的规范理论或许是一个有趣的例外；例如，某些美德伦理学版本强调的是存在而非行动。然而，这并未消解两者之间的区别，只是将其推向了兴盛与使智能体得以兴盛的性格特质之间的模糊地带。此外，美德伦理学无法提供行动指导这一观点也受到过批评；例如，参见 Hursthouse（1999年）关于行动指导的美德伦理学观点。[14]或者至少，智能体以 NT_1 作为评判某一行动是否道德的主要标准。[15]黄金法则或康德的绝对命令（康德，1785）或许是这类通用算法的典型样例，我们将在第5节讨论。

然而，我们能否询问它在道德意义上是否更成功？如果不进一步解决元理论问题，似乎很难给出答案。或许更成问题的是，这种解释似乎认为，伦理问题以某种有意义的方式独立于解决这些问题的各种方式。恰恰相反，算法解决方案（NT）似乎取决于计算问题本身的性质，以及该问题如何通过某种 NT 提供算法解决方案；这些可供性已经嵌入计算层之中。如果某个具体问题只有借助某一种 NT 才可判定或易处理，那么将其视为一个独立的计算层问题似乎更为公正。例如，我们可以设想一个伦理问题空间，其中只包含有关义务的信息，而不包含有关结果的信息；因此，基于义务的 NT 可以判定该问题，而基于结果的 NT 则无法判定。这自然引出了一个范围更窄的解释，也是我们在分析中将主要关注的解释：

(3) 由 *NT* 提出的、作为计算层问题的具体道德问题 ——我们不必将 NT 放在算法层，而可以根据某个特定 NT 的表述来定义具体的计算层问题。反过来，这使我们无需确定算法层实际执行的具体程序：只需假设存在这样一个程序即可。因此，(3) 为机器伦理学家提供了诸多便利，包括：(i) 明确什么是 道德的，或什么行为在道德上是良好的 （由建模理论所规定）；(ii) 为行动指导提供蓝图，帮助进行算法设计和计算方法的选择；以及 (iii) 提供评估性能的手段（例如，合适的义务论智能体会成功遵循道德义务和规则）。重要的是，(3) 的狭窄范围使我们得以忽略困扰着 (1) 和 (2) 的理论问题：与 (3) 不同，(3) 并无普遍性追求（因而可以用于特定情境或领域）；与 (2) 不同，建模的 NT 中相关的行动指导方面已嵌入计算层问题描述之中。也就是说，(3) 容纳了这样一个事实：不同的伦理问题——例如特定环境中提供的信息——会带来不同的伦理行为可能性，而 (2) 并未考虑这一点。或许最重要的是，(3) 允许我们分析伦理问题的算法层，而 (2) 将规范理论视为算法本身，这可能会掩盖对实际如何执行此类程序的分析。

从复杂性分析的角度来看，(3) 的另一项主要优势在于，它构成了伦理计算中不可或缺却复杂度最低的方面，因为 (1) 和 (2) 都预设智能体能够执行 (3) 类型的计算。换言之，由于解释 (1) 是对 (2) 的泛化，而 (2) 又取决于对 (3) 的具体实例化，因此它们构成了一个伦理计算层级（如图1所示）。也就是说，要依据某种规范理论解决广义道德问题（解释 1），

[16]同样，即使我们认为只有一种 NT 是正确的，也不一定意味着我们不能因某种替代理论的计算效率而发现它有用（即使从道德立场来看，我们通常并不喜欢该理论）。[17]例如， NT_1 提供的解决方案可能在计算效率上最高，但也可能违反了 NT_2中的某些原则；然而，对于 NT_2 而言，并不存在高效的解决方案（例如，需要指数时间）。因此，问题是：假设智能体认为 NT_2在道德上优于 NT_1，那么哪种 NT 最成功?

解释	**计算算法** 什么是 the 问题?	什么是所谓的 解决方案?	**实现** 它是如何实现的?
(1) 规范理论作为算法解决方案，用于广义化道德	Do what is 道德（一般行为）	NT_X （应用通常）	心智／机器
（2）规范理论或规范理论（复数）作为 al- 算法解决方案 to 具体 道德问题	具体道德问题 P	NT_X （应用于 P）	心智／机器
（3）特定道德问题 P 由.....提出特定规范理论$_X$	P 由.....界定 NT_X	计算方法	心智／机器

表1：在 Marr的三个层次中，规范理论下伦理问题如何被界定和分析的三种解释。

例如 NT_1，智能体还必须能够将 NT_1 应用于具体道德问题（解释 2）；要在特定情形下（解释 2）应用 NT_1，智能体就必须能够以 NT_1 所设定的方式应用 NT_1 （解释 3）。因此，如果某种具体的 (3) 类计算是不可判定的，那么同一问题的 (1) 类和 (2) 类计算也将不可判定；在特定情形（2）下，该问题对于规范理论而言不可判定，因而在一般情形（1）下也不可判定。表1进一步说明了解释之间的差异。

3.1.2. 第 *2* 级：算法

因此，我们认为，分析伦理学的计算复杂性的一种自然方式，是聚焦于规范理论所提出的问题（计算层），这些问题通过各种计算方法（算法层），由确定性图灵机加以解决（实现层）。当然，这仍然为算法层和实现层的具体运作方式留下了相当大的解释空间。要为一个定义明确的问题找到最有效的算法解决方案，通常是一个经验问题；随着编程技术（例如，通过动态规划将问题分解为更简单的子问题）或启发式方法（例如，利用问题中的规律）不断取得新的进展，相关答案也会持续修订。更重要的是，这还取决于我们将什么视为解决方案。如果我们认为规范理论严格要求系统必须找到某个问题的最优解，那么这就意味着算法层应遵循某种能够保证产生最优解的程序；这就是所谓的 精确算法。[18] 一种不那么严格的解释，是接受与最优解“足够接近”的解决方案；这类解决方案通常被称为

[18]例如，这类似于 Φ按照 CPG 为沙拉问题生成一个最优（尽管难处理）的解决方案。

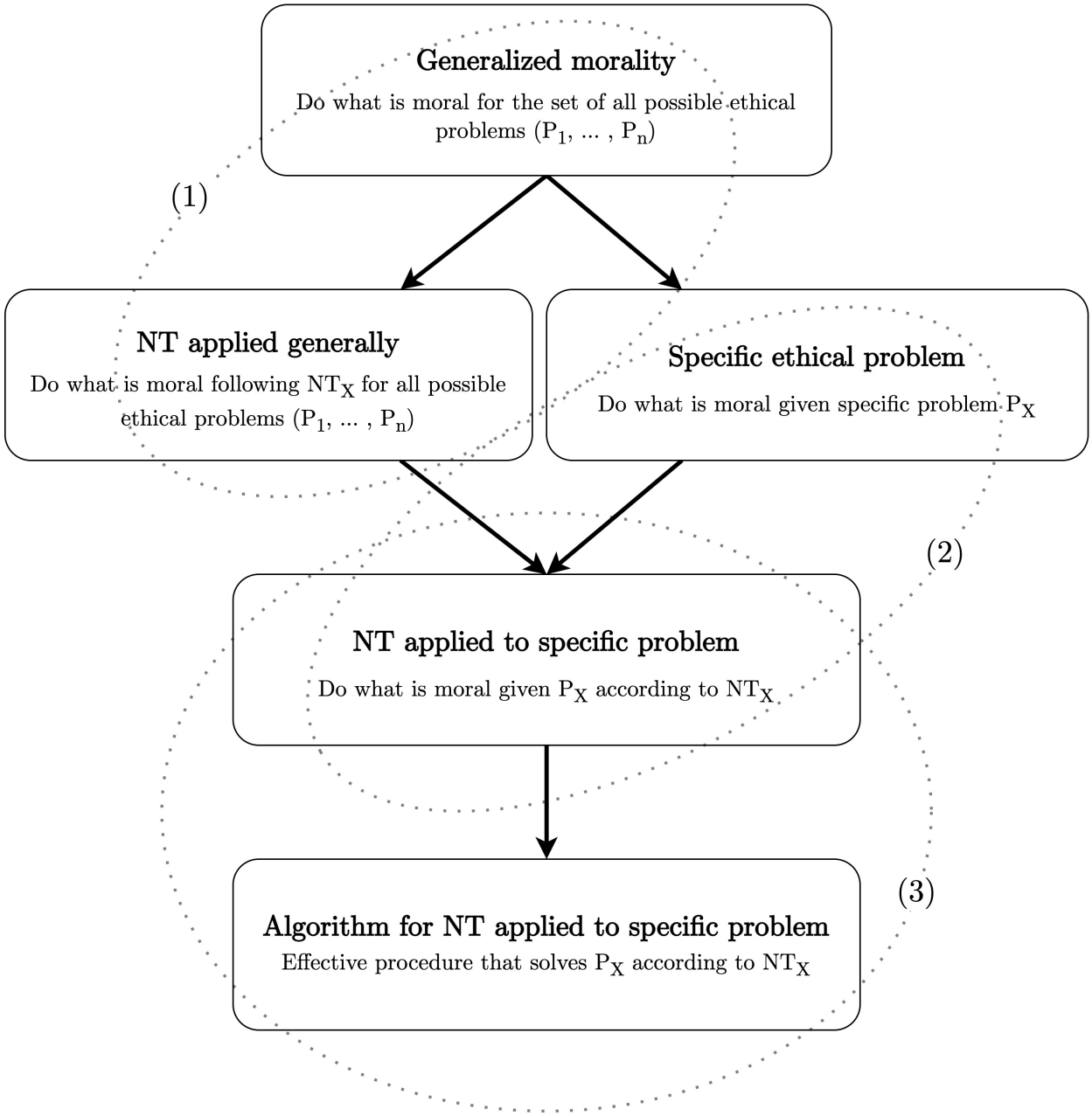

图1：伦理计算层级。箭头表示依赖关系（即，A $\rightarrow$ B 表示 A 的解决方案依赖于 B 的解决方案）。虚线椭圆圈出了三种解释（1）–（3）的计算层级（上方）和算法层级（下方）。

近似算法。虽然近似算法无法保证找到最优解，但它们能够保证所得解决方案与最优解之间的距离在某个固定范围内（即最优解与近似解之比存在一个可证的界限）。近似解与精确解之间的差异，对可处理性而言至关重要，因为许多精确解难处理的问题都可以在多项式时间内得到近似解（Williamson 和 Shmoys，2011年）。

我们主要关注精确解，原因有二且相互关联：（i）规范性理想要求如此（遵循严格的解释）；（ii）这使我们能够聚焦于问题本身，而非算法。第一个理由可由以下考量支持：如果近似解可以接受，我们如何论证某个解决方案与最优解之间的距离处于可接受范围内？需要注意的是，尽管近似能够给出关于距离的可证保证，但这个距离仍可能任意大。[19] 这似乎意味着我们还需要界定何谓可接受的距离，而这一界定可能因情况而异。此外，许多现实世界问题没有可供利用的可识别结构，因此无法通过高效的近似算法求解（Nievergelt 等人，1995）。这自然引出了第二个理由：与近似解相比，精确解更容易进行比较，因为我们无需界定近似解在什么条件下才算足够接近。

当然，道德理论家可能会合理地指出，我们不应关注道德问题的精确解或最优解，而应从“可许可”或“不可许可”的角度理解这些问题。例如，即使某项行动并非最优，它仍可能是可许可的；只要我们避免不可许可之事，道德并不要求我们做出超出可许可范围的任何事情。这一推理进而可能为采用次优近似提供正当性。然而，这会模糊最优性作为数学概念与作为道德概念之间的差异。例如，道德可许可性可以被理解为数学上的最优解；也就是说，将其视为用于评估行为的某个不动点或度量。或者，道德可许可性也可以被理解为对某种固定的道德最优性概念的数学近似。但如此一来，我们又回到了原本想要避免的同一两难境地：在每种情况下，我们都需要说明，在给定的道德可许可性阈值下，某种近似为何可以接受。不过，需要注意的是，这并不排除这样一种可能性：此类近似可以结合特定语境中的可许可性得到论证；而是说，这种分析超出了本文的范围。

3.1.3. 第 *3* 层：实现

在实现层面，我们将采用使用最广泛的计算模型：图灵机（TM）（图灵，1936年）。更具体地说，由于图灵机是一种数学计算模型，它指的是任何能够实现图灵机的物理系统（即图灵完备的系统）。图灵声称，任何可由算法计算的函数都能由图灵机计算。当图灵证明其可计算性概念等同于丘奇（1936年）独立提出的方案时，这一论题便获得了公信力。这构成了丘奇-图灵论题的基础，而该论题后来也被证明

[19]例如，我们可以设想这样一种两难境地：最优解的道德价值为 100，但最佳近似只能得到 50 的价值。

等同于许多其他计算形式（Herken，1995）。简而言之，这意味着任何通用系统（例如计算机或计算机语言）都能模拟其他任何通用系统的计算层面。我们还将假设 P $\neq$ NP（第2节讨论）。与丘奇-图灵论题一样，这是计算机科学家普遍接受的另一项猜想，尽管它仍有待证明。[20]

重要的是，如果我们能够证明某个问题是NP难的，这就意味着我们不能指望找到它的高效解决方案；这里的“高效”指“可由确定性图灵机在多项式时间内求解”（P可处理性）。[21] 因此，如果通过计算方法解决的伦理问题是NP难的，我们就不能指望计算系统高效地解决这些问题，这也会直接影响道德机器的可行性。然而，即使P $\neq$ NP和丘奇-图灵论题几乎获得普遍认可，我们仍须注意与P可处理性的局限及适用性相关的若干问题。例如，如果能够确保输入规模足够小，那么P不可处理性就不是什么大问题（例如，只有5种食材的沙拉吧仅有32种可能的组合）。重要的是，一个问题仅仅因为P不可处理，并不意味着它无法依据其他合理的可处理性概念得到有效求解。事实上，许多NP难问题都可以通过这样的算法求解：其运行时间仅在输入的某些 部分（输入参数）上呈超多项式增长，而相对于整体输入规模则呈多项式增长。[22]反过来说，多项式函数中的大常数，例如 n^{100}，仍属于P可处理，即使它们可能不符合任何直观的“有效”概念。此外，采用其他计算模型也可能显著缩短耗时，例如利用并行化、随机存取存储器或量子计算。尽管不变性论题——以及与之密切相关的扩展丘奇-图灵论题（Kaye等人，2006；Bernstein和Vazirani，1997）——指出，没有任何机器能够比确定性图灵机快超多项式倍，[23] 但未来计算技术进步能否以及在多大程度上证伪这一论题，仍有待观察。[24] 关键在于，尽管P可处理性是从理论和实践角度形式化研究有效计算不可或缺的工具，却不应将其解读为一条一概而论的明确界线，用来划分哪些问题可处理、哪些不可处理。尽管P可处理性对道德机器有直接影响，一个相关但更复杂的问题是，它能否为人类的道德认知提供任何有意义的洞见（我们将在第7节回到这个问题）。

[20]请注意，丘奇-图灵论题并非数学意义上的猜想，而是关于计算本质的一种假设；由于其对有效可计算性的概念是非形式化定义的，因此无法证明。[21]P类 $\neq$ NP意味着，我们不能指望找到NP完全问题，以及可转换为NP完全判定变体的NP难问题的有效解决方案。请注意，即使P类 $=$ NP，许多NP难问题仍然会是难处理的，例如，它们若属于被认为涵盖NP的复杂性类（如PSPACE或EXPTIME），则依然如此。[22]正是这一观察推动了参数化复杂性的发展（Downey和Fellows，2012），以及固定参数可处理问题类（FPT）的提出。另见Fellows（2002）和Niedermeier（2006）。[23]请注意，人们普遍认为，不变性论题适用于并行和串行计算模型，参见例如Frixione（2001）；Parberry等人（1994）；Tsotsos（1990）。[24]例如，未来某台计算机至少能够在多项式时间内解决某些目前尚难处理的问题，这种可能性并非微乎其微。

为了划分问题空间，我们将聚焦于三类道德机器：因果引擎（第4节）、规则遵循者（第5节）和道德学习者（第6节）。主要原因是，机器伦理学中的几乎所有实现都采用了这三种方法之一（Tolmeijer 等人，2020年）。另一个原因是，这三类机器分别对应一种重要的规范框架：结果主义关注预测未来事件（因果引擎），义务论关注遵守规则或履行义务，而美德伦理学则强调学习。[25]为了对伦理问题进行复杂性分析，我们还将假设这些问题可以表述为定义明确的计算问题，也就是说，它们具有明确定义的初始条件和目标（例如，以特定的输入和输出条件表示），并且能够通过常规数据类型（例如布尔值、整数、浮点数）和结构（例如数组、列表、图、树）来表示。尽管这些简化条件可能无法充分体现智能体在现实世界中可能面临的问题之丰富性及其潜在的定义不清，但考虑到现实世界中的伦理问题只要可判定，其信息丰富程度至少不亚于简化后的计算问题，这一假设便有其合理性。用技术术语来说，我们假设定义明确的计算问题构成了现实世界环境中伦理问题信息论本质的合理下界。最后，我们主要关注时间复杂度而非空间复杂度，理由很简单：访问和存储记忆需要耗费时间，这意味着记忆消耗通常有其上限，而该上限由时间消耗决定（Garey and Johnson，1979年）。

4. 结果主义与因果引擎

结果主义是一类将结果置于道德评价核心的规范理论。尽管所有后果主义者都认同结果具有道德重要性，但他们可能对何为良好结果存在分歧；或者说，他们可能对使结果变得良好的因素有不同看法。例如，功利主义——可以说是结果主义理论中影响力最大的分支——主张采取能够使效用最大化的行动。效用可以理解为受影响个体的整体福祉（边沁，1789年；密尔，1861年）、对其偏好的满足（Singer，2011年）、痛苦的减少（Smart，1956年），或其所处国家的福利（Sen，1979年）。结果在道德上为何重要，也存在许多细微差异，例如：意图产生的后果是否重要（而不只是实际后果），后果是否取决于行动主体的视角（即主体相对而非主体中立），间接后果是否重要（而非仅看行动本身造成的直接后果），后果对哪些对象重要（例如，对有限的一部分个体，还是对地球上所有有感知能力的生命），以及后果在多长时间范围内重要（例如，仅看即时结果，还是看永恒的结果）（Sinnott-Armstrong，2021年）。尽管如此，所有形式的结果主义都共同承诺重视未来事件的道德价值。因此，任何承诺践行结果主义的智能体——无论是人工的还是生物的——都必须能够对未来作出预测，才能在实践中遵循该理论的要求。这就是成功的结果主义智能体依赖所谓“因果引擎”的原因；我们用这个术语宽泛地指代支持因果认知的信息处理过程。

[25]当然，正如我们稍后将看到的，在许多情况下，这些理论与类型之间的界限会变得模糊。

请注意，绝大多数生物体以某种方式关注自身行动的后果，因为这会大大提高它们的生存几率。直观而言，因果认知似乎对许多基本能力至关重要，例如避免伤害、解决问题和规划。实验结果表明，年仅八个月的人类儿童就能根据因果关系作出推断（Sobel 和 Kirkham，2006）。这或许意味着，我们的生物本性中深深植根着某种因果推断的前反思能力，反映出许多人认为是神经系统的核心功能的预测处理（Friston，2010年；Hohwy，2013年；Keller 和 Mrsic-Flogel，2018年）。然而，与生物体不同，机器并未通过进化过程发展出因果引擎。相反，人工系统遵循结果主义的能力依赖于计算技术，这些技术通常源自统计建模、贝叶斯建模和马尔可夫建模等方法（Casella 和 Berger，2021年）。机器学习方法也常被视为一种"预测分析"，因为算法会从经验中学习如何作出更准确的预测；例如，监督学习通过人工生成的数据进行，强化学习则通过交互式的试错过程进行。但结果主义并不只是对未来作出预测，还涉及从可能结果的集合中评价哪些结果在道德上优于其他结果。也就是说，即使一个结果主义智能体能够以神一般的准确性和速度预测所有可能行动的结果，也不一定意味着它能以同样的速度轻易判断出哪种结果才是最优结果。

鉴于这些考量，本节将探讨三类一般结果主义问题的计算复杂性：确定最优结果的组合学（4.1）、因果推断（4.2），以及在不同时间范围下动态与部分可观测环境中的决策（4.3）。本节旨在循序渐进地向初学者介绍时间复杂度分析、概率论（贝叶斯网络）和随机方法（马尔可夫决策过程）。

4.1. 结果组合学

在最简化的情形下，我们可以设想：结果主义者面对的问题是，在智能体已经能够确定不同结果为何的前提下，比较这些结果的道德价值。如此一来，我们就可以忽略因果推断本身的复杂性，从而单独考察最优结果评估问题。用复杂性理论的术语来说，我们假设智能体可以访问一个所谓的预言机，它能够通过一次操作就提供有关因果事件的答案。例如，如果智能体询问"如果我执行行动 a会发生什么？"，预言机就会给出类似"行动 a 会产生一个道德价值为 v 的结果"的回答。[26] 这类问题中最简单的计算问题可以形式化如下：

C1 — Optimal outcome following consequentialism

输入： 一个环境，作为可能行动的集合 $E = \{a_1, a_2, ..., a_n\}$ n；以及一个价值函数 v，它为每个行动 $a \in E$ 赋予一个结果值。

[26]为涵盖功利主义的多种版本，我们将不对应该最大化的效用的确切性质作出判断；唯一重要的是，效用能够用数值表示。

输出： 一个行动 $a \in E$，使得 $v(a)$ 在 E 中所有可能的行动里达到最大值。

以下通用穷举搜索算法可以保证找到最优解：

算法 1： 使用因果预言机进行穷举搜索

```
1  h ← 0 // 结果值（默认设为 0）
2  j ← 0 // 动作索引（默认设为 0）
3  对于 a_i = a_1 到 a_n ∈ E，执行
4  |  h_i ← v(a_i) // 调用预言机
5  |  如果 h_i > h 则
6  |  |  h ← h_i // 更新最高价值
7  |  |  j ← i // 更新最高价值的动作索引
8  |  结束如果
9  结束循环
10 返回 j // 返回具有最高结果值的行动索引
```

简言之，该算法先为结果（步骤 1）和动作索引（步骤 2）初始化默认值。然后，它遍历环境中的每个行动（步骤 3），调用预言机（步骤 4），检查该行动的结果是否高于当前最高值（步骤 5）；若是，则更新最高结果值（步骤 6）及其索引（步骤 7）。最后，返回最高结果的索引后终止（步骤 10）。假设每条指令的执行时间相同（1），我们可以按以下方式精确计算该算法解决 C1 所需的机器操作次数：第 1、2 和 10 行只需执行一次（3 次），第 3 至 7 行各需执行 n 次（$5n$ 次），而第 8 和 9 行可以忽略（因为它们是流程控制语句），因此总计为 $3+5n$。用大 O 表示时，这可简化为 O（n）。换言之，时间复杂度随输入规模线性增长（O（n））。重要的是，无论机器执行其他指令的速度有多快，要确保最优性，它都必须向预言机提出至少等于可能行动数量的问题。也就是说，如果有 10 个行动，智能体至少必须调用预言机 10 次。

如果我们允许存在多个价值观，会发生什么？例如，我们可以假设智能体有一组包含两个或更多结果值的集合，需要针对每个动作-结果进行检查（例如快乐、公平性、信任等）。这就得到以下问题：

C2 — Optimal combination of values

输入： 与 C1 相同，此外还为每个 $a \in E$分配一组结果值函数 $V = \{v_1, v_2, ..., v_i\}$。

O输出： 一个行动 a，使得在所有 $v \in V$和 $a \in E$中，$v(a)$ 达到最大值 .

如果我们假设这些价值观之间的相互作用微不足道，也就是说，可以将价值观汇总$v_1(a) + v_2(a) + ... + v_i(a)$，从而得到单一的总价值 $V(a)$（即遵循可加性定律），那么最优动作 a^* 可以形式化地表示为：

$$a^* := \underset{a \in E}{\operatorname{argmax}} V(a) := \{a \in E : \sum_{m=1}^{i} v_m(\acute{a}) \leq \sum_{m=1}^{i} v_m(a) \text{ for all } \acute{a} \in E\} \tag{1}$$

如果智能体需要针对每种价值分别调用预言机，时间复杂度就是 n （行动数量）与 i （价值数量）的乘积，得到 $O(ni)$。如果 i 等于行动数量，那么运行时间相对于 n呈二次增长，但仍属于多项式 $O(n^2)$。[27]

到目前为止，我们只关注单个行动的道德评价。但在现实世界的伦理决策问题中，可能需要执行多个行动。然而，正如沙拉示例所展示的，将行动组合起来的可能性会带来可处理性问题，而这些问题根植于组合结构的排列之中。为了说明这会如何影响结果主义的计算复杂性，[28] 我们将行动计划$\varphi = \{a_1, a_2, ..., a_n\}$ 定义为环境所提供的一个非空且不同的 n 行动 $\varphi \subseteq E$集合，使得 $\varphi \subseteq E$。随后，我们扩展 C1 来描述以下问题：

C3 — Optimal plan of up to two distinct actions

输入： 一个环境，包含一个由 $E = \{a_1, a_2, ..., a_n\}$ 个 n 可能行动构成的集合，以及一个函数 v，该函数为每个行动计划 $\varphi \subseteq E$分配一个结果值。
输出： 一个行动计划 φ，使得在所有 $\varphi \subseteq E$中，$v(\varphi)$达到最大值；没有任何 $a \in \varphi$与其自身相同（即，同一行动不能执行多次），并且 $|\varphi| \leq 2$。

解决 C3 的唯一方法是向预言机进行若干次调用，次数等于可能的行动计划数量（最多包含两个行动）。随着行动数量增加，这个数字将呈三角形增长（$\binom{n(n+1)}{2}$）——也就是正方形的一半。[29] 这种易处理的程序符合行动组合原理（CPO），即行动计划产生的结果值不一定等于其各个个体行为单独执行时所产生结果值的总和。然而，它会违反因果关系的一项基本原则：两个因果事件——行动 x 和行动 y——所产生的结果取决于 x 和 y 发生的先后顺序。我们可以将其称为因果顺序原理（PCO）。为在解决 C3时满足PCO，结果主义者必须额外向预言机进行三角形数量的调用，从而使 n的增长达到二次增长（$n-1$）。不过，按照渐近大 O 记法，无论采用哪种方式解决 C3，其时间复杂度都是 O（n^2），仍远在可处理界限之内。

当我们将 C3类型的问题泛化（例如，考虑 n 个行动）时，动作-结果的计算复杂度就会成为结果主义者面临的问题：

C4 — Optimal plan of up to n distinct actions

输入： 与 C3 相同。**输出：** 一个行动计划 φ，使得在所有 $\varphi \subseteq E$ 中，$v(\varphi)$ 达到最大值，没有任何 $a \in \varphi$ 与自身相同，并且 $|\varphi| \leq n$。

[27]为了通过算法过程展示这一结果，我们只需扩展穷举搜索（算法 1），使其遍历 n 个行动和 i 个价值观，例如，为每个 i额外添加一个**for**循环，或者在遍历行动 1 到 n 的循环中嵌套一个遍历价值观 1 到 i 的循环。[28]关于基于 SAS$^+$ 形式化体系的行动计划，类似分析可参见 Lindner 等人（2020）。[29]正如那个著名故事所说，卡尔·弗里德里希·高斯年幼时，老师让他把 1 到 100 的所有数字相加，他很快就找到了这个数列的求和公式。

遵循 CPO 并解决 $\mathrm{C4}$ 的精确算法的时间复杂度为$O(2^n)$。换言之，对于试图解决 $\mathrm{C4}$ 类型问题的结果主义者而言，不存在多项式时间可解的过程。[30] 更糟糕的是，如果结果主义者还应遵循 PCO，那么精确算法将产生 $O(n!)$ 的阶乘增长（N. J. A. Sloane，2022）。[31] 更广泛地说，众所周知，许多规划任务都是 PSPACE 完全的（Bylander，1991，1994年；Littman 等人，1998年）。[32]

请注意，尽管这种不可处理性在实践中或许并不构成有害问题——例如，对于较小的输入，比如有四种可能的行动，遵循 CPO 求解 $\mathrm{C4}$ 需要调用 15 次，而同时遵循 CPO 和 PCO 则需要调用 64 次—— $\mathrm{C4}$ 但这仍然预设了许多其他重要假设，而这些假设在现实世界情境中未必成立。例如，它假设智能体不能重复执行同一行动，并且在智能体计算解决方案期间，问题空间保持静态。相比之下，现实世界环境可能会在一个仅部分可观测、且随连续时间变化的状态空间中，呈现出潜在无限的可能行动集合（我们将在第 4.3 节回到这一点）。最重要的是，智能体无法调用因果预言机，而必须依赖自身的因果引擎；这就引出了因果推断的复杂度。

4.2. 因果推断

一旦进入不确定性的领域，我们就无法保证任何系统的性能都能达到最优。相反，我们所能期望的最佳结果，是依据我们的“最佳猜测”达到最优，也就是说，依据我们的信念或所知达到最优（贝叶斯最优性）。这也是许多版本的功利主义经过修订、强调最大化“期望”效用而非实际效用的原因之一（Broome，1987年）。它也构成了期望效用假说（冯·诺依曼和摩根斯坦，1947年）的基础；该假说广泛应用于决策理论和经济学，用于在收益未知时对理性选择、偏好和风险偏好（即对风险的接受与厌恶）进行建模。[33] 然而请注意，对概率进行建模的不同方式，为具有道德分量的解释留下了空间；换言之，不同的规范原则可以指导我们如何应对不确定性下的决策。

以下问题展示了这一点，其表示为一个有向无环图（图 2）。该图展示了一个包含三种行动的环境，每种行动都有一定概率产生两种可能结果值中的一种。如果我们只想不考虑风险地最大化期望效用，只需将每个结果值与其相应概率相乘后求和即可——例如， a_1 对应的 $0.2(-2) + 0.8(4)$——然后选择期望效用最高的行动。另一种风险规避程度更高的做法，是选择最坏情况下结果最优最好的行动（这是一种称为“极小极大”的决策规则，即最大化最小收益）。尽管这两种决策程序在运行时间方面差别不大——比如

[30]这类似于后文的沙拉问题 Φ （其中，CPO 被替换为等价的 CPG）。[31]这个数列称为“ $\{1, ..., n\}$的非空子集的排列数”（N. J. A. Sloane，2022），可表示为 $en! - 1$ 的下取整函数，其中 e 表示欧拉数。[32]另请参阅 B¨ackstr¨om 和 Nebel（1995），了解 SAS+ 规划的一些易处理结果。[33]具体而言，该假设指出，智能体通过比较期望效用值来选择不同选项；通常，这些值由效用值 U 与其对应概率 P相乘后求加权和计算得出，即 $\sum U(x_i)P_i$。

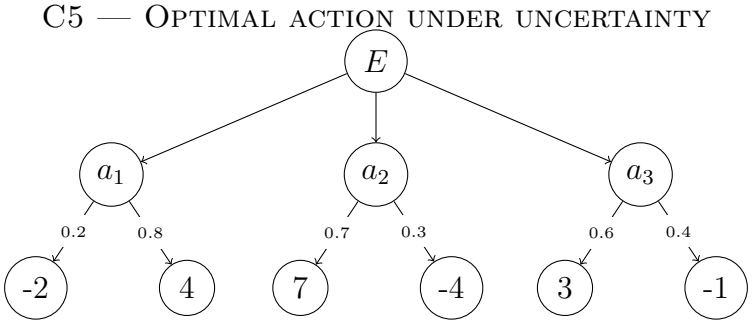

图 2 ：一个有向无环图（DAG），表示不确定性下的决策问题。

我们针对 C2 的解决方案，两者的运行时间都是 $O(no)$，其中 o 指每种行动的结果数量——但它们在道德层面有着显著差异。

然而，与 C1-C4一样， C5 仍假定存在某种贝叶斯预言机，能够根据特定原因（行动）推断特定事件（结果）发生的精确后验概率。更广泛地说，因果推断可以理解为识别什么导致什么的能力，例如："现象 X的原因是什么（或有哪些）？"，"Y会产生什么影响（或哪些影响）？"，以及"X与Y之间存在什么因果关系？"。这些问题都并非易事；事实上，科学研究在很大程度上由回答因果问题所推动，而这需要结合精心收集的数据、大量统计建模技术，以及诸如演绎（根据给定前提进行推导）、归纳（根据观测进行推断）和溯因推理（推断最佳解释）等因果推理能力。

因果推断的一个关键方面是根据先验知识确定后验概率，也就是说，确定在给定证据或信念 B的情况下， A 发生的可能性。在统计建模中，概率的贝叶斯解释为这一挑战提供了一种广受欢迎的应对方式。贝叶斯方法——例如贝叶斯推断、网络和统计学——都基于托马斯·贝叶斯定理，该定理指出，给定 B 时 A 的概率由以下等式给出：$\Pr(A|B) = \frac{\Pr(B|A)}{\Pr(A)\ \Pr(B)}$.[34] 贝叶斯建模已被用于处理~~和建模广泛~~的认知现象，例如运动控制（K¨ording 和 Wolpert，2006）、符号推理（Oaksford 和 Chater，2001）、动物学习（Courville 等，2006）、因果学习与推断（Steyvers 等，2003；Griffiths 和 Tenenbaum，2005）、归纳学习（Tenenbaum 等，2006）、目标推断（Baker 等，2007）和意识（Lau，2007）。

贝叶斯定理最强大且应用最广泛的扩展之一，是构建称为贝叶斯网络（BN）（Pearl，1985年）的图模型；它能够将大量变量及其条件依赖关系简洁地表示为一个单一的DAG（图3）。贝叶斯网络在研究人类因果关系学习方面尤其有用（L Griffiths 等，2008年）。虽然贝叶斯网络的节点表示所关注的贝叶斯变量

[34]应用贝叶斯法则来确定某个因果事件的可能性 A，在给定 B 的情况下，这是一个简单的过程，前提是我们具备：(i) 某个先验概率，即 A （在 B之前） P（A）；(ii) 对 P（B）在没有 A的情况下的概率 P（B）的某种估计 $\neq 0$；以及 (iii) 对 B发生（给定 A）这一相反似然的估计。

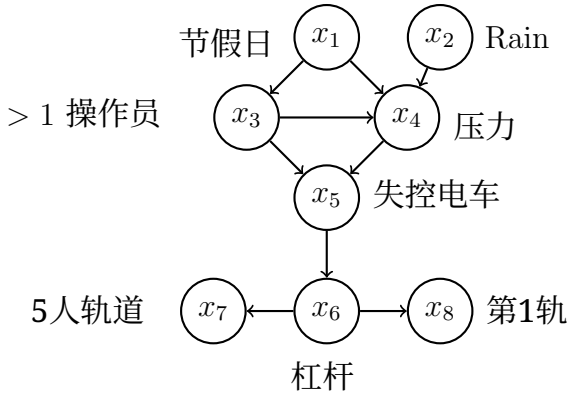

图3：表示贝叶斯电车难题中八个布尔变量之间因果关系的贝叶斯网络（C6）。自20世纪80年代引入以来，贝叶斯网络一直有助于在复杂领域进行基于证据的预测，例如医学诊断（Lucas 等，2000年）和天气预报（Cofino 等，2002年）。

——例如假设、可观测量、事件的发生情况、对象的特征——但连线（或边）表示变量之间的条件依赖关系。每个节点都有一个概率函数，根据其父变量返回一个变量（遵循贝叶斯定理），而未连接的节点彼此条件独立。例如，图3所示的贝叶斯网络描述了八个变量之间的因果关系：是否为公共假日（x_1）、是否降雨（x_2）、当前是否有两名或更多火车站操作员在火车站工作（x_3）、操作员是否承受压力（x_4）、是否存在失控电车（x_5）、是否拉动杠杆（x_6），以及电车是否正驶向与5人（x_7）或1人（x_8）相撞的路线。由于贝叶斯网络支持对任何可能变量子集进行概率推断（即基于有关这些子集的证据），因此它可用于支持网络中任意方向的因果推理过程。利用概率链式法则，[35] 联合概率——即所有可能价值观组合上的概率分布——由下式给出：

$$P(x_1, ..., x_n = \prod_i P(x_i|\psi_i) \tag{2}$$

其中，ψ_i 表示 x_i的父节点的取值。因此，图3中该网络的联合分布为：$P(x_1, ..., x_8) =$

$$P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2, x_3)P(x_5|x_3, x_4)P(x_6|x_5)P(x_7|x_6)P(x_8|x_6) \tag{3}$$

现在，我们可以描述一系列适用于结果主义的贝叶斯推断问题，例如：

C6 — Bayesian Trolley Problems

（a）似然——在关于其父变量拥有完整、部分或没有证据的情况下，P x_7 为真的概率是多少?

[35]链式法则使我们能够仅使用条件概率来计算一组变量的联合分布，即 $P(A \cap B) = P(B|A)P(A)$。

(b) 条件概率——已知证据表明今天是公共假日（$x_1 = true$），x_7 为真的概率是多少？(c) 因果推理——例如，拉动杠杆（$x_6 = true$）会对 x_7 或x_8产生什么影响？(d) 最可能解释（MPE）——已知关于该集合补集的完整证据时，变量集合最可能的配置是什么？(e) 最大后验假设（MAP）——已知关于补集部分证据时，变量集合最可能的配置是什么？[36]

贝叶斯网络非常适合回答这类因果探究，可使用变量消除（Zhang and Poole，1996年）和消息传递（Pearl，2022）等算法进行精确推理，并使用随机抽样（Pearl，1987年）进行近似推断。然而，即使像$\mathrm{C}6$ （a）—（e）这样的问题可以在受限网络中于合理时间内解决，业已证明，大多数贝叶斯网络推理问题在一般情况下都是难处理的。更具体地说，任意图上的精确推理是NP难的（Cooper，1990年），[37] 这意味着，推断某个事件的精确概率（或某个命题表达式为真）至少与NP类中最难的问题一样困难。此外，寻找最大概率解释（MPE）的判定变体是NP完全的（Shimony，1994年），而相关的最大后验假设（部分MAP）是$\mathrm{NP}^{\mathrm{PP}}$完全的（Park and Darwiche，2004）。[38] 或许更耐人寻味的是，这些问题的近似也同样难处理：精确推理（Dagum and Luby，1993年）、MPE（Abdelbar and Hedetniemi，1998年）和部分MAP（Park and Darwiche，2004）的近似均为NP难。[39]

这些结果带来的一个重要启示是，贝叶斯推断的复杂度取决于网络的结构：受约束的图能够限制每个节点的条件依赖关系和父变量的数量，而无约束的图则无法用于有效计算。例如，对于 $x_1 \to x_2 \to x_3 \to x_4 \to x_5$类型的链状图，可以通过逐步消除父变量的消除算法，在多项式时间 $O(nv^2)$内确定 $P(x_5)$的精确推理，其中 n 是变量的数量，v 表示变量可能取值的数量。然而，随着依赖于其他变量的变量数量增加，贝叶斯网络中的推断开始类似于判断任意布尔公式是否可满足（SAT）这一问题：这是已知的第一个NP完全问题。[40]

[36]MAP（有时称为部分或边缘MAP）可视为MPE的泛化，因为它可能需要对未观测或未作解释的变量进行边缘化。此外还要注意，在贝叶斯网络文献中，MPE和MAP都有许多变体：相关澄清请参见（Kwisthout，2011年）。[37]更确切地说，精确推理是#P 完全的（Roth，1996年）；#P是与NP相关联的计数问题类。[38]PP是指由概率图灵机在错误概率小于 $1/2$ 的条件下可判定的问题类（Gill，1977年）；$\mathrm{NP}^{\mathrm{PP}}$则是指可由能够访问PP类问题预言机的非确定性图灵机求解的问题类。[39]关于MPE和MAP问题诸多变体的复杂性结果综述，请参见（Kwisthout，2011年）和（de Campos，2020年）。[40]SAT在推导贝叶斯网络的复杂性结果中发挥了重要作用，这不足为奇。

总之，如果一个结果主义智能体要使用贝叶斯网络解决因果推断问题，我们就不能指望任何易处理的程序能针对任意图给出精确解，甚至近似解。同样的不可处理性结果也困扰着认知科学中的贝叶斯建模，因为贝叶斯规划（K¨ording and Wolpert, 2006）、学习（Kemp and Tenenbaum, 2008）和决策（Vul et al., 2014）都预设了NP难计算。作为一种可能的补救办法，我们可以转而找出能够实现易处理解决方案的约束条件（Kwisthout et al., 2011）。例如，如果排除极端条件概率（即接近0的值），有界方差算法（Dagum and Luby, 1997）就能在多项式时间内生成推断的近似解。同样，研究表明，当底层图的树宽较低，[41] 或者最大概率解释的概率较高时，MPE是易处理的（而当这两个条件同时满足时，部分MAP是易处理的）（Kwisthout, 2011）。然而，这又带来了另一个令人不安的权衡：我们无法保证这些约束条件能反映现实。对于机器而言，这意味着受约束的图可能无法正确建模因果关系。对于人类认知的贝叶斯建模——例如不确定情境下的伦理决策建模——这也意味着我们必须追问，相对于所建模的现象，这些约束是否合理。对于结果主义哲学家而言，这还提出了一个问题：对于遵循结果主义的智能体，因果推断要想成功，必须满足哪些约束条件?

4.3. 动态与部分可观测环境中的决策

到目前为止，我们只研究了将问题的整个状态空间作为输入的问题，例如，将其表示为集合中的元素或图中的节点。但现实世界中的伦理问题提出了一系列额外挑战，可能会限制结果主义者得出最佳结果的能力，包括：（i）部分信息和可观测性；（ii）不断变化的动态且连续的环境；（iii）用于作出决策和执行行动的有限时间范围（例如紧急情况）；（iv）用于评估结果的可能无限长的时间范围；以及（v）可能无限的行动集合（例如，在高于一维的维度中移动）。每项挑战都反映了结果主义者众所周知的认识论问题（Lenman，2000年），例如：在信息永远不可能完备的情况下，作出合理知情的伦理决策所需的最少信息量是多少？或者，应当考虑一项行动结果的时间范围是多长（即，我们需要预测的未来有多长）？[42] 仅时间本身就可能引入混沌不可预测性。正如气象学家兼数学家爱德华·诺顿·洛伦兹的著名论述：一只蝴蝶扇动翅膀，几周后可能导致一场龙卷风。[43]

[41]树宽是图论中的一个概念，非正式地说，可以将其理解为衡量给定图比简单的树（任意两个顶点之间恰有一条路径）“宽”多少；更正式地说，它指图的树分解中最大顶点集合的大小（Bodlaender，1994）。[42]不过请注意，无论是当代还是古典的结果主义者——例如边沁（1789年）、密尔（1861年）和Sidgwick（1907年）——都没有将其原则主张为严格的决策程序，而是视为一项标准。参见 Bales（1971年）。[43]在混沌理论中，“蝴蝶效应”指的是这样一种现象：非线性确定性系统的某个状态发生微小变化，可能会导致后续阶段出现巨大差异（Lorenz，1963）。

尽管如此，人们已经发展出许多数学工具，能够成功应对这些问题。在缺乏解析解或证据的情况下，随机方法让我们能够通过掷骰子（例如蒙特卡洛方法）或将其视为无记忆的事件链（马尔可夫过程）来探索复杂现象。马尔可夫过程是任何满足马尔可夫性质的过程，这意味着某一未来状态的似然仅 取决于当前状态（即"无记忆"）。[44] 从复杂度理论的角度来看，以马尔可夫方式研究过程的吸引力在于，它能够使原本难处理或不可判定的随机建模变得易处理（Vanmarcke，2010年）。蒙特卡洛方法指的是另一类基于重复随机抽样的通用算法，例如，在某一分布或区间内抽取若干伪随机变量。[45] 反过来，这些相当简单的思想已经发展成熟，形成了一整套随机方法，并成功应用于广泛的科学问题，例如统计物理学（Binder 等人，1993年）、工程学（Hajek，2015年）和贝叶斯统计（Gelman 等人，2013年）。

随机方法在自动化决策领域的一个有效应用是强化学习（RL）。强化学习的基本思想很简单：智能体通过与环境交互来学习，并根据收到的奖励更新其行为，例如策略或动作策略。强化学习智能体通常形式化为一个马尔可夫决策过程（MDP），即五元组 $\langle S, A, R, P, \gamma \rangle$，其中：

- S 是一个状态集合（称为状态空间）
- A 是一个行动集合（称为动作空间）
- $R_a(s, s')$ 是智能体执行行动 a，从状态 s 转移到 s' 时获得的奖励
- $P_a(s, s') = Pr(s' \mid s, a)$ 是在智能体执行 $a \in A$ 的条件下，从 $s \in S$ 转移到 $s' \in S$ 的概率
- γ 是折扣因子（$0 \leq \gamma \leq 1$），用于说明智能体偏好长期还是短期奖励。

强化学习智能体的目标是在指定的时间范围内最大化奖励（R）。为此，它需要找到一个策略，即函数 $\pi(s,\ a)$，该函数根据给定的状态 s 决定执行什么行动 a。如果目标是无限期地最大化期望折扣奖励（称为无限时域目标），则最优策略 π^* 可形式化为：

$$\pi^* := \underset{\pi}{\operatorname{argmax}}\, E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \middle| \pi\right] \tag{4}$$

[44]因此，马尔可夫过程构成了连续型和离散随机过程的一大类；例如，泊松过程、维纳过程或布朗运动，以及随机游走，都可以表述为马尔可夫过程的特殊变体。[45]据说，现代蒙特卡洛方法由斯坦尼斯瓦夫·乌拉姆在洛斯阿拉莫斯参与曼哈顿计划期间发明；事实上，该方法对于推导出项目成功所需的仿真至关重要（Haigh 等，2014年）。

强化学习与蒙特卡洛、深度神经网络及其他技术相结合，在 Dota 2 和围棋等复杂游戏环境中取得了超越人类的性能（Berner 等人，2019年；Silver 等人，2018年）；近来，这些技术也在核聚变等离子体控制方面取得了显著进展（Degrave 等人，2022）。更广泛地说，有人认为，奖励足以驱动与自然智能和人工智能相关的各种行为形式，例如学习、知识、感知、语言、社会智能和泛化（Silver 等人，2021年）。鉴于强化学习具有广泛的适用性，有人提出，它为构建理想的伦理人工智能体理论提供了恰当的框架（Abel 等人，2016年），也可用于构建人工德性智能体（Stenseke，2021年）。

重要的是，强化学习能够应对许多可能限制伦理智能体决策的因素：连续动态（Serfozo，1979年）、部分可观测性（Cassandra 等人，1994年），以及针对不同时间范围设定的目标。[46] 强化学习面临的一项关键挑战，是探索与利用之间的权衡。也就是说，当我们无法获得完备信息时，应该依据已有知识作出决定（利用），还是冒险探查可能更好的选项（探索）？ 理论上，探索-利用困境可以通过部分可观测性的概念来解决，它提供了一种对智能体能够直接观测和无法直接观测的内容进行建模的方法。部分可观测马尔可夫决策过程（POMDP）[47] 在马尔可夫决策过程的五元组基础上增加了两个要素：一个观测集合（Ω）和一个条件概率集合（O），用以表示智能体执行 a 且环境转移至隐状态 s' 时，观测到 $\omega \in$Ω的似然，即 $O = Pr(\omega \mid s',\ a)$。简言之，求解 POMDP 的核心在于计算智能体可能所处状态的概率分布（信念状态）；在这些状态中，最优策略通过将行动映射到观测历史来最大化期望奖励。原则上，由于 POMDP 的最优解纳入了行动从信息论视角看所具有的工具价值——以及如何利用信息作出更好的未来决策——因此它为解决探索-利用困境提供了一种方案。

遗憾的是，在无限时域下，为部分可观测马尔可夫决策过程寻找最优解是不可判定的（Madani 等人，2003年）。此外，尽管有限马尔可夫决策过程和有限部分可观测马尔可夫决策过程的解决方案是可判定的，但通常并不易处理。Papadimitriou 和 Tsitsiklis（1987年）的结果表明，有限部分可观测马尔可夫决策过程是 PSPACE 完全的，而 Mundhenk 等人（2000年）的结果证明，各种马尔可夫决策过程问题的复杂度范围从概率对数空间（PL）完备到 EXPSPACE 完全。[48] 强化学习中的其他复杂性结果也显示出类似趋势：到达目标状态在最坏情况下可能需要数量呈状态空间大小指数级增长的行动（Whitehead，1991年）。直观而言，当无法利用有关状态空间的先验知识时，无偏搜索可能导致过度探索。然而，仅凭最坏情况时间复杂度结果，不足以评估强化学习作为不确定性下序贯决策框架的理论可行性，因为这取决于诸多因素，例如任务表示（如状态和行动的数量（Koenig 和 Simmons，1993年））、环境所提供的反馈类型（如可观测性）、

[46]例如，折扣因子 γ 为 0 时只考虑短期奖励，而 $\gamma = 1$ 没有终止状态时则会考虑无限时间内的奖励。[47]用于不确定性下的控制的部分可观测马尔可夫决策过程框架最初由 ˚Astr¨om（1965年）提出，随后 Cassandra 等人（1994年）将其应用于人工智能领域的问题。[48]有关马尔可夫决策过程和部分可观测马尔可夫决策过程复杂度的详细综述，另请参见 Littman（1996年）。

策略类型（如固定策略或依赖历史的策略（Mundhenk 等人，2000年）），或对智能体资源的限制。[49] 与贝叶斯推断的结果类似，尽管不存在强化学习成功的可靠理论保证，但通过简化任务表示（前提是能够实现这种简化表示）、提高奖励的可观测性以及利用先验知识，可以显著提升其实用可行性。强化学习尤其擅长于游戏环境，这并不令人意外，因为这类环境通常能够提供简单的状态空间表示（例如国际象棋和围棋的二维网格）以及清晰可辨的奖励（例如 Dota 或星际争霸）。

然而，强化学习还存在其他问题，可能会妨碍其在后果主义中的应用。例如，即使智能体找到了能在其他智能体所处的环境中最大化自身效用的动作策略，其他智能体的策略或偏好的效用也可能导致冲突。这类博弈论考量颇具挑战性，尤其是在与部分可观测性以及有关其他智能体策略和目标的不完全信息相结合时（我们将在第5节第1小节回到这一问题）。其他问题涉及样本复杂度这一概念，即学习算法学习目标函数（或学习到与最优函数误差在一定范围内的函数）所需的训练样本数量。不过，正如我们将在第6节讨论的，样本复杂度不仅受训练数据分布可能任意“糟糕”这一存在性问题的困扰，还引出了有关归纳的更深层哲学问题。

或许最关键的是，强化学习——以及广义上的随机方法——都预设了试错过程。虽然这在模拟游戏中可能不是什么大问题，但对于未必能提供同等随机探索条件的现实世界环境而言，这却是一项挑战；尤其是某些行动可能造成灾难性后果时。此外，鉴于这一过程具有随机性，强化学习智能体可能会找到一种提高其激励奖励的方式，而这种方式与人类设计者的初衷相悖（在安全人工智能研究中，这被称为“奖励黑客”（Amodei 等人，2016））。[50]

类似地，多臂老虎机问题催生了大量研究，探讨不同条件下的探索—利用困境（Slivkins，2019）。最基本的形式会提出这样的问题：给定 n 个可能的行动（臂），每个行动都会产生一个从固定（但先验未知）分布中随机抽取的奖励，如何才能最大化期望收益？许多版本的多臂老虎机不以奖励最大化为目标，而是从遗憾最小化的角度研究学习问题；遗憾以实际采取的行动与最优动作之间的差异来衡量（例如，利用事后信息进行比较）。目标是找到能够在探索与利用之间取得平衡，同时将遗憾降至最低的策略。其变体包括在不完全信息下进行遗憾最小化（Zinkevich 等人，2007年）、智能体会收到与奖励相关的上下文信息的上下文老虎机（Bouneffouf 和 Rish，2019；Langford 和 Zhang，2007年）、识别最佳臂的问题（Bubeck 等人，2013年），以及寻找均值高于某个阈值的臂（Locatelli 等人，2016年）。对多臂老虎机解决方案的研究通常基于以下两种假设之一：（i）随机——每个臂的奖励分布未知但固定；或（ii）对抗性——奖励由计算资源不受限制的对手选择（Auer

[49]例如，参见 Chatterjee 等人（2016年），了解关于有限记忆策略下部分可观测马尔可夫决策过程可判定解的最新综述。[50]另请参见 García 和 Fernández（2015年）关于安全强化学习文献的综述。

等人，1995）（例如，在被操纵的赌场赌博）。在最一般的形式中，两种假设都被放宽，由此产生了无休止多臂老虎机问题（Whittle，1988年）。这意味着，即使臂未被操作，其收益也会随时间变化。举例来说，设想你是某幼儿园的园长，有 n 名儿童和 m 位保姆，而且 $m < n$。由于儿童人数多于保姆人数，任务就是分配保姆的注意力，以尽量减少孩子们的恶作剧。当有保姆看护某个孩子时，就能获得有关其位置、活动和情绪的信息；若无人看护，则会失去这些信息，而孩子可能会捣乱（从字面意义上说，他们会躁动不安）。[51] 虽然针对老虎机问题的不同变体已有许多易处理的解决方案，但已证明，即使是近似求解无休止多臂老虎机也难以处理。Papadimitriou 和 Tsitsiklis（1994年）给出的证明表明，对于n 个臂，且无论臂处于无人看护还是被主动操作状态，其转移都是确定性的（也就是说，被看护臂和无人看护臂的所有转移值都为0或1），寻找最优策略是 PSPACE-困难的。[52] 此外，由于该证明还表明，判定最优奖励是否非零也是 PSPACE-困难的，因此排除了近似解。[53]

当然，像无休止多臂老虎机这样的问题没有有效解决方案，这并不令人意外；但这表明，除非问题空间（环境）的性质本身允许利用，否则就不存在任何算法方法，能够确保在不确定性下的序贯决策中取得最优性能（或将遗憾降至最低）。无休止多臂老虎机问题的不可处理性结果，优雅地说明了在不断变化的世界中进行决策所面临的这一点。更一般地说，尽管马尔可夫链和蒙特卡洛掷骰可以帮助建模并挖掘复杂空间中的统计倾向，但这类方法要成功，前提是这些倾向确实存在。然而，这或许更多地反映了现实世界动态过程的复杂度，而非智能体的计算限制。正如Hofstadter所指出的：深蓝战胜Garri Kasparov，更多地说明了国际象棋的特性，而非人类智能（Hofstadter，2002年）。Bro˙zek和Janik（2019年）最近也针对道德理论作出了类似评论：“机器可能比任何康德式人 或边沁式人 都更像智人，这一事实对人类道德几乎说明不了什么，却充分说明了哲学关于道德能动性的构想具有理想化性质”（第103页）。不过，本节讨论的复杂性结果意味着更深一层的结论：尽管我们或许会预期人工智能方法在一系列与伦理决策相关的任务中表现优于人类，但它们也受世界复杂度的限制，这不可避免地制约着构建完美道德机器的一切尝试。

总而言之，我们探讨了三种复杂度来源，它们给计算后果主义智能体带来了可处理性问题。这可能意味着，任何受多项式图灵可处理性限制的计算主体，在实践中都无法遵循规范性理想的指引。由此推论，结果主义或许更适合作为一种理论理想，而非能够指导伦理决策的可行决策策略。但如果道德理论探讨无法指导实践中的道德决策，那么它又有何意义？从更务实的角度来看，我们可以考察那些限制因素——例如，在

[51]其他几个直观的样例请参见 Whittle（1988年）。[52]请注意，PSPACE困难性意味着比 NP难更强的性质，因为无论 P = NP 是否成立，PSPACE-困难问题都仍然难处理。[53]不过，如果放宽问题，允许采用线性规划，就有可能获得近似保证，即平均而言每个时间步选择一个臂（Guha等人，2010年）。

可能的行动、行动组合、条件概率、时间范围、任务表征和近似的空间中——在这些限制条件下，结果主义决策会变得易于处理；并确定在每种情况下，这些决策与最优决策有多接近；或者与某个道德价值不动点有多接近。

5. 义务论与规则遵循者

结果主义以结果为核心，而义务论伦理学关注行动本身：某项行动是否道德，取决于它是否遵守一套道德规则、义务或职责。但一条规则最初为何正当？又如何确保对某条规则的特定解释经得起其正当性所依据的原则的检验？根据神命论，道德规则的正当性与普遍有效性源于上帝的权威。基督教十诫为此类规则提供了典范，例如上帝在西奈山赐予摩西的“不可杀人”或“不可偷盗”。相较之下，在伊曼努尔·康德的义务论伦理学中，理性存在者凭借自身的自主意志受道德法则约束，而道德义务的根本原则则体现在绝对命令中：“只按照你同时能够意愿其成为普遍法则的准则行动”（康德，1785）。这意味着，作为理性存在者，人们有义务只按照理性代理组成的共同体会接受为法律的准则行事。

需要强调的是，规则及规则体系早已深深植根于大多数人类社会之中；它们由社会制度生成并以法律的形式得到执行。事实上，道德与法律之间存在复杂且互补的关系，因为二者都是旨在规范人类行为的规范体系，例如促进社会和谐与社群稳定。一方面，法律可以弥补道德在功能上的脆弱性，因为后者缺乏执行自身规范要求的机制。另一方面，在法律难以适用的情况下，道德可以通过责任、团结和公平性等观念，协调社会期望。此外，许多法律思想家认为，法律若要成功履行规范行为的功能，就必须与其所规制对象的道德规范和道德情感产生共鸣。[54]

鉴于义务论以规则为基础，同时人们普遍认为机器本质上是自动遵循规则的系统，因此有人可能会得出结论：义务论为道德机器提供了绝佳的方案。毕竟，义务论规则与机器代码中无处不在的条件语句优雅地相对应，例如：“若输入 $X \rightarrow$ ，则执行行动Y”。在大众文化中，这一观点最著名的探索（及其问题化）体现在艾萨克·阿西莫夫的小说《机器人三定律》中（Asimov，1942年）。计算义务论的吸引力也充分反映在机器伦理学文献中；事实上，Tolmeijer 等人（2020年）的调查显示，机器伦理学领域50项实现中有22项纳入了义务论伦理学的某些要素。这种吸引力部分源于一种普遍看法：从计算角度而言，义务论的复杂度低于其他理论（Brundage，2014年；Wiegel

[54]当然，这一问题在法律实证主义者与非实证主义者之间引发了分歧；前者认为，即使法律在道德上不公正，也可以是有效的；而后者认为，只有与道德规范相容的法律才是有效的（Moka-Mubelo,2017年）。

和 van den Berg，2009；Tolmeijer 等人，2020年）。这一观点涉及技术、心理和哲学三个层面：

（1）从技术角度来看，相较于计算后果（例如，考虑第4节讨论的复杂度），遵循硬编码规则显然更容易。例如，在对行动计划进行道德评价的研究中，Lindner 等人（2020）发现，根据行为义务论和目标义务论判断某个计划是否在道德上可允许，可以在线性时间内完成。相比之下，他们还发现，按同一度量衡量，功利主义是 PSPACE 完全的，而基于避免伤害的原则则是 co-NP 完全的。

（2）其次，Kahneman（2011）关于决策心理学的研究催生了一种颇具影响力的理论，该理论提出，人类推理包含两个截然不同的方面："系统1"快速而直觉化，"系统2"则缓慢而审慎。在道德心理学领域，这促成了道德认知双加工理论的发展（Greene，2007）；该理论认为，道德判断既依赖意识控制过程（通常对应功利主义判断），也依赖自动化的情绪过程（通常对应义务论判断）。以该理论为基础的实证研究发现，认知负荷增加（通过施加一项需要额外控制的任务）会延长功利主义判断的反应时间，却不会延长非功利主义判断的反应时间（Greene 等人，2008）。另一项研究表明，认知负荷可能提高作出义务论判断的频率；如果受试者被提醒自身终有一死，他们作出功利主义回应的可能性则会降低（Tr´emoli`ere 等人，2012）。尽管这并不能证明，作为一种规范理论，义务论本身在计算上比功利主义更高效，但它表明，在人类道德判断中，与自动化过程（可被描述为具有义务论判断特征）相比，意识控制过程需要更多认知资源，也更容易受到认知负荷操控的影响。

(3) 规则能够减轻道德判断的认知负担这一观点，也广泛存在于道德理论中；尤其是作为一种应对经典功利主义所面临挑战、从而挽救结果主义作为决策程序的做法（Brandt，1979年）。例如，现代版本的结果主义区分了会产生最大善的行动（行为功利主义），以及如果得到遵守就会产生最大善的规则（规则功利主义）。前者若被用作决策程序，就会对智能体提出不切实际的要求——例如容易受到偏差影响、缺乏关于行动后果的完备信息，或缺乏作出正确判断的时间——因此，许多后果主义者会采用某种形式的规则功利主义作为决策程序（Hooker，2016年）。[55]

然而，在机器伦理学文献中，义务论——以及更广义的道德规则遵循——的某些计算层面相对受到忽视。机器执行"如果输入 $X \rightarrow$，则采取行动 Y"这一类型的指令或许看似简单，但它们建立在以下条件之上：

[55]举一个直观的例子：我们在红灯前停车，是因为我们可能相信这条特定的交通规则能够带来最大的总体善（例如，对社会而言），而不是因为在这一特定情境下，该行动本身能够带来最大的总体善。需要注意的是，这并不一定意味着行为功利主义与规则功利主义相互冲突：行为功利主义仍然可以作为一种标准，用于评估规则功利主义所规定的遵守规则的后果。

(a) X 确实是 X，例如，智能体对某个输入（例如，某种情境）具有某些性质这一点有适当的理解（或者说，X 的“语义奠基”（Harnad, 1990年）），并且

(b) 在 X 的每一种可能情形下，执行 Y 都是对 X 的恰当回应（或者说，Y 本身就是善的）。

如果这两个条件都得到满足，义务论智能体就能够在道德上采取行动，常数时间 $O(1)$，前提是我们为每一种可能的输入都存储了相应的行动规则（但这会增加对空间复杂度的要求）。[56]

然而在实践中，条件 (a) 和 (b) 极难满足：(a) 假定对任何可能的状态空间都具备完全知识；(b) 则依赖于某种普遍能力，即理解某个行动是对特定状态的恰当回应。[57] 在此背景下，本节将探讨义务论的复杂度，讨论规则的普遍性（5.1）、逻辑与语义学的复杂度（5.2），以及结果主义与义务论混合理论的可能性（5.1.3）。此外，第 5.1.2 节将基于算法博弈论，探讨策略动态的复杂度。我们将论证，规则的道德力量在于其普遍适用性、普遍辩护和计算简洁性。然而，要在运行时实现计算效率，前提是人们已经收集了实践中使这些普遍性得以成立所需的大量知识。

5.1. 规则的普遍性

道德规则的普遍性可以从几个方面来理解。我们首先区分两个重要维度，即适用和辩护。前者指的是一种普遍能力：给定任何可能的输入，都能依据一套规则判断某个行动是否恰当。换言之，就是普遍地判断某个行动是否遵循了其所依据的规则、原则或义务。智能体若要普遍地遵守“不得伤害他人”，就意味着它从不以直接伤害另一智能体的方式行动。对于人类而言，这类规则似乎很简单，但从计算视角来看，要处理它们则更为困难。例如，显然，这类规则预设了复杂的能力，以判断行动（或者更可能是行动链）是否确实遵守规则：例如，知道什么是另一智能体；了解动态环境中可能导致伤害其他行动者的可能行动（及行动序列）的集合；等等。在规则适用方面，情境会造成极大差异：自动驾驶车辆若高速撞上他人，很可能会造成伤害（也就是说，“不得伤害他人”很容易转化为避免碰撞），而社交机器人在教室中可能对可能造成伤害的行动和原因的无限集合浑然不觉。

[56]例如，我们可以设想，每个输入都是一个索引，能立即执行行动列表中的某个行动。另一种做法是，我们可以用道德义务来进行义务论行动评估，例如通过约束满足来评估：给定输入 X，我们只需遍历行动列表，并检查它们是否违反了任何道德义务（这类似于 4.1 中的问题 C2）。[57]正如我们稍后将讨论的，这还可能涉及理解某个行动是否符合其赖以成立的规范原则。另一种满足 (b) 的方法是要求某个行动Y 本质上是好的，不受具体情境影响（或者要求某些输入—行动配对本质上是好的）。作为一项思想实验，我们可以设想一台机器，其中硬编码的全都是本质上好的行动，并让它不断重复执行一系列好的行动。既然这些行动事实上本质上都是好的，那么这台机器就是一台完美的道德机器。这种做法似乎适用于烤面包机这类简单机器。但在动态与部分可观测环境中，这似乎预设了近乎上帝的全知。

如果智能体以高速与他人发生碰撞，就会伤害他们（即“不要伤害他人”很容易转化为避免碰撞），而教室中的社交机器人可能不会意识到可能导致伤害的潜在无限集合的行动和原因。

第二个维度是规则本身是否具有普遍的正当性。对于神命论者而言，规则可以神圣权威为依据而获得普遍正当性。在机器伦理学中，这一观点启发了神命逻辑的发展：人类输入会被解释为对机器发出的神命（Bringsjord and Taylor，2012）。对于源自康德思想的契约主义者而言，原则要获得正当性，必须由所有人共同认可（Rawls，1980），或者不存在任何人能够合理地拒绝这些原则（Scanlon，2000）。契约主义强调智能体的理性；按照康德的道德理性主义，这要求我们尊重他人，也就是说，原则必须对每个人都具有正当性。与之相反，源自霍布斯思想的契约论者——例如 Gauthier（1987）和 Narveson（2001）——强调智能体的自身利益，也就是说，道德规则应当使共同利益最大化。因此，对于契约论而言，遵守道德规范或规则的正当性依据在于智能体自身的利益；当智能体认识到，实现自身利益最大化的最佳方式是与他人合作时，利他主义便由此产生。

5.1.1. 法律与自由主义语境中的人类规则遵循

分析源自这两种一般性概念的规则的计算复杂度，第一步是考察人类实践中的规则遵循。一个值得注意的现象是，立法实践受到不知法律不免责 （“不知法律不免责”）这一原则的指引，其含义是法律应当易于理解和遵守。否则，被告就可能利用故意视而不见的策略。这也意味着，人类法律的制定会考虑人类能力：如果法律的制定方式使任何人都无法遵守，那就违背了法律的宗旨。然而，区分道德规则与法律规则之处在于，后者有司法管辖权立法机制作为支撑，并且可以由国家（例如警察）强制执行。重要的是，如果难以判断某项法律在特定情境中应如何适用，我们就会依靠专家（例如律师、法官以及各级法院的顾问）来解释法律，确保其得到公正适用，并可能为未来的适用创造新的实践。事实上，许多立法机构有意将法律表述得含糊，以便在新情境出现时不断赋予其意义。[58] 因此，在这类情况下，法律的规范性内容主要由其解释赋予，而非法律本身。[59] 反过来，如果法律的适用未能体现其所适用对象的道德情感，就可能构成法律失效：被告也可能有正当理由选择对裁决提出上诉。关键在于，尽管在人类情境中，守法意义上的规则遵循似乎相对简单，但只有依托相当复杂的既有机制，才能确保法律得到成功实现。因此，很难理解遵循人类法律所涉及的计算复杂度，因为这已经预设了构成人类特征的认知能力（或某些行为标准）

[58]例如，欧盟法院。[59]也就是说，法院提供保障，确保某一具体适用符合其规范性依据所依托的原则。Stenseke（2023）对这一论证作了更为详尽的阐述。

法律人格，同时将解释和辩护的过程交由法律从业者负责。

另一个与第一点相关的观察是，法律通常以肯定性或否定性规则的形式表述，即“做这件事”或“不要做那件事”。从人类的角度看，否定性规则至少乍看之下比肯定性规则简单：记住自己不该做什么，比记住自己该做什么更容易（换言之，这降低了空间复杂度）。尤其是在西方自由主义社会的背景下，这一点可以体现在“伤害原则”中，例如法国《style id='3'>人权和公民权宣言 </style》（1789年）中的表述：“自由就是能够做任何不伤害他人的事情”（Johnson，1990年）；又如密尔在《style id='5'>论自由</style》中所说：“[.....] 对于一个文明社会中的任何成员，违背其意愿行使权力的唯一正当目的，是防止其伤害他人”（密尔，1859年）。然而，从计算视角来看，这类原则节省记忆的特点，只有在自由主义社会背景下，以自主公民及其可能采取的无限集合行动为前提时才有意义。举例来说，对于一台机器，遵循“不要伤害他人”的最佳方式或许就是完全不采取行动：遵守这一原则的复杂度与它能够采取的行动集合成正比。

5.1.2. 通用规则及其辩护

遵循法律规则与以神命论为基础的道德规则，其共同点在于，行动主体不一定要理解规则背后的道德理由，就能做出道德行为；智能体只需遵循规则，而不必反思该规则在道德上是否站得住脚（因为上帝或法律就是这样规定的）。当然，对于正当规则应当是什么，可能存在巨大的分歧。相比之下，在康德伦理学、契约论者和契约主义者看来，辩护可能是行动本身的一部分，因为只有在受影响的参与者（或契约所适用的道德共同体成员）接受某一行为时，该行为才具有道德性。这一点在那些力求既普遍适用又普遍获得辩护的规则中体现得淋漓尽致，例如康德的绝对命令（CI）或黄金法则（GR）：“以你希望别人对待你的方式对待别人”。

值得注意的是，涉及辩护的通用规则也可以进行复杂性分析，前提是规则本身为判断某项行动在道德上是否正当提供依据。举例来说，根据 GR，对于智能体 A 而言，问题“在输入 Y 的情况下，行动 X 是否道德上可允许？”是可判定的，前提是可以判定：如果另一智能体 B 在输入 Y 的情况下实施行动 X，A 自己是否会接受。比如，如果在给定 Y 的情况下，X 会造成伤害，而 A 不希望别人伤害自己，那么答案显然是否定的。当然，这种分析忽略了确定任意可能输入、行动与结果之间事实映射所涉及的复杂度，但它给出了一种可以加以分析的一般结构：

(GR1) 基于自身偏好的黄金法则[60] — 例如，黄金法则可能涉及行动集合以及某个行动中的智能体的个人偏好 A，其含义是 A

[60]这里，我们对“偏好”的确切性质持不可知立场——它也可以指义务或责任——以涵盖义务论的多种版本；唯一重要的是，它可以表示为数值或布尔值。

必须检查个体行为是否符合她的个人偏好：她会根据自己希望他人如何对待自己来对待他人（即，如果某项行动由他人实施，会如何影响她自己的偏好）。假设智能体能够快速检查行动对其偏好的影响，那么判断任何行动在道德上是否可允许的最坏情况时间复杂度为 $O(np)$，其中 n 是可能的行动数量，p 是偏好数量。[61]

作为一条通用规则，GR1 凭借其计算效率极具吸引力：只需检查你 是否会基于自己的偏好接受某项行动。乍看之下，它也能涵盖自然世界中利己行动者普遍共有的许多偏好，例如增加快乐、减少痛苦。因此，它也有助于防止诸如直接造成伤害之类的行动。然而，GR1 的主要问题在于，它没有考虑智能体之间偏好的差异，也没有考虑行动对不同智能体的偏好会产生何种不同影响。因此，它容易遭遇反例：例如，法官不应把罪犯送进监狱，因为法官自己也不想进监狱。[62] 作为补救措施，Reinikainen（2005）提出，GR 的普遍适用性必须经受“公开性检验”，这意味着一项行动必须“从所有受影响者的想象视角来看都可以接受”（第155页）。由此引出了第二种表述：

(GR2) 基于他人偏好的黄金法则 ——假设行动主体完全了解行动如何影响他人的偏好，且每个智能体拥有相同数量的偏好，那么判断任意行动在道德上是否可允许的最坏情况时间复杂度为 $O(npo)$，其中 o 指其他行动者的数量。[63]

比较 GR1 和 GR2 的表述后可以清楚地看到，关键并不是时间复杂度略有增加，而是 GR2 要发挥作用所需的知识要求有所提高。即使在规模较小的智能体群体中，要完全了解行动如何影响他人的偏好也可能很困难。然而，GR2 所缺失的是，它未能捕捉到黄金法则大多数表述中至关重要的递归特征。重点不在于智能体只需将他人的偏好纳入考量；智能体还应考虑他们希望他人如何将自己的偏好纳入考量。这甚至未必涉及具体的行动或偏好本身，而是指智能体通常应以一种他们希望他人通常对待自己的方式对待他人。Stace（1937年）、Wattles（1996年）和 Singer（2002年）等哲学家探讨了 GR 的这一特征；他们都指出，针对黄金法则的反对意见只对黄金法则的具体适用方式有力，但若我们将黄金法则理解为智能体的一般行为，则这些反对意见并不成立。遵循 Wattles（1996年）的观点，这可以涉及个人应用黄金法则的方法：

(GR3)将黄金法则应用于你的一般行为，并以你希望他人将黄金法则应用于其一般行为的方式来应用。

[61]请注意，$O(np)$也是寻找最令人满意的行动（例如，最大化偏好的行动）的紧界，这类似于 4.1 中的 C2。[62]这一反例由康德（1785 年，脚注 12）提出，广为人知。[63]例如，如果有两个可能的行动、五种偏好和 80 亿个智能体，那么要确定这两个行动中的任何一个是否满足世界群体在 2022 年的偏好，将需要进行 800 亿次计算。

这一表述具有自我修正性，因为它会考虑智能体 A希望他人在对待 A时予以考虑的潜在无限种方式；例如，基于偏好、义务、对因果结果的敏感性、正直或尊重。它也体现了道德理性主义计划的某些方面，即确定道德的普遍先验原则（例如康德的绝对命令）。尽管如此，仍有两个主要问题会扭曲对 GR3 等规则的复杂性分析。第一个问题是，这种分析可能预设智能体的认知能力存在极大差异，预设这些能力使智能体一般行为的哪些特征成为可能，也预设智能体之间共享哪些认知能力和一般行为特征。以绝对命令为例，它假定了一种理想化的自主性，即智能体依照自己施加于自身的规则行动（不受任何外部影响）。同样，正如我们在讨论“伤害原则”时所指出的，指导自身一般行为的复杂度与一个人能够采取的行动空间成正比。在自由民主国家的语境中，它可能指一组大体上由“通常具备行为能力的”成年人士共享的能力（例如，有意图、有理解地行动，且不受外部因素的控制性影响）。在这一语境中，对 G3 的复杂性分析因而预设我们拥有对成年人的计算规范。

第二个问题是博弈论层面的：对于其他智能体会有什么期待，人们之间可能存在很大差异。在这里，契约论者（遵循霍布斯）与契约主义者（遵循卢梭）之间的冲突便显现出来，因为他们对“原初状态”及其与道德的关系有着不同理解。前者假定人性主要受自身利益驱动，因此道德成了一个合作问题；只有当道德（合作）能增进共同利益时，做一个有道德的人才是理性的。后者则持有更乐观的看法，因为其出发点是相互尊重：道德是从一种承认每个理性自主行动者具有同等道德重要性的立场出发，达成有约束力的协议后所产生的结果。这种冲突可能导致计算资源需求出现显著差异。例如，如果你完全自私，并且认为其他人也都完全自私，那么实现个人的黄金法则就相对简单：你不会帮助有需要的人，因为别人也不会帮助你。相反，如果你相信每个自主存在者都具有同等的道德重要性，那么你可能会帮助他人，这并非出于共同利益（例如直接互惠或间接互惠），而是因为经过推理，你得出结论：自由平等的公民就应该帮助他人。

5.1.3. 策略博弈中的道德行为

尽管很难想象对一般规则进行复杂性分析，能够摆脱一般行为与行为预期的极端差异，但博弈论的洞见可以揭示其中一些方面。1950年，数学家约翰·纳什著名地证明，对于每个有限的 n人博弈，至少存在一个固定策略组合——纳什均衡（NE）——即没有任何博弈参与者能够通过改变策略获益（Nash 等，1950年）。在随后的七十年里，博弈论及其诸多扩展已成为生物学、社会科学和经济学中进行数学建模的标准工具（Smith 和 Price，1973；Nowak，2006；Holt 和 Roth，2004）。在哲学领域，当混合动机博弈转变为协调博弈时，博弈论对社会规范理论的建构发挥了重要作用（Ullmann-Margalit，2015年；Bicchieri，2005年），也有助于契约论者

		玩家 2	
		$Cooperate$	$Defect$
玩家 1	$Cooperate$	$(R,\ R)$	$(S,\ T)$
	$Defect$	$(T,\ S)$	$(P,\ P)$

表 2：囚徒困境的经典收益矩阵，其中 $T > R > P > S$.

pro致力于从自身利益推导出道德（Gauthier，1987年；Skyrms，2004).

从博弈论视角来看，一个值得探讨的问题是：智能体需要具备哪些计算资源，才能决定如何行动？以囚徒困境（表 2）为典型案例，最高效的做法是采用以下两种纯策略之一：

(S1) 纯背叛——始终采取自私的行动，不管不采取这种行动是否能让自身收益（自身利益）更高。

(S2) 纯粹合作——始终采取利他行为，不管自己可能因此蒙受多大损失。

假设智能体在每种相关情境下都能判断何为合作或背叛，那么 (S1) 和 (S2) 的效率相同， O(1)。从某种意义上说，(S1) 可视为对霍布斯自然状态下行为的一种朴素解释（ Hobbes，1651），而 (S2) 则代表康德假想的目的王国中的行为（Kant，1785）。在前者中，智能体的生活“肮脏、野蛮而短暂”，缺乏促成共同繁荣的机制。[64] 在后者中，智能体彼此视为目的（而非手段），从而得以兴盛。当然，采用 (S1) 的智能体会错失道德本身所蕴含的博弈论理性，即相互合作的收益大于相互背叛时所体现的理性。同样，采用 (S2) 的智能体可能会做出极端且看似不必要的自我牺牲行为（用康德术语来说，即违反对自身的义务）。此外，在混合群体中，合作型智能体会成为搭便车者的目标，后者利用他人的善意（Fehr 和 Fischbacher，2004），除非存在惩罚机制（Fehr 和 G¨achter，2000）。尽管如此，从机器伦理学的角度看，(S2) 值得关注，因为它体现了神命论和法律实证主义的某些方面；也就是说，如果人类已经积累了某一论域的完备道德知识（例如，确定了哪些行动有助于人类福祉），并能将其作为输入—行动指令加以实施。那么，举例来说，如果认为机器不过是实现人类目的的工具，那么极端形式的机器牺牲或机器剥削或许无关紧要。[65]

虽然（S2）对于道德机器的前景或许颇具吸引力，但从人类的角度来看就没那么合理了。简言之，如果没有自我导向的激励，人们就不太可能合作（即使康德的目的王国也假定人对自身负有某种义务）。更常见的策略，也是经济学和行为生态学中研究得最广泛的策略，是以最大化自身利益的方式保持理性：

[64]基于他关于自然状态的理念，霍布斯提出了著名的论点：人们最好服从一种绝对政治权威，因为这种权威有能力保护人们免受自身伤害。[65]从这个角度来看，设想由自身利益驱动的道德机器似乎也荒谬至极。

（S3）混合理性 ——采取任何能最大化自身利益的行动。

该策略是 $mixed$ 而非纯策略，意即可以将其表示为对每种纯策略的概率分配。它是（S1）的一个更复杂版本，因为它将合作可能带来的自身利益纳入考量（Axelrod and Hamilton，1981年）。然而，依照（S3），何为理性选择，最终取决于博弈的特征。在一次性囚徒困境（只进行一次）中，如果双方都知道以后不会再进行博弈，那么占优策略也是唯一可能的纳什均衡，即背叛。无论玩家 2采取什么行动，玩家 1背叛都会更有利，反之亦然：由于双方都无法报复对方，背叛对任何一方都没有损失。根据归纳推理，只要众所周知该博弈恰好进行 n 次，重复囚徒困境也同样如此（Luce and Raiffa，1989年）。然而，如果回合数未知或无限，理性玩家之间的合作便可能在非合作情境中出现；背叛仍可能是纳什均衡，但不再是严格支配策略（Aumann，2016年）。仅仅知道将进行多少回合，就会对行动选择产生如此大的影响，这似乎有悖直觉；而且，这也不符合人类在实验环境中的表现（Heuer and Orland，2019年）。人们为这一“悖论”提出了若干解决方案。Radner（1986年）指出，放宽严格的理性概念（例如，满足于“足够接近”的收益）可以延长合作时期。Kreps等人（1982年）则基于这样一种前提得出了类似结果：智能体对其他博弈参与者的选项、动机或行为掌握着不完全信息。从复杂性视角看，Neyman（1985年）提出了一个更有意思的解决方案，他指出，如果博弈参与者的记忆足够有限，合作便会成为一种均衡。具体而言，如果将智能体建模为具有固定大小s的有限自动机，并进行 n 场博弈，那么当 $2 \leq s < n$时，相互合作就会成为一个不动点。[66]直观地说，由于博弈参与者没有足够的记忆来数到 n，实际上他们会将其视为一场无限博弈。

这带来了一个与迄今讨论的复杂性结果相悖的见解。第4节的结果表明，智能体由于自身的计算约束而无法做到道德最优；而这些博弈论考量则说明，智能体会因缺乏某些计算资源而表现出道德行为（合作）：例如，Neyman（1985年）通过限制空间复杂度，Kreps等人（1982年）通过限制信息，或Radner（1986年）通过限制理性来实现这一点。这或许表明，规范理论还具有另一种作用，这与3.1.2节的论述相矛盾：规范理论的目的并非实现道德最优性本身，而是在新颖且复杂的情境中提供行动指导。结果可能并非最优，无论从理性视角还是道德视角来看都是如此，但考虑到智能体的资源受限，这些结果尚可奏效。Alexander（2007年）对此进行了深入探讨，他认为，道德原则的出现是为了提供“快速节俭启发式”，使具有有限认知能力的智能体能够协调一致地达成次优结果。

然而，这也是一种乐观的预设结论：它预先假定，促进合作的道德规范和原则会带来某种实质性的共同利益（例如契约论视角），而智能体本身可能缺乏认知能力

[66]关于计算复杂性语境下这一悖论的有趣讨论，请参阅 Aaronson（2013）。

来验证这种利益。因此，除非将共同利益与其他替代方案（或事实上的最优方案）进行比较，否则它就会成为自身的乐观固定点。与神命论、法律实证主义或自然法一样，这种观点可能要求智能体盲目遵循规则，因为“这些规则是好的”，却不证明这些规则为何优于其他替代方案。当然，从进化视角来看，它或许能为利他主义问题提供有启发性的事后解释：既然自然选择有利于最适者生存，为何许多生物体会表现出利他行为——通过降低自身的繁殖潜力来提高其他行动者的繁殖潜力？观察各种生物体的合作行为，人们或许会得出结论：只有当利他行为能带来某种替代性的繁殖收益（例如亲缘选择、直接互惠、间接互惠和网络互惠（Nowak，2006））时，它们才可能得以维持。同理，人们或许会为“快速而直觉式”的道德行为辩护，认为它们凝缩了进化史或文化史中的道德智慧。然而，这种观点依赖于一种乐观保守主义：即事物之所以如此自有其充分理由，即便我们可能无法完全了解这些理由（我们将在第5.3节回到这个问题）。[67]

5.1.4. 不完全信息博弈中的道德行为

与其寻找能够促成合作的计算约束，不如研究在更现实的博弈论情境中阻碍理性自利最大化的因素。例如，通过 Harsanyi（1967年）关于贝叶斯博弈的研究，我们可以对不完全信息情境进行建模：

（S4）贝叶斯理性 — 做任何能够使预期自身利益最大化的事情。

贝叶斯博弈放宽了经典博弈论中的一个基本假设：博弈的特征（例如每个博弈参与者的行动和收益函数）都为所有博弈参与者所知。例如，在 NE 的最初表述中，假定每个博弈参与者都知道其他博弈参与者的均衡策略。而在贝叶斯博弈中，博弈参与者转而对博弈的特征持有信念，这些信念可能包括对他人关于博弈特征之信念的信念。这自然会导致无限的高阶信念层级（关于信念的信念的信念，依此类推），在数学上处理起来十分繁琐。Harsanyi 的贝叶斯博弈模型通过类型这一概念部分解决了这个问题；类型概括了某个博弈参与者关于博弈本质的信念（以及其关于信念的无限层级）。[68] 如果假设博弈参与者的私人要素取自一个共同知识分布，那么无限回归就得到解决，从而可以进行贝叶斯均衡分析；这种分析规定了每个博弈参与者的行为，即该行为是对其所相信的其他博弈参与者行为的最佳回应（也就是说，在给定博弈参与者自身类型的情况下，是对其他博弈参与者策略的最佳回应）。

然而，在实践中，Harsanyi 的设定难以用于建模智能体与未知智能体交互的多智能体系统。我们如何确定一个陌生人采用的概率池与我们所知的相同？在自主智能体与多智能体系统的研究中

[67]作为一个有趣的思想实验，如果托马斯·霍布斯生活在21世纪，他还会试图说服民主社会将权力交给某个主权者吗？[68]关于 Harsanyi的无限信念层级模型的两种互补构造，参见 Mertens 和 Zamir（1985年）以及 Brandenburger和 Dekel（1993年）。

人们转而通过递归推理的概念来研究嵌套信念。[69]递归推理方法通常会将嵌套信念近似到预先设定的递归深度。例如，如果 A 试图预测 B 的行为，A 就会根据自己对 B 的信念，模拟 B 的决策，从而预测 B 的下一步行动。反过来，这又要求从 B 的视角预测 A 的行为，依据是 A 认为 B 对 A 有何信念，依此类推。随后，递归会在固定深度处终止，并从某个概率分布（例如均匀分布）中抽取行动预测。第 0 层的底层预测随后向上传递至更高层（$0+1$），在这一层选出最优动作；之后再逐层向上传递至最高层（l），由 A 实际作出选择。[70] 上述过程的一种重要形式是交互式部分可观测马尔可夫决策过程（I-POMDP）。它通过将其他智能体的模型加入状态空间，对部分可观测马尔可夫决策过程（见第 4.3 节）进行了扩展（Gmytrasiewicz 和 Doshi，2005年）。简而言之，为了让智能体 A 选出最优动作，A 必须针对 B 的每个模型求解 B 的交互式部分可观测马尔可夫决策过程；这又涉及针对 B 所拥有的 A 的每个模型求解交互式部分可观测马尔可夫决策过程，如此递归，直至第 0 层；在这一层，其他智能体的模型是标准的部分可观测马尔可夫决策过程（也就是说，它们是从某个概率分布中抽取的“噪声”）。

尽管已有若干针对 I-POMDP 的精确解和近似解，[71] 但显而易见，它们很难计算。由于第 0 层由 POMDP 构成，智能体可以递归地利用这些 POMDP 来求解更高层级的 POMDP，因此求解 I-POMDP 的复杂度等同于求解 $O(N^l)$ 个 POMPD，其中 N 是智能体在每个层级考虑的模型数量上界，l 是递归深度（Gmytrasiewicz 和 Doshi，2005年）。例如，对于包含 4 个智能体模型和 4 个层级的 I-POMDP，这相当于求解 256 个 POMDP；对于有限时间范围，这些 POMDP 单独来看均为 PSPACE 完全问题（Papadimitriou 和 Tsitsiklis，1987年）。更广泛地说，Bernstein 等人（2002年）研究了若干去中心化控制问题——多个智能体协同控制一个过程，且各自掌握的状态信息可能不同——并证明去中心化 MDP 和 POMDP 都是 NEXP-困难的；也就是说，它们至少与可在非确定性指数时间内求解的最难问题一样困难。更具体地说，尽管 POMDP 和 I-POMDP 同样受到“历史诅咒”的影响——也就是说，策略空间与由时间范围决定的未来可能信念数量成正比——但由于 I-POMDP 可能涉及更多（可能嵌套的）信念，因此还会受到“维数灾难”的进一步阻碍，因为信念表示的复杂度与信念结构的维度成正比。

然而，历史和维度并非策略互动中的唯二诅咒。例如，如果博弈参与者不确定其他博弈参与者的收益函数，预测与理性之间就存在内在张力。Foster and Young（2001年）证明，在某些情境下，理性玩家不可能既根据自己的信念采取最优行动，又同时拥有正确的信念。其推理很简单：如果 A 试图预测 B 在 t_2 的行动，而 A 必须在 t_1 采取一个 B 可以观察到的行动，那么 B 的观察可能会使 A 对 B 在 t_2 的行为预测失效。此外，在某些

[69]关于递归推理方法的最新综述，请参见（Albrecht and Stone，2018年）中的第 4.5 节。[70]事实上，零和博弈中著名的极小极大算法就是这种方法的一种特殊变体，其中对手 B 的评估函数被设定为博弈参与者 A 的评估函数的反函数（ Campbell and Marsland，1983 ）。[71] 例如，使用蒙特卡洛抽样（ Doshi and Gmytrasiewicz，2009）或模型等价（Rathnasabapathy 等人，2006）等方法。

情境中，更高阶的推理可能不会带来任何额外收益。事实上，de Weerd 等人（2013年、2017年）的研究表明，在一些情境中——例如序贯谈判和石头剪刀布——推理层级高于 2 并不会给计算代理带来任何显著优势。

然而，尽管递归推理的不可处理性结果会直接影响计算系统，但人类如何在类似问题中进行推理，以及机器应如何以稳健的方式与人类互动，仍是悬而未决的问题。实验心理学对策略性人际情境进行了广泛探索，而这些研究往往结合了“心智理论”这一概念（Yoshida 等人，2008年）。例如，在两项以人类为参与者的实验中，Hedden 和 Zhang（2002年）表明，参与者会采用一种关于其他博弈玩家心智的短视“默认模型”，并根据新证据动态调整该模型。此外，Camerer 等人（2004）提出的“认知层级”模型认为，博弈玩家会假定自己的策略最为复杂，即他们会在递归层级 l上使用最佳回应，来预测层级 $l-1$ 上玩家的行为（也就是说，处于嵌套信念层级中“高出”一步的位置）。通过用来自多种博弈的大量实证数据拟合模型，他们发现，人类的平均递归推理深度为 1.5。[72]

鉴于该主题涉及诸多相互冲突的维度，要对计算约束是否以及在多大程度上影响道德语境中的策略互动得出普遍结论，似乎并不容易。不过，这些结果初步表明了一种趋势：考虑到信息可得性（关于状态空间以及其他博弈玩家的信息）、有限理性（例如记忆和递归层级）以及博弈设定（例如一次性或重复、随机或确定性），人们会以最优性换取效率（或仅仅是可行性）。直观而言，这些复杂性结果或许能支持纯策略（如 S1 和 S2）的吸引力：在复杂情境中，简单地相信每个人都属于某种类型（例如自私或利他）会更容易。这也可能解释了为什么在“局部”范围内（例如较小的朋友圈中）更容易合作，因为博弈特征是共同知识；大家拥有共同的目标（例如最大化共同利益），并共享用于发现和惩罚搭便车的机制。[73] 同样，这也揭示了相反的一面：为什么在更大的群体中难以实现合作——博弈特征并非人人共有，智能体也不知道彼此会有何种预期。

5.1.5. 计算道德均衡

从道德计算视角来看，博弈论中或许最有意思的部分，并不是探究智能体需要多少资源才能决定如何行动，而是寻找能够使所有人的利益最大化的策略（例如，假设所有人都遵循同一策略）。这有两个原因。首先，从道德角度看，这大致对应于第 5.1.2 节讨论的通用规则，例如，

[72]然而，Goodie等人（2012）对“人类递归推理是悲观的”这一观点提出了质疑。该观点认为，人类递归推理的层级相对较低，且建立在低估对手的基础上。Goodie等人发现，在竞争性和简单情境中，所有参与者都运用了最高可用推理层级。[73]反过来，合作群体可能会受益于“群体智慧”这一现象（Yi等人，2012）：多种解决方案的聚合比单个解决方案表现更好。

GR 和 CI。也就是说，如果某个行动规则（或策略组合）对于其他所有行动都是相容的最佳回应，并且 能够带来某些实质性的共同利益，那么道德共同体就会乐于找到这些规则。其次，这些通用规则又对应于博弈论中典型的解概念——即纳什均衡——且具有从道德立场看颇具吸引力的性质（例如，使共同利益最大化）。

因此，自然会提出这样的问题：计算纳什均衡有多困难？早在1928年，v. Neumann（1928年）就提出了极小极大定理，由此可知，双人零和博弈中的均衡可以通过线性规划在多项式时间内计算出来（Khachiyan，1979年）。然而，对于至少有两名博弈参与者的非零和博弈，已证明该问题是PPAD-完全的（Chen等人，2009）。PPAD由Papadimitriou（1994）提出，是一类函数问题：这类问题可由非确定性图灵机在多项式时间内求解，并且基于有向图上的奇偶性论证，保证存在解决方案（“PPAD”）。奇偶性论证基于一项图论洞见：在所有有限无向图中，与奇数条边相连的节点数为偶数（这定义了PPA类，PPAD是其子类）。对于有向图，也有类似的洞见：给定一个有向图和一个源节点（没有前驱节点的节点），必然存在一个“到达线路尽头”的节点，它没有后继节点。[74] PPAD复杂度类的特别之处在于，它体现了纳什均衡的特殊性：由于纳什均衡总是存在（纳什，1951年），决策问题“这个博弈是否存在纳什均衡？”的答案总是是，因此它不可能是NP难的。类似地，它也体现了布劳威尔不动点定理的特殊性：对于任何将紧致凸欧氏空间映射到自身的连续函数 f，至少存在一个点 x_*，使得 $f(x_*) = x_*$。打个比方：如果你站在某个区域内，并展开一张该区域的地图，假设地图上没有空洞，那么地图上至少有一个点与你所在的位置相对应。但仅仅知道这样的点存在，并不一定意味着很容易找到它。假设可以高效地检查地图上的某个点是否确实对应你所在的位置，那么定位问题就能在非确定性多项式时间内解决（NP归属性）。同理，如果能够高效验证某个策略组合是否为最佳回应，那么寻找纳什均衡的问题就属于NP（而NP又包含PPAD类）。

然而，如果纳什均衡具有任何特殊性质——例如，使博弈参与者效用之和最大化，或使每个人都获得至少达到某个数值的期望收益——问题就会变成NP完全问题（Gilboa和Zemel，1989年）。[75] Conitzer和Sandholm（2008年）进一步强化了这些结果，并证明平等主义优化问题（例如，最大化共同收益）是不可近似的；也就是说，不可能在多项式时间内找到一个能够近似最大联合收益的纳什均衡。事实上，最优纳什均衡是众多已知的与纳什均衡相关且难以近似的NP难问题之一（Austrin等人，2013年）：例如，计算是否存在多个纳什均衡，或寻找具有最小支持集的纳什均衡。[76] 更准确地说，尽管线性时间算法可以获得 $1/2$-近似的

[74]“终点”问题是一个典型的 PPAD-完全问题。[75]可以通过将其归约为在无向图中寻找大小为 k 的团的问题来证明这一点（Gilboa和Zemel，1989年）。例如，寻找大小为 3 的团，就是寻找 3 个彼此相连的节点。[76]最小支持问题问的是：给定一个数 $k \leq 1$，是否存在一个纳什均衡，使得博弈玩家使用的正概率策略不超过 k 个？

最优纳什均衡（Daskalakis 等人，2006），但实现优于 $1/2$ 的近似与植入团问题一样困难，而后者或许可以在准多项式时间内求解（Lipton 等人，2003年）。[77]

自然地，在不完全信息或随机动态的情境中，问题可能会更加棘手。由于标准式博弈是贝叶斯博弈的特殊情形，在贝叶斯博弈中寻找混合策略均衡仍然是 PPAD 困难的。Conitzer和Sandholm（2008年）指出，即使在对称型双人贝叶斯博弈中，判断博弈是否存在纯策略贝叶斯均衡也是 NP难。相比之下，在标准式博弈中，只需检查是否存在某种纯策略组合构成纳什均衡，就能在多项式时间内判断纯策略纳什均衡是否存在。然而，这种方法不适用于贝叶斯博弈，因为策略空间会随着类型数量呈指数级增长，而这些类型包含玩家偏好的私人信息。此外，Conitzer和Sandholm（2008年）还证明，在具有概率状态转移的重复博弈中，判断是否存在纯策略纳什均衡是 PSPACE困难的。这类博弈也称为“马尔可夫博弈”，因为它们将马尔可夫决策过程扩展到了包含多个决策者的情形；而在有限博弈中，这一问题仍然是 NP难。

虽然纳什均衡是博弈论中研究最广泛的解概念，但它并非唯一具有道德意义的解概念。一种重要的替代方案是 Aumann（1974年、1987年）提出的相关均衡 （CE）。简单来说，相关均衡可以看作共同共享贝叶斯理性的结果：它只是在给定博弈参与者所掌握信息的情况下最大化效用。与纳什均衡的理性假设不同，相关均衡不假设博弈参与者知道其他博弈参与者会按照纳什均衡所规定的方式采取行动，也不假设每位博弈参与者都知道其他人的策略。相反，均衡意味着，根据自己掌握的信息（无论是私人信息还是共同信息），没有任何博弈参与者会认为偏离自己的策略能带来更高收益。一个有用的类比是，想象博弈参与者通过观察某个随机事件来选择行动：在混合策略情形中，假设每位博弈参与者观察到的事件彼此独立；而在相关情形中，这些事件则未必独立。例如，博弈参与者 A 和 B可能会私下观察到相关信号——比如两个交通信号灯——分别指示 A 等待、B 通行。如果这些信号是从相关分布中抽取的，那么 A和 B 都不会想违背信号的指示（因为闯红灯可能导致他们相撞）。因此，相关均衡是策略组合上的一般分布，而混合策略纳什均衡则是“非相关”策略空间上的分布（各博弈参与者的策略独立分布）。由于每个混合策略纳什均衡都可以定义为各博弈参与者的混合策略之积，因此纳什均衡实际上是相关均衡的一种特殊情形；并且每个博弈都存在一个相关均衡。

除了保证始终存在之外，相关均衡还有多方面的吸引力。它似乎会在纳什均衡不会出现的自然情境中涌现（Hart and Mas-Colell，2000年）。[78] 与纳什均衡不同，相关均衡更适合容纳对结果起决定作用的外部因素，例如遵循可信来源的建议。从道德视角来看，相关均衡涵盖了许多情境——例如道德或遵循法律规则的情境——在这些情境中，显然

[77]这里，植入团问题被用作困难性假设，因为人们猜想，不存在多项式时间算法能够以优于随机猜测的准确率区分植入团与随机图（Hazan 和 Krauthgamer，2011年）。[78]另请参阅 Cesa-Bianchi 和 Lugosi（2006）第7章。

人们受到激励，不偏离规定的行动（例如遵守交通规则）。最后但同样重要的是，相关均衡比纳什均衡更容易计算。由于相关均衡可以用一组线性不等式来定义（Hart and Schmeidler，1989年），而且不必基于纳什的结果，因此可以通过线性规划求解寻找相关均衡的问题；对于参与者人数不限的博弈，该问题都能在多项式时间内计算（Gilboa和Zemel，1989年）。Papadimitriou and Roughgarden（2008年）证明，对于许多自然类别的简洁表示的多人博弈，包括多矩阵博弈、图博弈、超图博弈和调度博弈，都存在可在多项式时间内计算任意相关均衡的算法。遗憾的是，尽管任何相关均衡似乎都易于求得，Papadimitriou and Roughgarden（2008年）也表明，在几乎所有类别的简洁博弈中，计算出使预期联合收益最大化的相关均衡仍然是NP难问题。

除了这些不可处理性结果之外，人们或许也会质疑计算道德上良好的均衡是否有意义。当然，在系统层面，这些结果为建模和衡量复杂的行为动态提供了不可或缺的分析工具。例如，无政府代价 通过将“最坏情况”分散式均衡下的福利值除以最优集中式配置下的福利，来衡量自私行为导致系统共同福利的损失（Koutsoupias and Papadimitriou，2009）。[79] 不过，从个体智能体的角度来看，这些工具的用处或许不大。我们可以设想，某个智能体已经成功计算出一组“黄金法则”策略组合（例如，一组行动规则），无论发生何种互动，这组策略都能使其社会的共同利益最大化。然而，只有当社会中的其他 智能体遵循她通过均衡计算确定的策略组合时，她的社会才能繁荣。她或许会诉诸自己非凡的计算能力[80]，试图说服其他智能体听从她的引导；她或许会像康德那样，主张这对于理性存在者的意志是至关重要的；也可能会像霍布斯那样，主张这对于预期自利收益的最大化是至关重要的；或者她可能只是希望其他所有人——例如，凭借共同的理性能力和相关分布——都能计算出相应的策略组合。只要手段得当，她或许会通过惩罚所有不遵循其策略组合的人，来强制维持这一均衡。极端情况下，我们可以设想一位超级智能利维坦，它经过计算后发现，即使将自由受限和惩罚异见者所造成的潜在福利损失考虑在内，若以强制手段推行策略组合，共同福利仍能达到最优。

核心观点是，分散式合作问题的解决方案要在实践中发挥作用，就需要在某种非平凡的意义上具备集中式特征。共享认知能力、理性、交流（Crawford and Sobel，1982年）以及信息（例如来自相关分布的信息）固然能够让智能体更容易集体达成道德上理想的均衡，但合作的复杂度会受到地方性及智能体特定语境中大量特征的扭曲；例如或多或少公开的权力动态（如通过法律、政治和宗教制度体现出来的权力动态）、心理启发式（例如信任、羞耻和内疚），以及

[79]相反，稳定性价格 衡量的是“最佳可能”分散式均衡与最优集中式解之间的差异比率。[80]遵循 Cummings 等人（2016）最近提出的一项方案，协调复杂度 可以定义为：拥有关于博弈的完备知识的集中式协调者，为了协调博弈参与者实现近乎最优的解决方案而需要公开传达的最少信息量。

内群体/外群体偏好之间的冲突。

此外，在现实世界中，我们常常不清楚自己正在参与哪种博弈，甚至不确定自己是否在参与博弈。博弈可以是合作型或非合作型、离散型或连续型、一次性或重复型（例如，与陌生人或朋友进行），同时进行或序贯进行、零和或非零和、对称型或非对称型；信息可能在不完美和不完全的程度上各不相同，群体规模也可能不同，而且还允许存在潜在无限多个不同策略的集合（例如，未知长度的序贯马尔可夫过程中，特定的行动组合）。换言之，即使假设存在"集中式特征"（如共享的认知、记忆和理性观），博弈论模型——进而，均衡解的适用性——仍在很大程度上无法由数据确定。小型且相互隔离的群体或许倾向于合作——例如，容易识别共同利益的最大化；拥有发现搭便车行为的机制；知道他人掌握着相同的博弈信息；等等——而较大的社会则可能需要数千年才能变得平等主义，例如，智能体无法就自己正在参与的协调博弈的性质达成一致（比如，由于群体内偏好之间存在冲突）。[81]

5.1.6. 道德系统设计

博弈论模型具有高度不确定性，而且在这类模型中计算良好均衡十分困难，这表明通过设计促进集体福利或许更容易；也就是说，构建能够自然产生易处理且在道德上值得称许的均衡的系统（而不是在任意情境中寻找良好均衡）。这直观地体现了合法性与道德之间的关系。尽管智能体可能出于道德理由而遵守规则（例如，如果人人都遵守规则，就能带来共同利益），但若存在责备、责任追究和惩罚机制，他们遵守规则的意愿可能会进一步增强。同样，制度化道德规则（例如"不要伤害他人"）在设计上或许能体现自然形成的策略组合，既易于理解、在道德上有正当依据，也便于实际应用。在经济学中，机制设计（或"逆向博弈论"）的目标同样是设计能够实现预期目标的分散式经济机制，例如优化社会福利（Hurwicz and Reiter, 2006）。在算法博弈论领域，这也进一步推动了算法机制设计的发展。该领域旨在设计兼具计算和博弈论视角下优点的博弈，典型例子是具有良好最坏情况均衡（即无政府代价较低）且能在多项式时间内计算出来的博弈（Nisan and Ronen, 1999）。本质上，由于计算代理难以在开放式环境中找到并遵循道德均衡，复杂性考量有助于指导人们设计在道德上具有吸引力的分散式系统。[82]

总而言之，我们对道德规则普遍性的探究发现：（1）尽管人类情境中的道德规则遵循看起来在计算上

[81]明确一下术语的含义：通常所说的协调博弈，是指博弈参与者合作会更有利的博弈，例如猎鹿博弈或性别之战。因此，这类博弈不同于鹰鸽博弈（斗鸡博弈）这类反协调博弈；在反协调博弈中，博弈参与者选择不同策略才更有利。[82]算法机制设计的一些著名例子包括交通路由（Roughgarden，2005年；Roughgarden 和 Tardos，2002年）、拍卖设计（Cai 和 Papadimitriou，2014年）以及互联网问题（Feigenbaum 和 Shenker，2004）。

十分简单，但它依赖于“通常具备能力”的人类行动者之间共享的认知能力与共识（通常还需要制度来确保规则得到公正适用）。尽管诸如“不要 X”之类的规则可能降低空间复杂度，但这只有通过限制自主智能体能够采取的潜在无限的行动空间才能实现。

（2）如果计算主体应当以确保自身获得规范性辩护的方式遵循普遍道德规则——例如，通过理性协议或最大化共同利益来顾及其他存在——那么算法博弈论的结果表明，这需要解决难解问题（但有少数值得注意的例外）。请注意，这些不可处理性结果适用于显著博弈；这里的意思是，假定智能体拥有其（据称）正在参与的博弈模型，即使有关博弈某些方面的信息可能并不完整。在许多现实世界情境中，即便能够获得这类模型，其有效性仍可能无法由数据充分确定（也就是说，模型可能无法捕捉现实世界的关键动态）。换言之，如果我们假设，道德规则遵循可以 被数学地建模为遵循博弈中的策略，那么这类博弈的解决方案可能是可判定的（例如，通过计算均衡），尽管一般而言并非多项式时间可解。还必须强调的是，这些挑战绝非义务论所独有，而是影响任何具有普遍性追求、旨在解释多主体动态的道德理论。因此，这些挑战自然也指向结果主义（因为它关注结果如何影响他人），但并不针对神命论和法律实证主义等理论。

(3) 鉴于寻找良好均衡的难度、计算他人信念（例如嵌套信念）所面临的问题、博弈模型的不确定性，以及合作可能因缺乏某些计算资源（例如信息、记忆或自身利益）而涌现这一事实，这项研究还表明，道德规则和原则或许更适合作为“快速节俭启发式方法”，引导认知有限的智能体在开放式环境中取得虽非最优、但切实可行的结果。颇具讽刺意味的是，从计算视角来看，这反过来可能使“天真”的道德策略（例如“有疑虑时，就采取利他行动”）变得可行，甚至成为必要；尽管这并未从博弈论角度解释，为什么 这些策略从自身利益的角度来看是可行的。[83] 此外，与其在复杂环境中计算道德均衡，一种有前景的替代方法是利用复杂度约束和均衡度量，指导设计在道德视角下具有吸引力的分散式系统。

5.2. 道德逻辑与语义学的复杂度

分析义务论的计算复杂性，另一种方式是考察旨在捕捉道德推理和规则遵循的形式语言的句法与语义学。例如，我们可以注意到，以自然语言表达的道德规则具有某些逻辑特征，例如“给定事实 a，行动 b 是义务”。因此，我们可以

[83]为了说明这一点，我们可以设想两个群体——A 和 B——它们都形成了相同的一套合作行为，并由此实现了最优的共同福利。在群体 A 中，均衡是自利个体经过数千次迭代的重复互动后形成的。在群体 B 中，均衡从一开始就已固定，因为智能体的母亲告诉它们“永远要善良”。关键在于，尽管 A 中的智能体理解合作背后的自利逻辑，但它们的境况与 B 中对这种逻辑一无所知的智能体同样好。

设想构造一台机器是可能的：这台机器的记忆中存储着若干规范（x，y），这些规范将事实（x）与义务或许可（y）联系起来。问题在于，是否存在任何复杂性考量，可能限制这台机器在规范推理方面的实际成效（这将是第5.2.2节的主题）？更广泛地说，用于道德推理的逻辑的表达能力与复杂性类之间有何关系，例如，就其能够描述的问题而言（这是第5.2.1节的主题）？本节将讨论这类问题，以及语义学中一些更为基础的问题（第5.2.3节）。

5.2.1. 可判定性与描述复杂性

若不重申构成可计算性基础的经典结果，关于逻辑复杂度的讨论便是不完整的。1929年，库尔特·哥德尔的完备性定理确立了一阶逻辑（FOL）中的语义真与句法可证性相对应；也就是说，存在针对一阶逻辑的完备、可靠且有效的演绎系统（哥德尔，1930年）。两年后，哥德尔提出了著名的两条不完备定理（哥德尔，1931年），指出任何能够进行初等算术的相容形式系统——例如，使用自然数、加法和乘法——都是不完备的（第一定理），并且这样的系统无法证明自身的一致性（第二定理）。受哥德尔定理启发，图灵（1936年）和丘奇（1936年）分别对一阶逻辑的“判定问题”（Entscheidungsproblem）给出了否定答案；丘奇的证明利用了判定 λ 演算中两个表达式是否等价这一问题的不可判定性，而图灵则构造了图灵机的停机问题。他们的结果确立了不存在能够判定任意一阶逻辑公式是否为逻辑有效式的决策程序。另一方面，命题逻辑（PL）与一阶逻辑不同，不包含关系和量词，因此是可判定的。例如，命题逻辑的可满足性问题（SAT）是首个被证明为 NP完全的决策问题（库克，1971年），此后一直处于计算复杂性理论的核心。利用库克关于 SAT 的结果，卡普（1972年）证明了另外21个问题是 NP完全的，这不仅展现了 NP类直观上的吸引力，也表明许多自然的计算问题都难以处理。

然而，我们不必确定检查某个输入是否满足某个性质 X 所需的资源，而是可以问：表达 X 的复杂度是多少？后一个问题是描述复杂性理论领域的核心；该领域根据表达其中语言所需的逻辑类型来定义复杂性类。事实上，得益于描述复杂性领域的大量研究成果，我们发现，表达与检查密切相关。[87]该领域的第一个重大成果是 Fagin 定理（Fagin，1974年），它确立了 NP 恰好等于能够用存在二阶逻辑（SO∃）表达的性质集合；与一阶逻辑不同，存在二阶逻辑能够对性质和关系进行存在量化。换言之，每个可在 NP 中计算的查询（包括 NP完全问题）都

[84]一个系统——或者更确切地说，数学逻辑意义上的一种理论，即由形式语言中的句子构成的集合——如果不会导致矛盾，就是相容 的。若一个公理系统中的系统语言里的任何陈述都能由公理证明，则该系统是完备 的。因此，不完备性意味着系统中存在既不能证明、也不能证伪的陈述。[85]Rosser（1936年）进一步完善了这些定理，他证明这些定理时不必假设 ω-相容性。[86]有效性是指，真实的前提能够保证论证结论为真。[87]关于描述复杂性的权威性介绍，参见 Immerman（1998年）。

等价于 SO∃ 中的一个查询。[88] 此外，由于存在公式（对论域中的某些成员进行量化）的补是全称公式（对论域中的所有成员进行量化），[89] 因此可以直接推出，co-NP 可由 SO∀ 表达，而允许全称量化和存在量化的无限制二阶逻辑（SOL）等于多项式层次（PH）中所有复杂性类的并集。其他值得注意的结果包括：一阶逻辑对应于对数时间层级（LH）以及电路复杂度类 AC^0,；带有最小不动点算子的线性有序一阶逻辑系统给出 P 类（Immerman，1982年；Vardi，1982年）；带有传递闭包的 SOL 给出 PSPACE（Immerman，1989年）；而带有最小不动点的 SOL 对应于 EXPTIME（Immerman，1998年；Abiteboul 等，1997年）。

这些结果带来了这样一种认识：一种语言的表达能力与其能够描述的问题直接相关。然而，应当注意的是，尽管计算复杂度与描述复杂度密切相关，但二者也存在一些关键差异。例如，一个重要的不一致之处在于，描述复杂度研究的是有限的数学结构，而计算系统则在问题的有序编码上运行，因此能够枚举逻辑上可能无序的对象。举例来说，从逻辑角度看，我们可能会将图中的一组节点视为无序的；但一旦它被转移到图灵机的磁带上，就不可避免地变成有序的（因而可以用于各种形式的操作）。[91] 尽管如此，复杂度可以用表达能力来刻画——无需借助某种抽象机器——这一事实进一步彰显了复杂性类的自然吸引力。而对于计算型道德规则遵循者——例如，对包含道德规范的数据库执行查询的系统——其后果意义深远，却又因其仅仅具有一般性而在某种程度上显得晦暗不明。例如，用一阶逻辑可描述的语言对应于 AC^0 （多项式规模的有界深度电路），这使得人们能够执行整数加法和减法，但不能进行乘法。另一方面，加入一个能够计算二元关系传递闭包的运算符后，就可以生成能够回答可达性查询的结构（例如，是否可以在 n 个步骤内从节点 A 到达节点 Z？）。也许最有趣的是，表达涉及对性质和关系进行不同形式量化的查询（而不仅仅是像在一阶逻辑中那样对对象进行量化）具有不可处理性，而这在自然语言中似乎是直观的。例如，即便是如下这种形式的简单道德查询："对于所有可观察事实、可能的行动和义务，是否存在某个不违反任何道德义务的行动？"，在最坏情况下，也可能只能用难解复杂性类来表达（例如，假设义务是事实与行动之间的一种关系）。关键在于，尽管对性质和关系的量化、传递闭包以及不动点为经过充分研究的形式系统集合提供了强大的表达能力，但它们也引发了关于不可处理性的担忧；这仅仅是因为语言帮助我们简洁地表示和传达有关道德规则遵循者的查询

[88]详细证明请参见（Immerman，1998年）第7.1章。[89]这是因为 $\exists x \neg P(x)$等价于 $\neg \forall x P(x)$。[90]简言之，不动点逻辑通过一个运算符扩展了一阶逻辑，该运算符能够构造关系变量的不动点。例如，如果我们把含有自由关系变量的公式视为确定关系空间上地图的公式，那么这个运算符就能在该地图上定义不动点。[91]然而，这种不一致性并不适用于 Fagin 定理，因为 SO∃ 可用于声明某种所需序的存在性。

r者，但这并不一定会使此类查询变得易于计算。[9] [2]

5.2.2. 模态逻辑与道义逻辑的复杂度

一阶逻辑中应用最广泛的片段之一是模态逻辑；它是表示必然性（□）和可能性（◊）的首选逻辑。[93] 从 20^{th}世纪中叶起，模态逻辑的标准语义学便采用可能世界方法，[94] 其中，$\Box P$ 表示 P 在所有可能世界中都为真，而 $\Diamond P$ 表示 P 在某些可能世界中为真（假定这些世界彼此可达）。模态逻辑的一个优点是，与一阶逻辑不同，它具有稳健的可判定性。[95] 然而，尽管模态逻辑是可判定的，其典型问题通常并不易处理。例如，模型检查问题询问给定公式在给定克里普克结构的给定状态下是否为真，[96] 该问题可在线性时间内求解；而有效性问题询问一个公式是否在所有克里普克结构的所有状态下都为真，该问题是 PSPACE 完全的（Ladner（1977））。更确切地说，Ladner（1977）证明，模态逻辑的有效性问题对于 K系统、T系统和 S4系统是 PSPACE 完全的，而对于 S5系统则是 NP完全的。[97] Halpern 和 Moses（1992）将这些结果推广，证明了 K系统、T系统、S4系统和 S5系统的多智能体版本（即至少两个智能体所用逻辑的“合并”）的有效性问题是 PSPACE 完全的。Halpern 和 Moses（1992）的研究还表明，添加分布式知识算子——使一个“全知”智能体能够综合其他所有人的知识——不会改变复杂度；而添加共同知识算子——使每个人都知道 P，并且每个人都知道每个人都知道……P 成立——则会使该问题成为 EXPTIME 完全问题。[98]

当然，其他流行版本的模态逻辑也同样受到不可处理性问题的困扰。例如，Spaan（1993）证明，有效性问题的 PSPACE 完全性同样适用于时态情形（时序逻辑），也就是说，加入表示“总是会发生……”和“过去总是会发生……”的操作符。类似地，Sistla 和 Clarke（1985）证明，线性时态逻辑（LTL，由 Pnueli 于1977年提出）中的可满足性问题，其复杂度取决于所使用的操作符：对于包含“下一个”和“直到”操作符（不包括过去时）的情况，该问题要么是 PSPACE 完全，要么是 NP完全。[99] 一般而言，PSPACE 完全性也

[92]关于相关问题“逻辑全知”的有趣讨论，参见 Aaronson（2013）。该讨论运用复杂性考量，对这样一种观点提出质疑：如果智能体知道某些事实，那么它也知道这些事实的所有逻辑推论。[93]或者，由于模态逻辑的核心思想远早于一阶逻辑——例如亚里士多德的模态三段论——因此，将模态逻辑描述为命题逻辑的一种扩展或许更为准确。[94]这种方法最初由 Carnap（1947）提出，之后由克里普克（1963）发展为其现代形式。[95]深入阐述请参阅 Vardi 的报告《为什么模态逻辑具有如此强的可判定性？》（Vardi，1997）。[96]克里普克结构是一种图：它以节点表示可达状态，以边表示状态转移，并通过标记函数记录每个状态所具备的性质。[97]对于不熟悉模态逻辑的读者，K、T、S4 和 S5 指的是添加到系统中的公理和规则的不同选择；例如，K——最弱的版本——只包含必然化规则（即 $(\models P) \implies (\models \Box P)$）和分布公理（$\Box(P \to Q) \to (\Box P \to \Box Q)$）；T 除了 K 之外还包含自反公理（$\Box P \to P$）；S4 和 S5 则在 T 的基础上分别加入迭代公理 4（$\Box P \to \Box\Box P$）和 5（$\Diamond P \to \Box\Diamond P$）。[98]关于模态逻辑复杂度的全面论述，另请参阅 Spaan（2016）。[99]关于时态逻辑和命题逻辑复杂度的更详细研究，参见 Vollmer 等人（2009）

适用于 若干版本的 LTL 的模型检测问题（Schnoebelen，2002年） .

对于道德机器的前景而言，或许更值得关注的是动态逻辑（DL，由 Pratt 于1976年提出）和道义逻辑等模态逻辑变体的复杂性结果。DL 增添了模态算子 $[a]$和 $\langle a\rangle$，使其能够刻画程序行为的性质；例如， $[a]P$ 表示执行行动 a之后， P 必然为真（即 a 会导致 P），而 $\langle a\rangle P$ 表示执行 a 之后，P 有可能成立（即 a 可能会导致 P）。请注意， a 可能指整个程序，因此动态逻辑能够形式化多个程序构成的复杂算法系统的动态特征——例如转移、序列和结果。随后，命题动态逻辑（PDL，由 Fischer 和 Ladner 于1979年提出）以命题逻辑（PL）为基础（而非以一阶逻辑（FOL）为基础；后者是 DL 第一个版本所依托的逻辑（Pratt，1976）），旨在描述计算机程序的正确性、终止性和等价性。有趣的是，要判定 PDL 的公式 F 是否可满足，可以让两个子过程并行运行：一个枚举所有可推导的公式（R1），另一个枚举 PDL 的有限模型，并检验它们是否满足这些公式（R2）。这样一来，如果 F 可满足，最终必定会找到一个满足 F 的模型。如果 F 可满足，R2 会在某个时刻回答“是”；如果不可满足，R1 则会在某个时刻回答“否”（任一子过程给出答案后，该过程即停止）。尽管如此，PDL 虽适用于程序行为的形式化验证，但对于受多项式时间限制的计算道德智能体而言，吸引力较小，因为 PDL 的 SAT 问题是 EXPTIME 完全的（Fischer 和 Ladner，1979；Pratt，1980）。[100]

另一方面，道义逻辑（由 Von Wright（1951年）提出）旨在刻画道德概念的逻辑特征，例如许可（通常用算子 P 表示）和义务（通常用 O 表示）。例如，公理 $O(A \to B) \to (OA \to OB)$ 表示：“如果必须满足 A 蕴含 B，那么只要必须满足 A ，就必须满足 B 。”同样，康德的“应当蕴含能够”可以表示为 $OA \to \Diamond A$。因此，道义逻辑一直是机器伦理学实现中广受欢迎的框架，这并不令人意外，例如用于自动化伦理推理（Arkoudas 等人，2005年；Wiegel 和 van den Berg，2009；Furbach 等人，2014年；Malle 等人，2017b）。然而，与其他模态逻辑相比，道义逻辑的计算复杂性仍相对缺乏研究。原因之一在于，规范性概念（例如能动性、权利、责任和承诺）本身的复杂度及其相互关系，可能会被展开到任意细致的程度 [101] 。另一个原因是，人们难以就规范的恰当语义达成一致，例如，这些语义究竟应基于可能世界、公理化构造，还是操作性执行。道义逻辑语义面临的一个更根本的挑战体现在“约根森困境”（Jørgensen，1937年）中：它提出，包含规范（例如命令句）的论证是否表达真值。在标准的逻辑蕴涵观念中，真结论必须从真前提推出；也就是说，结论和前提都可以为真或为假。然而，由于命令句——据称——既非真也非假，

线性时态逻辑中的操作员。[100]De Giacomo 和 Massacci（2000年）提出了一种更实用的命题动态逻辑可满足性算法——尽管其最坏情况下仍为 EXPTIME 完全。[101]有关阐述，请参见 Sergot（1998年）；该阐述也收录于 Gabbay 等人（2013年）编著的道义逻辑手册中。

它们无法在论证有效性中发挥任何有意义的作用，因此我们无法基于逻辑推理来证明命令句的合理性。与此同时，在前提和结论均为命令句的情况下，基于命令句进行推理似乎在逻辑上是有效的（J¨ørgensen，1937年，第290页）：

(P1) 像爱自己一样爱邻人
(P2) 爱自己 (C) 爱你的邻人

因此，两难境地意味着，我们要么必须接受规范性陈述没有真值，要么否认论证的前提和结论具有真值。反过来，这些考量催生了若干类道义逻辑，大致可分为两派：可能世界方法——例如标准道义逻辑、“确保”逻辑（STIT）（Horty，2001年）和动态道义逻辑（Meyer 等，1988年；Van Der Meyden，1996年）——以及不以可能世界为基础的方法，例如输入／输出逻辑（Makinson 和 Van Der Torre，2000年）、祈使逻辑（Hansen，2008年）、可废止道义逻辑（Governatori 等，2013年）和优先级默认逻辑（Horty，2012）。

关于复杂度，第一条路径意味着必须应对常见于可能世界语义中的 PSPACE 完备性（Halpern 和摩西，1992年），或动态逻辑的 EXPTIME 完全性（Fischer 和 Ladner，1979年）。事实上，研究表明，用于构建道义 STIT 的 STIT 片段，其可满足性问题是不可判定的（Schwarzentruber 和 Semmling，2014年）。[102] 作为可能世界语义的一种替代方案，Sun 和 Robaldo（2017年）基于一种“规范基础”的操作语义，研究了输入／输出逻辑的复杂度。在这一观点下，陈述性语句可以为真或为假，而规范则不能：它们只会被违反或得到遵守。更确切地说，规范是有序对（x，y），对应于输入／输出系统的演绎操作：以事实（x）作为输入，并产生义务（y）作为输出。然而，他们的研究发现，输入／输出逻辑的判定问题是 NP/co-NP-困难的，并且处于多项式层次的 2nd 层（即 $\mathrm{NP}^{\mathrm{NP}}$ 和 $\text{co-NP}^{\mathrm{NP}}$）。相比之下，Governatori 等人（2013年）证明，一种能够计算“弱”许可（除非明确禁止，否则允许）和“强”许可（仅在明确许可时才允许）的可废止道义逻辑具有计算可处理性。然而，所构建的对象逻辑可表达性相当受限，因为它仅包含命题符号及其对应的否定，并且只允许使用模态文字（通过对命题文字应用表示义务和可许可性的模态算子而得到）。

总之，综述的结果支持这样一个结论：使用模态逻辑——例如时序逻辑、动态逻辑、道义逻辑——来表示或自动化道德推理，通常会引入计算不可处理性。自然，为了捕捉伦理生活丰富而复杂的面貌，逻辑形式的规范性概念必须能够涵盖知识、时间、能动性、程序行为以及多主体动态，这不可避免地会增加复杂度。然而，必须指出，尽管这些结果看起来可能表明，对于受多项式时间限制的道德机器而言，其前景存在根深蒂固的问题，但它们同时也

[102]另请参见 Balbiani 等人（2008年）、Herzig 和 Schwarzentruber（2008年）以及 Xu（1998年）关于 STIT 的复杂度和可判定性的其他研究结果。

对其他实际用途极具价值。例如，知道某个问题已被证明属于 PSPACE，或许就能将它与大量其他 PSPACE问题联系起来，并借助为解决这些问题而开发的算法、近似技术和启发式方法。动态逻辑的可满足性问题是 EXPTIME 完全，这一事实并不妨碍它在程序行为的形式化验证中发挥作用。更广泛地说，从设计角度来看，复杂性结果或许能极大帮助人们为特定的实际目的选择形式系统——包括其表达能力、操作员、句法和语义学。

5.2.3. 道德语义学问题

语义问题有多种不同形式，其中一些数百年来一直受到哲学界的广泛探讨。本节前面曾指出，如果智能体应遵循一条形式为"'如果输入 $X \rightarrow$ ，则执行行动 Y'"的规则，我们就假定智能体对 X 和 Y 的含义有一定程度的理解，并非一无所知。在命题逻辑的语义学中，我们只需说，相对于某个模型（或在模态逻辑中，相对于可能世界），这些项非真即假，并给出规则（例如真值表），以确定由这些项和逻辑联结词构成的句子的真值（例如，$\neg$，$\wedge$，$\vee$，$\rightarrow$，$\Leftrightarrow$）。类似地，编程语言的语义学或许只需定义这样一个过程：句法中有效的字符串（指令）如何引发（解释）特定的状态转移（执行），例如通过操作某种数据结构。在这类情况下，语义学可以被视为纯粹机械的操作，而建立模型与现实之间联系的工作——例如，说明语言能够表达的任何有效句子如何对应于现实世界中的某种事态——则被绕开了。

然而，人类道德话语——例如关于美德、原则和判断的话语——充斥着语义上所谓的"厚"描述和概念。例如，道德谈论中包含"厚描"，其意义中嵌入了主观性，例如通过内在动机来解释个体的行为（Geertz等人，1973）。也就是说，当我们将某个智能体描述为勇敢或公正时，自然会认为该智能体具有体现勇气和公平性的某些主观特征（例如心理倾向和经验）。反过来，只有当智能体也理解这些主观特征如何与其自身主观性（例如感觉运动经验、信念、欲望和意图）相关时，这些特征对该智能体而言才可能有意义。道德谈论还包含"厚概念"的典型样例，它们兼具描述性和评价性（Blackburn，1998）；例如，"慷慨"和自私 等词语既可以描述某些行为（分享或不分享食物），同时又表示一种评价性品质（分别是好 与坏）。

遗憾的是，厚概念和厚描引出了长期以来一直处于形而上学和元伦理学争论核心的问题。要阐明厚描的意义，尤其是当它们依赖主观性时，或许还需要妥善回答其他一些根本性难题，例如符号奠基问题（Harnad，1990年）、意识的困难问题（Chalmers，1997年），以及关于如何确定道德术语中"意义"与真的意义这一元伦理学问题。第一个问题与任何使用符号的系统都有关——例如，用符号进行交流或推理的系统——因为它关乎符号如何获得意义。四十多年来，塞尔的中文房间实验（Searle，1980）一直为关于以下问题的哲学讨论提供了场域：

计算系统是否能够以及在多大程度上能够“理解”。Searle 的著名论点是，能够让一位中文使用者相信自己理解中文的计算机程序，并没有真正“理解”中文，因为它只是在遵循句法规则。但即使我们反对 Searle 对计算主义的广泛否定，[103]我们仍然需要说明厚重道德术语如何获得意义，或者更务实地说，它们对我们的道德实践“起到”什么作用。然而，符号的根基在多大程度上依赖于意向性（Brentano，1874年）、知觉经验（Barsalou，1999年）或某些感觉运动能力（塔代奥和弗洛里迪，2005年），仍是一个悬而未决的问题；而这些能力能否由计算系统实现，则是另一个尚无定论的问题。[104] 同样，要阐明厚概念，自然必须假定其评价性成分与描述性成分可以区分开来。[105]

那么，这些更深层的语义学问题与计算复杂性有何关联？当然，这最终取决于我们在这些争论中持何种立场，因为这将决定我们如何看待道德语义学，以及道德术语在道德行为中发挥的作用。一方面，如果我们遵循元伦理认知主义，认为道德术语的意义可以得到恰当的剥离，并界定为真值承载实体，那么，任何以“厚实”道德语言表达的问题，原则上都可以按照第5.2.2节讨论的方式进行计算（例如，检验其相对于某个给定或所有可能的克里普克结构的有效性）。因此，道德语义学不会带来任何超出道德语言表达能力本身所蕴含范围的额外复杂性负担。然而，尽管认知主义者认为道德术语能够表达关于世界的、独立于心灵的事实，但这并不意味着我们已经找到了这些事实；正如神命论一样，这种进路预设

[103]计算主义是一系列观点的统称，主张人类心智——包括意识和认知——是某种形式的计算。[104] Aaronson（2013年）从复杂性考量出发，提出了相关“瀑布论证”的一个技术版本，以补充中文房间论证；该论证声称，意义总是相对于某个外部观察者而言的（Searle，1992年）。论证始于这样一种观察：任何具有足够大状态空间的物理系统，原则上都能实现任意其他系统的语义；例如，对于某种从瀑布初始状态 I 到最终状态 F的映射 M，总能以某种方式标记任意给定的排列 P，使其从 I 和 F 出发，从而 M 实现 P，因此 P 或许可以表示我们想要的任何“语义”，例如国际象棋程序。然而，如果我们真的试图用瀑布来计算国际象棋着法，就需要找到一条从国际象棋程序到瀑布的归约，例如，说明如何将国际象棋局面和棋步高效地追踪到瀑布的初始状态和最终状态。由此，Aaronson推测，对于任何一个能够访问瀑布预言机的国际象棋程序，都存在另一个性能同样出色、资源需求相近，却不访问瀑布预言机的国际象棋程序。换言之，任何从国际象棋到瀑布的归约算法，极有可能只是在解决国际象棋问题，而没有以任何有意义的方式利用瀑布。对Aaronson而言，这反映了完备性这一更实质的概念：问题被归约到的类别本身不能足以解决同一个问题。也就是说，尽管借助 X_1的预言机，NP问题 X_1 可以在多项式时间内解决，但P类中的问题不能归约到P类中的问题（否则就意味着P类中的每个问题都是P完全的）。因此，瀑布计算与国际象棋计算之间所谓的等价关系并无多少实质内容，除非能够在一种本身无法解决瀑布问题或国际象棋问题的计算模型中证明这种等价关系。

[105]粗略地说，这大概就是布莱克本（1992）和黑尔（1952）的立场。相比之下，像普特南（2004）和威廉斯（2006）这样的思想家认为，厚概念是事实与价值不可分割的结合体。例如，普特南指出，“厚”伦理概念“根本不理会所谓的事实/价值二分，并欣然允许自己有时用于规范性目的，有时用作描述性术语”（Putnam, 2004, p. 35）。

人们已经收集了所有道德术语的详尽知识——例如，通过接触柏拉图式实在——并成功将其连同所有语义关系一并编码。[106] 然而，就目前而言，人们对道德术语的意义仍存在严重分歧。[107]

一种怀疑论的替代方案是接受不可判定性的事实。例如，赖斯定理指出，图灵机识别的语言的所有非平凡语义性质都是不可判定的；语义性质是关于程序行为的性质，而非平凡性质既不对所有部分可计算函数都为真，也不对所有部分可计算函数都为假。[108] 举例来说，我们可以把一种道德语言 ML 想象成一个由图灵机描述组成的集合，而图灵机属于 ML 的标准是，其语言 L(TM) 至多接受三个字符串（$|\text{L(TM)}| \leq 3$）。如果一种语言 M_1 属于 ML（$M_1 \in$ ML），这意味着它 M_1 满足性质 ML。ML 是图灵机语言的一种性质，这一点可以通过以下事实加以证明：具有相同语言的两台机器， M_1 和 M_2，要么都属于 ML，要么都不属于 ML；由于它们的语言相同，它们所含的字符串数量也相同。为证明 ML 是非平凡的，我们令 M_3 为一台接受所有字符串的机器，并令 M_4 为一台拒绝所有字符串的机器。由于 $M_3 \notin$ ML（接受的字符串超过三个），而$M_4 \in$ ML（0 为 ≤ 3），根据赖斯定理，ML 是不可判定的。当然，我们可以让 ML 表示许多不同的事物；鉴于赖斯定理的广泛适用性，我们可以证明与机器行为有关的一大类问题具有不可判定性。[109]同理，对于道德语言的前景，我们也可以得出怀疑论的结论：只要人们对道德术语的意义存在分歧（例如，不同智能体计算出不同的输出），就不可能存在可判定的道德语言。

当然，还有许多其他显而易见的理由使人对道德语义学的可判定性持怀疑态度。例如，可计算性和意义理论都面临同一个共同的克星：自指性的说谎者悖论——“这句话是假的”。它在哥德尔第一不完备性定理中发挥着核心作用——将“假的”替换为“不可证”——因为任何相容的数学系统都无法证明关于其自身的真命题。同样，图灵停机问题表明，无法判定一个计算机程序是否停机，因为对于任何能够对“停机”作出“是”判断的程序 P，都可以构造另一个程序，将P 作为输入，从而给出相反的“否”判断。[110] 在意义理论方面，阿尔弗雷德·塔斯基发现，说谎者悖论只会出现在“语义封闭”的语言中，即能够表达自身句子之真的语言（Tarski，1944）。塔斯基本人提出的解决方案——在一个构造出的层级中，将指称的元语言与被指称的对象语言分离开来——后来又被克里普克（1976）发现是不完备的；克里普克还运用自指技巧构造出了一些打破该层级的陈述，等等。[111]

[106]当然，对于极其有限的状态空间而言，这并非不可行。例如，我们可以设想一台烤面包机，它利用传感器读取吐司是烤得不够、恰到好处还是烤焦了，并理解评价句“这片吐司很糟糕！”如何表达有关吐司状态的某些事实。[107]回想上一节提到的约根森困境，我们还注意到，道义逻辑中关于有效推理存在严重分歧。[108]反过来说，如果一个部分函数对所有部分可计算函数都成立，或者对其中任何一个都不成立，那么它就是平凡的。[109]例如，可以很容易地运用赖斯定理证明，判断给定图灵机是否计算常值函数、全函数、恒等函数、两个自然数之和，或某个可计算函数，是不可判定的。[110]进而，可以通过从停机问题归约来证明更一般的赖斯定理。[111]克里普克给出了以下例子，该例由琼斯和尼克松表述：(*J*)“尼克松的大多数（即过半数）.....

当然，克里普克的解决方案——其利用了部分定义的真谓词（“未定义”）——反过来也可能遭到一种强化说谎者悖论的攻击：“这句话要么是假的，要么是未定义的”。归根结底，任何解决方案似乎都会产生新的自指问题，而这些问题又会针对该新方案，因此，任何以真为核心的语义理论都被一个伺机复仇的悖论所纠缠（Beall，2007）。此外，正如Dahl（2022）所论证的，说谎者悖论并不满足于针对建立在真之上的语义理论，而是会扩展到所有试图为任何语言提供统一意义解释的理论。简言之，既然任何统一意义理论都要求一种足以断言自身意义的语言，而任何语言都无法在阐述统一理论的同时自洽地为自身赋予意义，那么普遍意义理论就是不可能的。

一种更务实的做法是，否认道德术语有真假之分，转而遵循维特根斯坦（2010）的观点，认为“意义即使用”。这不仅能将道德语义学从柏拉图主义和悖论中剥离出来，还可能使道德话语更贴近其发源于其中的多元社会实践；在这些实践中，偶尔发生争吵、误解和意见分歧在所难免。依照这种观点，道德表达不应依据任何意义的形式说明来理解，而应看它们如何以或多或少令人满意的方式服务于我们的道德实践。倘若我们承认任意程序执行“这个或那个”并无不妥——从务实角度看，它们要么运行良好，要么运行不佳——那么这种做法甚至或许能够规避非平凡语义学和停机问题的不可判定性。我们可以不采用真值条件，而采用一种或几种著名的道德语言非认知主义进路，例如认为道德陈述的功能是表达情绪并引发他人的情绪（Stevenson，1937年）、提出规定性判断（Hare，1952年），或传达态度（Schroeder，2010年）。[112]然而，这类理论或许适用于情感丰富、善于判断且富有感受力的人类，但不太适合进行复杂性分析，因为它们往往预设了人类特有的心理学。

尽管如此，分析潜在的“无心”智能体使用道德术语时所涉及复杂度的一种方法，是考察 David Lewis 和 Robert Brandom 等现代维特根斯坦主义者。在研究反事实之前，Lewis 就对社会惯例进行了早期的博弈论分析（Lewis，1969）。在这一观点中，沿袭Schelling（1960）的思路，语言惯例和道德惯例都可以被视为反复出现的协调问题的自我延续型解决方案，因为利己行动者协调其行动能够互惠互利。更具体而言，语言意义随后在信号博弈这一范式下得到了探索。在一个简单的信号博弈中，发送者试图让接收者相信自己属于某种类型——例如，自己有能力——而实际类型只有发送者自己知道。直观地说，诚实的发送者不会因被误解而获益（例如，传达关于自身类型的虚假信息）；相反，说谎对能力不足的博弈玩家有利，而玩家的收益取决于接收者采取的回应行动（例如，雇用智能体 a 或 b）。鉴于其解释力，信号博弈已被用于建模交流和语言意义的发展（Skyrms，2010；Huttegger，

关于水门事件的断言是假的”，（N）“琼斯关于水门事件所说的一切都是真的”（克里普克，1976年，691）。在塔尔斯基层级中，N 需要处于高于琼斯所说一切内容的层级，而 J 需要处于高于尼克松所说一切内容的层级。[112]应当指出的是，这些非认知主义立场不应与维特根斯坦式意义观本身等同，而应理解为它们拒绝这样一种信念：道德术语具有真值。

2007年）。

由于信号博弈通常被建模为序贯贝叶斯博弈，其均衡解难免受到第5.1.5节所讨论的不可处理性问题困扰；尤其是，判断一个贝叶斯博弈是否存在纯策略贝叶斯均衡是NP困难的，而在马尔可夫博弈中计算纯策略纳什均衡则是PSPACE困难的（Conitzer和Sandholm，2008年）。然而，Lewis提出的均衡概念，即协调均衡，具有这样一种特性：只要至少除一名以外的所有博弈参与者都遵循某种规律性R，每位参与者也都更愿意其他所有参与者遵循这种规律性。当然，这预设了某种共同知识概念，后来Aumann的相关均衡将其加以泛化（Aumann，1974年）。如第5.1.5节所述，相关均衡在计算上很有吸引力，因为对于任意数量博弈参与者的博弈，都能在多项式时间内找到相关均衡（Gilboa和Zemel，1989年）。[113] 然而，在Lewis最初的定义中，惯例必然具有任意性，即存在一种相冲突的规律性 R'，它本来也可能成为稳定惯例（而且 R 和 R' 互不相容）。当然，任意性这一概念确实具有解释价值：互惠有利的协调问题解决方案成为惯例并非任意之举（例如，汽车在道路的某一侧行驶），但选择某个特定方案而非另一个方案则是任意的（例如，选择右侧通行而非左侧通行）。不过，任意性标准并未排除这样一种可能：相冲突的规律性或许能带来更大的共同利益。因此，既然稳定惯例会自我延续，其潜在的道德价值便依赖于乐观保守主义——这种立场也贯穿于关于利他主义和社会契约的事后进化论解释之中；也就是说，惯例之所以如此，必有其合理之处。相反，某种惯例可能具有极高的无政府代价（即极糟糕的最坏情形均衡），而惯例的稳定性却会确保无人相信违背惯例会让任何人获益。然而，将惯例与其他方案进行比较，例如通过计算一个能使预期共同利益最大化的相关均衡，会引入NP困难性（Papadimitriou and Roughgarden，2008年）。

从更乐观的角度看，交流中的合作原则或许能够缓解这些计算难题，例如遵循 Grice（1975）的研究。这一策略体现在 Lewis 对惯例的精细分析中（Lewis，1975）；该分析纳入了与个人自身信念有关的偏好："遵循惯例的预期通常会让每个人都有充分理由使自己也遵循惯例"（Lewis，1975，第8页）。同样，Lewis 写道："当且仅当某个群体 P 中存在一项诚实与信任惯例，且该惯例由交流利益所维系时，L 语言才会被该群体使用"（第10页）。对 Lewis 以及 Schiffer（1972）和 Bennett（1976）等其他格赖斯学派学者而言，意义最终源自说话者的交际意图与接收者的交际预期之间的协调。反过来，这些意图和预期也可能包含合作惯例。

Brandom（1994）提出了另一种与"意义即使用"相关但有所区别的阐释，适合用于解释道德语言。Brandom 基于话语实践这一概念，提出了关于智慧的理论——即人类所拥有的理性类型——该理论可以

[113]另请参阅 Urbano 和 Vila（2002）的研究，该研究展示了如何通过对未经中介的交流施加计算限制来实现相关均衡。

概括为“给予理由和要求理由的博弈”（第6页）。话语实践的参与者承担权利与承诺，前者和后者分别可被视为具有许可与义务的规范力量。话语实践的核心是推理关系，这些关系将承诺与权利保留至其他陈述之中。也就是说，句子只有就其功能而言才承载内容，而该功能是在与其他句子的关系中推导出来的。就道德话语而言，有趣之处在于，语言表现的特征在于其能够改变话语实践成员的规范地位。这体现为一种对话式“记分”，参与者在对话语境中追踪承诺与权利，例如通过作出、承认、质疑或撤回断言。[114] 从本质上说，语言表现的规范语用学决定了语义学的推理层面，而不是反过来。然而，尽管 Brandom 的推理语义学或许有助于阐明言语行为的规范承诺，但我们只能推测成功参与话语实践所依赖的计算层面，以及这种参与是否、在多大程度上缓解或增加了计算要求。一方面，给予理由和要求理由的博弈似乎旨在有效促进规范生活中的合作性沟通。另一方面，Brandon 对智慧的观念似乎具有深刻的人类特征；除了与非语言动物共有的感知能力之外，它还涉及对概念内容的理解，而概念内容又可能涵盖自身及他人的意向状态、信念和欲望。相应地，智慧主体的理由博弈或多或少可以对应于一种康德式的道德理性主义话语观（Brandom，2006），这种观念预设了复杂且理想化的道德自主与自由能力；而这些能力——至少在目前——还无法在计算模型中得到构造。

主要的启示似乎是：尽管交流或许能够缓解博弈论问题——包括计算不可处理性和令人厌恶的均衡所带来的威胁——但挑战不仅在于说明交流如何运作，也在于解释它为何如此有效。对 Lewis 而言，交流之所以成功，似乎依赖于说话者的交际意图和预期，以及由此产生的合作惯例。对 Brandon 而言，关键则在于有理性者参与的“给出理由并要求理由”的博弈。不幸的是，这使两种路径在计算层面都相当晦涩：由于它们旨在解释人类交流，因此和非认知主义理论一样，可以诉诸独特的人类特征，而这些特征从计算视角看来仍在很大程度上难以理解。不过，这类理论仍有可能带来显著价值，因为它们或许能弥合这样一道鸿沟：一边是促进高效合作的认知心理资源——例如理由、信任和交流；另一边则是可能使这些资源得以实现的计算架构。反过来，复杂性考量或许有助于阐明日常道德互动中，博弈论动态、社会心理能力与规范理论之间丰富的相互关系。

总而言之，本节讨论了道德语义学中更为深刻的问题。这些问题虽然在机器伦理学中相对受到忽视，却仍处于元伦理学和元语义学争论的核心位置，也同样是交际理论关注的重点。当然，一个简单的做法

[114]Lewis（1979年）首次提出了对话中的记分这一概念；在 20 世纪 70 年代，他曾担任 Brandon 博士论文的导师。

是坚持某种反对“强人工智能”的形而上学论证（Searle，1980）、某些关于符号 grounding 的高级要求（塔代奥和弗洛里迪，2005），或某种人类独有的情绪心理学，并由此得出结论：“机器永远无法真正理解道德语言。”但即便决定忽视这些问题，计算道德语义学仍面临许多复杂的问题。

5.3. 结果主义—义务论混合理论

在考察了结果主义和义务论各自的一系列计算层面之后，我们现在可以对二者的差异及其潜在结合做出更为实质性的论述。首先，应当得出这样的结论：“从计算角度而言，义务论比其替代理论的复杂度更低”这一主张，无法得到简单直接的回答；而且在许多情况下，这一主张根本就是错误的。基于本节所讨论的材料，我们可以勾勒出一个更为细致的回答：

（1）正如第5.1.1节所论证的那样，把在法律和自由主义语境中的规则遵循视为计算上简单，是一种错误。尽管法律规则可能有效地压缩大量的道德智慧，但这种有效性只有在以下复杂关系的背景下才成立：一方面，是激励人们遵守这些规则的机制（例如警察和惩罚）；另一方面，是确保对这些规则进行公正解释的机制（例如法院）；此外，还包括法律所适用对象的道德情感。例如，在自由主义语境中，“不得伤害他人”原则的复杂度，只有联系完全自主公民的能力来分析，才具有真正的意义。

（2）另一方面，如果义务论规则以神命论或法律实证主义为依据获得辩护，其效率可能显著高于其他方案。然而，除非是在极度简化的情形下，这类方法若要在实践中奏效，所需的知识量仍然大得难以想象，而这些知识本身也可能极具争议。

（3）另一种方案是，将辩护纳入道德计算，作为义务论规则遵循的一部分，例如，考量遵循规则的行为会如何影响他人。如 5.1.2 节所述，这可以涵盖从（i）个人偏好（“我按照自己个人的偏好对待你”）、（ii）他人偏好（“我根据自己对你偏好的了解来对待你”），到（iii）一般行为（“我通常以自己希望他人通常对待我的方式对待他人”）等不同考量。较弱的版本——例如（i）和（ii）——从复杂性视角来看可能相对简单，但也会引发一系列其他问题。相比之下，较强的版本——例如（iii）的契约论或契约主义版本——则会因行为预期以及构成一般行为观念的能力存在巨大差异而变得晦涩难明。

(4) 任何旨在解释多主体动态的一般规范理论，都面临博弈论问题。这包括理性问题（5.1.3）、不完全信息与递归推理（5.1.4），以及计算具有道德吸引力的纳什均衡时遇到的不可处理性（5.1.5）。

(5) 任何采用形式逻辑的计算系统，例如用于遵循义务论规则、道德推理或交流，都会受到可表达性（5.2.1）、不可处理性（5.2.2）和可判定性（5.2.3）问题的制约，而这些问题贯穿逻辑的句法与语义学。

（6）最后，我们应当承认，义务论是一个涵盖多种伦理学理论的体系，而这些理论又可能强调不同的认知能力和计算资源。它可以是一项道德理性主义计划，旨在以共同自主性和理性为基础，寻找普遍可辩护且适用的规则。它也可以是一项

契约论项目，旨在基于自身利益寻找互惠互利的行动规则。在最简单的情况下，规则可以充当“快速而节俭”的启发式方法，帮助认知能力有限的智能体在复杂或新颖的情境中采取任何行动，即使无法判断所采取的行动在特定情境中是否恰当；这种行动可能只是自动的“默认”行动，也可能依赖乐观保守主义（“过去效果不错，将来也可能如此”）。在更为宏大的情况下，义务论规则遵循可能要求人们为存在不完全信息的大规模协调问题计算道德上理想的均衡。

基于这些考量，比较义务论与结果主义的复杂度可能会产生误导，因为任何比较都取决于我们如何理解这些理论所规定的内容。例如，只要两种理论都将其他行动者纳入考量，它们便同样受到博弈论不可处理性的影响（5.1.3）。同样，困扰逻辑系统的许多更深层的道德语义学问题（5.2.3），也可以转而用于考察结果主义智能体：例如，计算系统“理解”某种效用究竟意味着什么？另一个原因是，义务论和结果主义可能得出相同的解决方案，比如在义务论所认为的道德正当行为同时也能产生最优结果的情形下。因此，一种合理的替代观点是将它们视为互补理论，分别强调不同的认知能力、计算资源或伦理生活层面。

因此，混合理论——结合结果主义与义务论的某些方面——广泛见于道德哲学（Brandt, 1979年；Hare, 1981年）、道德心理学（Kahneman, 2011年；Greene, 2007年）和机器伦理学（Bauer, 2020年；Arkin, 2007年；Dehghani 等, 2008a；Stenseke 和 Balkenius, 2022；Pontier 和 Hoorn, 2012；Azad-Manjiri, 2014年；Tufiş 和 Ganascia, 2015年；Govindarajulu 和 Bringsjord, 2017年；Pereira 和 Saptawijaya, 2009），这并不令人意外。然而，值得注意的是，几乎在所有情况下，义务论在这些架构中都扮演着相当狭窄的角色——通常处于更广泛的结果主义框架之内——这并未体现其在伦理理论中的深厚传统。例如，义务论可以充当先验约束，阻止某些内在有害行为（Pereira 和 Saptawijaya, 2009；Dehghani 等, 2008a ）。另一种做法是运用义务论规则来提升运行时效率，例如，将结果主义计算或探索转化为可利用的规则（Stenseke 和 Balkenius, 2022）。[115] 同样，在道德认知的双加工理论中（Greene, 2007年），义务论判断被视为快速且本能的反应，而结果主义判断则指缓慢且审慎的推理过程。

关于复杂度，一个新兴趋势是，道德规则遵循可以：（i）用于防止内在有害行为；（ii）降低道德计算（通常是结果主义计算）的运行时复杂度；或（iii）在智能体没有其他依据来做出决策的情境中，充当“原则优先”或“可废止的”默认模式。从道德心理学的角度看，（ii）和（iii）可以由这样一种观点加以支持：自动道德判断压缩了道德智慧，例如，这种智慧可能源于促进合作的进化适应、培育集体福祉的文化特定规范（如显著的相关均衡），或个人通过

[115]由于后果会促使人们采用某些规则，因此，将这一解决方案视为双层功利主义的一种形式更为合适（Hare，1981年）。

经验学会并内化的道德教训。对机器伦理学家而言，后果主义—义务论架构因而提供了颇具吸引力的情境特定模块化组合；其范围从基于既有道德知识的自上而下先验约束（神命论），到自下而上的后验经验，再将其转化为新的自上而下约束（例如，通过经验学习高效规则）。然而，这种看法并非毫无缺陷，因为运行时复杂度的任何降低都意味着存在以下一个或多个问题：

乐观保守主义的问题——从本质上说，无法保证“快速且直觉式”的道德行为会产生良好的道德结果，除非将其与其他选择进行评估（例如，依据经验或道德推理）。因此，直觉的道德价值取决于一种乐观保守主义：遵循直觉式义务论规则（或规则结果主义规则）所带来的效率，可能要以错失更好结果为代价，而这种结果只有通过进一步反思才能实现。遵循 Hare（1981）的说法，这正是如何判断何时该像“平民”一样思考（运用道德直觉），何时该像“大天使”一样思考（进行批判性反思）的难题。从计算视角来看，这体现了一个久经研究的探索-利用困境（4.3）。它也令人联想到规范和惯例所存在的问题重重的事后理由。如 5.1.3 和 5.2.3 所述：即使我们认为某种道德惯例——或规则、规范、原则——之所以形成，是因为它帮助有限认知的智能体协调行动，以实现互利目标，这一惯例的稳定性也可能取决于对惯例本身抱有的乐观保守主义，即便可能存在更好的替代惯例。同样，某些道德直觉可能源自曾使我们的进化祖先获得繁殖成功的适应性，但这并不意味着这些直觉在当今就必然是道德上正确的（Greene，2014年）。

速度与性能问题 ——另一个相关问题是如何在道德决策速度与道德表现之间做出选择。例如，即使我们像 Hare（1981年）和 Greene（2014年）那样，认为缓慢的功利主义审议在规范和认识论上应具有优先性，道德决策速度与道德表现之间仍存在悬而未决的冲突。在什么情况下，我们会为了更慢但可能更好的选项，而选择快速的——例如可行但非最优，或令人满意的——选项?

推理与学习问题 ——“快速且直觉式”的道德直觉或许能提高运行时效率，但这是以忽略产生道德直觉的过程本身的复杂度为代价的。一方面，这可能指 Hare（1981年）所说的道德思维“批判层次”这一过程的复杂度（它支配着“直觉层”的原则）。另一方面，我们也可以纯粹从学习过程的角度来思考：例如，智能体需要多少学习样本和多少训练时间，才能成功内化一种道德直觉，使其能够在运行时有效解决特定的道德问题？原则上，既然人工神经网络可用于近似任意函数（Scarselli 和 Tsoi，1998年），那么只要采用恰当的学习方式，机器学习系统就应能在道德问题上实现出色的运行时性能。然而，正如我们将在第 6 节看到的，这反而会不断增加对样本复杂度和训练时间复杂度的要求，同时引发其他与归纳有关的问题。

知识问题 — 高效的道德直觉无需通过推理或学习获得，而可以建立在道德知识的基础上。例如，对于道德实在论者和元伦理认知主义者而言，一种可行的替代方案是收集并记录客观道德实在的训示；他们相信，道德直觉能够指称反映世界主体无关特征的真命题。当然，正如前文已经多次指出的那样

在本文中，这样的项目出于各种理由似乎存在严重问题。尽管如此，这并不排除这样的可能性：某些 道德直觉具有普遍（或近乎普遍）的共识，或者有一些道德直觉——沿用 Asimov（1942年）的说法——对计算代理尤其具有吸引力。

总而言之，如果我们仅将义务论视为“规则遵循”，那么从有限理性的角度来看，它显然能够产生高效的决策程序。然而，尽管它或许能带来出色的运行时性能，这种效率的道德价值却要么建立在乐观保守主义之上，要么源于其他某种复杂过程（推理或学习）的结果，抑或依赖于大量道德知识的积累。从本质上说，规则的道德力量——其普遍适用性、普遍辩护能力和计算简洁性——只有通过复杂的方式才能得到保障。毕竟，道德规则的简洁性不应妨碍人们质疑其权威；相反，它应帮助人们记住，规则和逻辑系统一样，只不过是对错综复杂现象某些理想化方面的呈现。

6. 美德伦理学与道德机器学习

义务论以行动为中心，结果主义以结果为中心，因此很自然可以将二者理解为道德决策程序。相比之下，美德伦理学关注的是存在而非作为：它不再聚焦于何为正当的行动，或何种行动能带来最佳结果，而是要求我们培养自己的道德品格，以及勇气、公平性和节制等内在倾向——这些美德使我们成为道德高尚之人。更广泛地说，美德伦理学指的是一大类强调道德品格作用的伦理传统；这一理论的多元源流可追溯至东方的孔子、孟子等思想家，以及西方的柏拉图、亚里士多德等思想家（Crisp 和 Slote，1997年）。在当代，得益于安斯康姆（1958年）、Nussbaum（1988年）、Hursthouse（1999年）和 Annas（2011年）的贡献，美德伦理学已成为英语世界道德哲学中的核心规范理论。

在机器伦理学中，美德伦理学曾多次被提出作为创建人工道德主体的适当蓝图，因为它强调了其他理论相对忽视的伦理生活层面（Coleman，2001年；Wallach 和 Allen，2008年；Howard 和 Muntean，2017年；Stenseke，2021年）。[116] 它将品格置于道德评价的核心——品格既涵盖理性审议，也涵盖心理倾向——从而更全面地描绘了何谓道德；它不仅捕捉我们理想上应当做什么，也捕捉促使我们以值得道德赞许的方式行动的动机。这幅图景的一个关键方面是实践智慧（“实践智慧”）这一概念，它可以被理解为智能体通过实践和经验习得的道德智慧或技能（Annas，2011年）。对发展和学习的关注进而启发机器伦理学家将美德伦理学与联结主义统一起来；既体现在现代机器学习方法上，也体现在作为更广泛认知理论的联结主义上（Casebeer，2003年；Wallach 和 Allen，2008年；Howard 和 Muntean，2017年；Berberich 和 Diepold，2018年）。乐观地看，凭借持续从经验中学习、对语境保持敏感并适应变化的能力，有德性的机器或许能够把握动态环境中人类规范的复杂性，而在这类环境中，单纯的效用最大化或规则遵循会失效。

[116]有关美德伦理学计算实现的最新综述，请参见 Stenseke（2022a）。

尽管有这些承诺，从计算视角来看，美德伦理学仍相对难以捉摸。原因之一在于，美德伦理学建立在一种深具人本色彩的道德生活观之上——关于兴盛、推理和情绪的观念——其背后是丰富的文化与传统。显而易见，这使得从复杂性视角分析美德伦理学十分困难，因为这要求我们对人的构成作出全面的计算描述，同时还要涵盖其全部历史和具有文化特色的差异。不过，我们仍然可以分离并分析任何有德行的智能体都必不可少的一个方面：它的学习能力。原则上，迄今讨论过的任何问题都可以重新表述为一个学习问题。我们不再问"我能否有效地解决问题 X？"，而是问"我能否学习如何有效地解决 X？"。如果答案是肯定的，那么在训练得当的情况下，实际运行时复杂度可能会低得微不足道。直观地说，如果某个具体问题已经得到解决——例如，已经发现某种情境下的最优行动组合（4.1）、某个环境中的最优强化学习策略（4.3），或某种策略情境中的行动规则（5）——那么，只要完全相同的问题再次出现，同一解决方案就可以在常数时间 $O(1)$ 内应用于未来案例。即使完全相同的问题不再出现，也可能存在一些模式或趋势，可以从足够相似的情境中进行外推。通过反复光顾沙拉吧（2），我们或许能够学会食材的良好组合。看到相同的食材出现在其他沙拉吧中，我们就可以利用先前的经验，高效地搭配出美味的沙拉。遵循美德伦理学理论对道德学习的强调，道德行为也应当如此。用亚里士多德的话说："[.....] 找不到一个具有实践智慧的年轻人。原因在于，这种智慧不仅关注共相，也关注殊相；而殊相会通过经验变得熟悉"（NE 1141b 10）。

反过来说，让机器学习是机器学习的研究领域。得益于近年来的大量进展，它已在 21st 世纪成为人工智能领域的主导力量。对于我们的研究——也是本节的主题——相关的问题是：计算复杂性考量是否会限制计算主体在道德领域中的学习能力？

6.1. 学习的复杂度

与任何成熟的数学分析领域一样，有许多相关的变体、设定和度量，可用于正式分析学习的复杂度。[117]首先，我们可以区分运行时复杂度（算法在运行时解决问题 X 所需执行的状态转移次数）、训练时间复杂度（训练算法解决 X 所需的步骤数）和样本复杂度（学习如何解决 X 所需的数据点数）。尽管如此，在实践中，这些类型之间的区别可能会变得模糊；例如，如果一个样本代表一个时间点，那么样本复杂度就等于训练时间复杂度；同样，我们也可以把训练时间复杂度理解为 学习算法的运行时复杂度。不过，重要的是要注意，这三种类型紧密相关。在完全已知的环境中（比如国际象棋），模拟（例如使用蒙特卡洛方法）可能是获取经验的有效方式，比如尝试许多不同决策的价值

[117]关于统计学习理论的权威性介绍及其丰富历史的全面论述，参见 Vapnik（1999）。

。在这种情况下，样本复杂度衡量的是找到一个好的国际象棋棋步所需的模拟量。如果样本复杂度较低，就意味着能够有效地学习，因为它给出了总计算复杂性的下界（Kakade，2003年）。在其他情境中，可能有大量数据和充足的时间来训练模型，而任务本身可能指的是某种一般能力——例如通过图灵测试来评估——而不是系统如何解决某个具体的判定问题。GPT-3（Brown 等，2020年）等大语言模型利用互联网的大量内容作为训练数据来生成新文本，是这类情况的恰当例子。在最极端的情况下——例如在线强化学习设定中——如果智能体对环境一无所知，也无法访问环境的模拟，试错探索可能就是学习的唯一途径。

我们对复杂度资源的选择，最终取决于学习智能体所处的情境，而每种情境都有各自的一系列具体子类型和相关度量。为此，通常会将机器学习情境大致分为三类：(i) 监督学习（SL），其目标是从预先标注的数据中学习；(ii) 无监督学习（UL），其训练数据没有标签；以及 (iii) 强化学习，其学习基于反馈。对于一个 SL 智能体，相关的问题可能是：为了在一定误差范围内准确判断新的图片中是否有猫，我需要看多少张带标签的猫的图片？从德性理论视角看，这可以改述为：为了将未来未观察到的行动分类为勇敢行为，我需要观察多少个带标签的“勇气”样例（例如勇敢行为）？在这种情况下，我们可能只关心样本复杂度——训练集中描绘勇气的图片数量——如何降低智能体在分类任务中的预测误差（例如，通过经验风险最小化）。对于一个 RL 智能体，问题则可能是：为了近似道德典范（例如有德之人）的奖励函数，我需要观察其整体行为多少？[118] 在这里，考虑智能体学习阶段可能发生的有害错误风险也很重要。如（4.3）所述，如果试错探索是唯一的选择，那么 RL 智能体的目标可能不只是最大化累积奖励，还包括相对于最优解最小化遗憾。

在探讨某事物能否被有效学习之前，我们首先需要确定它是否能够被学习。例如，设 X 为点的集合，Y 为标签的集合（如 0 和 1），$\mathcal{H}$ 为假设的集合 h（如二元分类器），这些分类器接收一个 $x \in X$并输出一个标签 $y \in Y$。我们的学习算法 $\mathcal{L}$ 的目标是：给定从某个分布 ρ 中随机抽取、定义在 X上的一系列带标签的训练样本（x，y），推断出一个假设 $h \in \mathcal{H}$，使其能够以一定的准确性 ϵ 和失败概率 δ正确分类 $x \in X$ 的未来实例。进而，如果存在一种学习算法，能够在给定有限数量的训练样本（x，y）$_n$的情况下，以至少 $1-\delta$的概率，将输入映射为与最优结果相差不超过 ϵ 的输出，那么我们就称假设空间 $\mathcal{H}$ 是可学习的。[119] 若满足这一条件，样本复杂度可定义为 $n(\rho, \epsilon, \delta)$，这表示我们需要 n 个训练样本来

[118]这种特定方法称为逆向强化学习，由 Ng 和 Russell（2000年）提出，并由 Berberich 和 Diepold（2018年）建议将其作为迈向人工德性智能体的一条路径。[119]例如，遵循经验风险最小化，我们可以将最优的 $h*$ 定义为 $\mathcal{H}$ 中误分类风险最小的假设。

学习一个目标函数 针对分布 ρ、错误率 ϵ 和失败概率 δ进行校正。

6.1.1. 弱样本复杂度与强样本复杂度

基于这些定义，我们还可以区分样本复杂度的弱变体和强变体，例如，可以询问：对于某个特定的输入输出分布 ρ，我们需要多少样本才能学习目标（弱）；以及对于任意可能的分布，我们需要多少样本才能学习它（强）。按照强方法，通常称为无免费午餐定理（NFL）的不可能性结果表明，总会存在一些不利的分布，使得样本复杂度任意大（Wolpert，1992年、1996年；Schaffer，1994年）。作为一个直观的例子，我们可以设想一个学习算法，其目标是根据前几天的天气来预测天气——在此例中仅限于晴天（S）或雨天（R）。收集三天的数据后，我们有 2^3 种可能的天气历史（即 SSS、SSR、SRR、SRS，.....）。然后，我们将学习算法的错误 ϵ 度量为错误预测的比例（例如，在下雨天预测为晴天）。接着可以证明，每个学习算法都会在恰好一种天气历史中取得完美的 $\epsilon = 0$，在另一种天气历史中取得最差的 $\epsilon = 1$，在三种天气历史中取得混合结果 $2/3$，并且在剩余的三种天气历史中取得$\epsilon = 1/3$。无免费午餐定理表明，对于每个可能的 ϵ，每个学习算法在数量相同的可能天气历史中都能取得 ϵ。例如，如果我们为每一种可能的天气历史赋予相同的概率（即均匀概率分布），无免费午餐定理则意味着每个学习算法的预期 ϵ 都相同，均为 $1/2$。

NFL 的结果随后被扩展到优化（Wolpert 和 Macready，1997年）、监督学习（Wolpert，2002年）、统计学习（Von Luxburg 和 Sch¨olkopf，2011年）、元学习（Giraud-Carrier 和 Provost，2005年）以及数据隐私（Kifer 和 Machanavajjhala，2011年）。[120] 从本质上说，这意味着不存在一种学习算法，能够在经过固定大小的数据集训练后，在每一项学习任务上都表现出色。因此，对于每一种学习算法，都存在一项它无法胜任的任务，因为只观察到现实的部分实例时，任何学习算法都无法泛化到所有可能的现实。为此，NFL 也被用于讨论归纳哲学中更根本的问题，例如与休谟归纳问题（Sterkenburg 和 Gr¨unwald，2021年；Schurz，2017年）或奥卡姆剃刀（Lattimore 和 Hutter，2013年）的关联。休谟著名地对归纳本身的辩护提出怀疑，指出仅凭演绎推理无法确立归纳推断的有效性；而由于存在循环论证，归纳也无法为自身提供非演绎的依据（休谟，1739）。

当然，用归纳的根本问题来质疑道德学习系统的可行性，或许并不公正。虽然这些问题可能会阻碍完美通用道德学习器的前景，但并不妨碍我们寻求较弱但合理且在实践中可行的替代方案。我们不必追求一种全局且独立于模型的辩护，来解释归纳推断为何似乎有效；我们可以选择局部且相对于模型的辩护，以解释为何某些学习算法比其他算法效果更好（Sterkenburg 和 Gr¨unwald，2021年）。然而，应当强调的是，任何此类替代方案都不可避免地会包含某种形式的归纳偏置；这些假设是我们用来实现并促进可学习性的。一种缓解任意大的样本复杂度之 curse的策略，是约束概率分布的空间，例如，对数据所来自的分布结构作出假设，

[120]有关 NFL 定理的系统综述，请参阅 Adam 等人（2019年）。

这些数据点便是从该分布中抽取的（统计学中称为“参数化”程序）。最直接的参数化假设可见于中心极限定理：该定理指出，当我们将随机抽取的独立变量相加时，它们的和会趋近于正态分布。[121] 事实上，机器学习领域的大多数进展都依赖某种形式的参数化假设，例如使用线性回归和逻辑回归，以及人工神经网络的参数。[122]

另一种选择是限制假设空间。在归纳哲学中，这可以用奥卡姆剃刀来解释；奥卡姆剃刀大致指出，较简单的假设通常比复杂的替代方案更好。但“简单”是什么意思？要推断出一个能够准确预测的简单假设，我们需要多少样本？在计算学习理论中，可能近似正确（PAC）学习模型可以有效地回答这类问题。该模型由 Leslie Valiant（1984）提出。按照前文所述的形式化体系，我们考虑一个学习算法 $\mathcal{L}$，它希望根据从样本空间 X上的分布 ρ 中独立抽取的样本（x），在有限的假设集合 $\mathcal{H}$中学习一个布尔函数 f：$X \to \{0,1\}$。如果假设 h 能够以某个误差 $\epsilon > 0$近似 f，我们就称由 $\mathcal{L}$ 推断出的假设良好，其含义是：

$$\underset{x\sim\rho}{Pr}[h(x) \neq f(x)] \leq \epsilon \tag{5}$$

接着，如果对于任意选择的 ρ 以及所有失败概率 δ 和错误率 $\epsilon > 0$，$\mathcal{L}$ 都能以概率 $1-\delta$作出良好的近似，我们就称 $\mathcal{L}$ “可能近似正确”。最后，如果存在一个 $\mathcal{L}$，能够以样本量 n 作出“可能近似正确”的预测，且该样本量是 $1/\epsilon$ 和 $1/\delta$的多项式函数，那么我们就称函数 f 是 PAC 可学习的。换言之，PAC学习将这样一种观点形式化：如果一个函数是可学习的，就意味着存在一种学习算法，它在随机选取的数据上进行训练后，有合理的可能性获得合理的泛化误差，同时样本数量有一个多项式上界。

从计算复杂性视角来看，有趣之处在于，PAC 学习给出了样本复杂度的上界。如果目标函数是 PAC 可学习的，那么学习该函数所需的样本数量 n 可以通过以下方式推导：

$$n = \frac{1}{\epsilon} \ln \frac{|\mathcal{H}|}{\delta} \tag{6}$$

特别是，方程（6）阐明了预测误差、置信度、样本与假设空间之间的密切关系，以及它们在计算上的权衡。例如，它体现了一种直觉：通过增加样本数量，可以降低错误率 ϵ 和泛化误差 δ（尽管永远无法将其降至 0）。它还表明，要从有限数量的样本 n 中成功学习，我们必须限制假设空间 $\mathcal{H}$的基数；例如，可以减少单个假设的数量，或者

[121]从哲学角度来说，这意味着人们接受世界趋向于正态分布。[122]需要注意的是，参数足够多的深度神经网络可以视为非参数模型；例如，Lee 等（2017年）证明，无限宽的深度网络等价于一个非参数高斯过程。

降低它们的描述复杂性。[123] 相比之下，更大的 $\mathcal{H}$ 可能导致过拟合，也就是说，假设与某一组样本十分贴合，却无法泛化到额外的观测。

从哲学角度来看，PAC 学习颇具意义，因为它不受休谟怀疑论的影响，界定了一类丰富的实例，在这些实例中，归纳有保证能够奏效（至少很可能近似奏效）。不过，这一模型的一个重要权衡是，它只适用于有限假设类，这不可避免地要求在近似精度与学习算法的能力之间做出妥协。例如，包含连续参数的假设可能需要转换为离散参数。问题在于，我们应当把无限连续统划分得多精细？VC 理论（由 Vapnik 和 Chervonenkis 于（1974年、2015年）提出）可以解决这一难题。重要的是，与假设空间不同，VC 理论引入了更丰富的概念——VC 维度（VCD），它通过学习者能够打散的最大点集的基数，衡量学习算法的表达能力。例如，给定二维空间中的三个点，我们可以考察线性分类器 LC 能否正确地区分负类（标记为 $-$）和正类点（标记为 $+$）。具体而言，对于这三个点 $2^3 = 8$所有可能的标记方式，我们要问：能否画一条直线，将正类点与负类点分开？发现这是可能的——也就是说， LC 打散了包含这三个点的集合——之后，我们尝试对四个点做同样的事。由于不存在任何能被直线打散的四点集合，[124] 我们得出结论： LC 的 VCD 为 3。类似地，如果存在一个假设 $h \in \mathcal{H}$，能够与所有2^m 可能的 h（x_1） ...， h（x_m）配置相符，那么一个假设类$\mathcal{H}$ 就能打散这些点 $\{x_1, ..., x_m\} \subseteq X$。因此， $\mathcal{H}$ 的 VCD 就是满足以下条件的最大 m 的基数：其中存在一个子集 $\{x_1, ..., x_m\} \subseteq X$，能被 $\mathcal{H}$打散。进一步来说，得益于 Blumer 等人（1989年）的工作，VCD 与 PAC 可学习性通过一项重要洞见得以统一：有限 VC 维度是实现无分布学习能力的必要且充分条件。换言之，如果一个假设集是 PAC 可学习的，其 VCD 就是有限的。当然，尽管这并未消除归纳的根本问题，却提供了一个框架，用于描述归纳何时可行；它还提供了衡量表达能力的丰富指标（VCD），以及一种实质性的简单性概念（即，VCD 的最小值）。关于复杂度，Hanneke（2016年）最近证明，类别 $\mathcal{H}$ 具备 PAC 可学习性所需的最优样本复杂度[125] 为：

$$n(\epsilon, \delta) = O\left(\frac{VCD(\mathcal{H}) + \ln \frac{1}{\delta}}{\epsilon}\right) \tag{7}$$

然而，关于可行学习性的理论面临着与纳什均衡相同的问题：仅仅知道存在一个假设 h 属于 $\mathcal{H}$ 且与数据相容，并不一定意味着很容易找到它。因此，PAC 可学习性忽略了实际找到良好假设可能需要的大量计算。为此，

[123]Aaronson（2013年）将 PAC 学习的这一特性解释为奥卡姆剃刀奏效的数学辩护。[124]这是拉东定理的推论，该定理可用于推断d维点的线性分离的 VC 维。[125]这里，最优意味着上界与已知下界（Ehrenfeucht 等人，1989年；Blumer 等人，1989年）仅相差数值常数因子。

针对 PAC 已经证明了大量困难性结果。在适当设定下，学习者必须输出 $h \in \mathcal{H}$；Pitt 和 Valiant（1988年）证明，诸如两个单项式的析取（仅含一项的多项式）、布尔阈值函数，以及每个变量最多出现一次的布尔公式等表示类，无法被高效学习，因为它们可以归约到已知的 NP完全问题。基于广泛使用的密码学假设——例如 Rivest-Shamir-Adleman系统和 Blum整数——Kearns 和 Valiant（1994年）在非适当设定下（学习者可以输出任意 $h \notin \mathcal{H}$），针对一系列表示类证明了与表示无关的困难性结果，其中包括多项式规模布尔公式、常深度阈值电路和无环确定性有限自动机。[126] 此外，尽管非适当学习的困难性依赖于密码学假设，Applebaum 等人（2008年）指出，一旦证明成立，就意味着多项式层次[127] 要么会“坍塌”，要么任何 NP 中的平均情形困难问题都可以转化为单向函数（这将是密码学领域一项重大突破）。

除了适当学习与非适当学习之外，学习理论中另一个重要区别是可实现（或“无噪声”）学习与不可知（“有噪声”）学习之间的区别。在可实现情形下，假设在假设空间$\mathcal{H}$ 中存在一个最优假设 $h*$，其满足 $\epsilon = 0$。在不可知学习中（Kearns等人，1992），我们不对目标函数作任何假设；我们只想从某种分布中找到尽可能好的 h。[128] 统计机器学习中的典型问题包括半空间（或线性阈值函数）、[129] 单项式和决策列表的学习。在适当的可实现情形下，通过线性规划可在多项式时间内完成半空间学习；而在适当的不可知情形下，其NP困难性已通过多种方式得到证明。[130]

如 4.3 节所述，强化学习也存在类似的计算难度（Mundhenk 等人，2000；Papadimitriou 和 Tsitsiklis，1987）。事实上，有许多理由表明，强化学习比监督学习设定难得多：学习者可能无法从环境中获得训练集；学习者可能只能收到“带噪声”的即时奖励（这可能会误导智能体学习无法最大化长期未来奖励的策略）；利用机会必然会牺牲探索（反之亦然）；甚至还可能需要考虑不利后果（最小化遗憾）。[131]

[126]在非适当学习的情形下，与表示无关的困难性意味着，无论算法以何种形式表示其假设，学习仍然是困难的；其前提是该假设能够在多项式时间内求值。[127]这意味着，如果NP = co-NP，那么就可以推出PH = NP。人们普遍认为，PH发生坍缩的可能性微乎其微。[128]关于不可知学习与可实现学习之间更深层次关系的阐述，参见（Hopkins等人，2022）。[129]形式上，半空间是形如 $f(x) = \text{sign}(w_1 x_1 + ..., \ w_n x_n - \theta)$的布尔函数，其中 w_i是“权重”，θ是“阈值”，而 $w_1..., \ w_n, \ \theta \in \mathbb{R}$。对于自变量 ≥ 0，符号函数返回1；否则返回–1。[130]更多近期综述可参见 Angluin 和 Laird（1988年）；Amaldi 和 Kann（1998年）；H˚astad（2001年）；Ben-David等人（2003年）以及 Feldman等人（2012）。非适当学习情形另见 Daniely等人（2014）。[131]关于强化学习中样本复杂度的详细研究，参见 Kakade（2003）。

6.1.2. 机器学习理论与实践

需要强调的是，此处讨论的理论考量对于机器学习在各个领域中的实际可行性，或许意义有限。巧妙运用归纳偏置——例如任务表示和参数化工具——再加上海量训练数据与计算能力，仍在不断突破十年前看似不可能的界限。例如，近年来，自然语言处理领域的深度学习模型性能借助 Transformer 架构（Vaswani 等人，2017）得到提升；该架构利用注意力机制处理输入序列中任意位置的词元，通过高效并行化提升上下文敏感性。同样，深度强化学习的进展表明，只需少量示范样本就能显著加快学习过程（Hester 等人，2018）。事实上，令人惊讶的并非学习在总体上计算困难这一趋势，而是我们缺乏严谨的解释，说明为什么某些学习系统在实践中似乎能够很好地泛化。这就是所谓的“深度学习悖论”，其核心在于理解深度学习在缺乏理论解释的情况下为何仍能取得经验上的成功（Kawaguchi 等人，2017；Neyshabur 等人，2017；Arpit 等人，2017；Zhang 等人，2021）。由此又引出一系列错综复杂的问题，这些问题或多或少是机器学习所独有的。例如，在互联网上大量数据上训练的大语言模型，成为伦理的模型究竟意味着什么？[132] 我们应如何理解对足够先进的“黑箱”系统在透明性、可解释性、鲁棒性、安全性和公平性方面的要求？（Gunning 等人，2019；Amodei 等人，2016；Gabriel，2020；Berk 等人，2021）。[133]

虽然这些问题本身值得关注，但我们的复杂性分析带来了一个特别需要注意的问题：即归纳偏置及其道德论证的作用。从某种强意义上说，学习系统的成功——例如训练效率和预测准确率——似乎与其所利用的归纳假设以及由此引入的归纳问题成反比。也就是说，要使道德学习发挥作用，我们需要相对明确地了解道德问题的性能度量——例如，通过某种预先定义的评分、目标或目标函数来界定——即我们希望学习系统处理的道德问题，或希望其表现出的道德行为。因此，人们可能会质疑，现有的道德学习系统是否能够产生任何“新的”或“真正的”道德洞见，因为它们只是在经过特定归纳偏置筛选的既定数据上进行训练。与神命论和法律实证主义类似，这种观点预设了我们已经有了所寻求问题的答案。因此，道德机器学习的问题并不在于学习本身的计算复杂性，而在于如何为归纳能够发挥作用所需利用的道德假设提供辩护。正如卡尔·波普尔（1962）精辟地指出：

> 在构建归纳机器时，我们这些机器的设计者必须先验地 决定什么构成它的“世界”；哪些事物应被视为相似或相等；以及我们希望机器能够在其“世界”中“发现”哪类 “法律”。

[132]关于使用多种度量和测试用例对大型语言模型进行“整体”评估，请参见 Liang 等人（2022）。[133]算法公平性领域的近期研究表明，不同的公平性概念之间，以及公平性与准确性之间，存在不可避免的权衡（Berk 等人，2021年）。

> 换言之，我们必须在机器中构建一个框架，以确定其“世界”中哪些内容是相关或有趣的：机器将拥有其“先天的”选择原则。相似性问题已由机器的制造者替它解决，因此，他们也就为机器诠释了这个“世界”。（第48页）

7. 心智与机器的道德可处理性

首先，我们依据马尔的三个分析层次，提出了分析伦理学复杂度的三种可能解释。随后，我们着手分析了因果引擎、规则遵循者和学习者面临的一系列伦理问题。结果汇总于表3。根据这些调查结果，计算复杂性能够教给我们什么关于道德的知识？本节将讨论其对道德机器（7.1）和人类道德（7.2）的影响、道德可处理性论题的解释前景（7.3），以及局限性（7.4）和未来研究方向（7.5）。

7.1. 对道德机器发展前景的影响

不可处理性结果对道德机器的发展前景有何影响？首先，也是最重要的一点，由行动计划组合学（4.1）、概率因果推断（4.2）、动态与部分可观测环境（4.3）、一般规则（5.1）、策略动态（5.1.3）、逻辑（5.2）、语义学（5.2.3）和学习（6.1）所导致的不可处理性（和不可判定性），使我们可以明确得出结论：完美道德机器是不可能实现的。在许多情况下，解决方案的次优近似也难以处理。因此，道德机器的开发者应当在受限资源的基础上，力求做到“尽可能好”。Brundage（2014年）、Mabaso（2021年）、Hew（2014年）、Hagendorff和Danks（2022）也曾得出类似结论——尽管没有如此形式化；对于熟悉有限理性（Simon，1955，1990；Rubinstein，1998；Russell和Subramanian，1994）或有限伦理性（Bazerman和Tenbrunsel，2011年；Tenbrunsel和Messick，2004）的学者而言，这些结论也不足为奇。不过，本文提出的研究也有助于明确有限计算主体在伦理决策领域所面临的复杂度类型，以及由此产生的最优性与可行性之间的相关权衡。因此，本文应能为人工道德主体的相关讨论提供参考，因为它使“人工道德主体在实践上是否可行、在规范上是否值得追求”这一问题，更贴近人工智能方法的实际层面；而不是将讨论重心放在现有人工智能系统所缺乏、且为人类独有的能力上（Stenseke，2023）。

同样，这些结果也应有助于进一步发展在道德领域中实现的人工系统。具体而言，伦理问题的复杂度凸显了智能体的认知能力与时间、记忆、知识、交流、学习和启发式方法等道德资源之间的密切关系。然而，这也体现出道德资源方面显著的实现差异，可能会模糊实践道德能力的一般概念。也就是说，尽管我们或许能够确定特定情境下某个智能体可用的资源，以及如何有效利用并组合这些资源以产生胜任的伦理行为，但很难将这些见解泛化到任何智能体在任何情境中的表现。例如，对社交机器人而言，在教室中何种行为可被视为胜任的伦理决策，与高速交通情境中所要求的伦理能力大相径庭。

问题	结果
组合学（4.1）	
无序动作的最优计划	$\Theta(2^n)$
有序动作的最优计划	$\Theta(n!)$
STRIPS 与命题规划	PSPACE 完全（Bylander，1991，1994年）
贝叶斯推断（4.2）	
精确推理	#P 完全（Roth，1996年）
最大概率解释（MPE）	NP完全（Shimony，1994年）
最大后验假设（MAP）	NP^{PP}完全（Park and Darwiche，2004）
近似精确推理	NP难（Dagum and Luby，1993年）
近似最可能解释	NP难（Abdelbar and Hedetniemi，1998年）
部分MAP	NP难（Park and Darwiche，2004）
序贯决策（4.3）	
有限马尔可夫决策过程	从命题逻辑到 EXPSPACE 完全（Mundhenk 等人，2000年）
有限部分可观测马尔可夫决策过程	PSPACE 完全（Papadimitriou 和 Tsitsiklis，1987年）
无限部分可观测马尔可夫决策过程	不可判定（Madani 等人，2003年）
无休止多臂老虎机	PSPACE-困难（Papadimitriou 和 Tsitsiklis，1994年）
策略动态（5.1）	
有限交互式部分可观测马尔可夫决策过程	PSPACE 完全（Papadimitriou 和 Tsitsiklis，1987年）
去中心化马尔可夫决策过程	NEXP-困难（Bernstein 等，2002年）
双人纳什均衡（纳什均衡）	PPAD-完全（Chen等人，2009）
最大平等纳什均衡（最大纳什均衡）	NP完全（Gilboa和Zemel，1989年）
近似最大纳什均衡	NP完全（Conitzer和Sandholm，2008年）
纯策略贝叶斯纳什均衡	NP难（Conitzer和Sandholm，2008年）
无限马尔可夫博弈中的纯纳什均衡	PSPACE-困难（Conitzer和Sandholm，2008年）
有限马尔可夫博弈中的纯纳什均衡	NP难（Conitzer和Sandholm，2008年）
相关均衡（相关均衡）	P类（Gilboa和Zemel，1989年）
最大相关均衡	NP难（Papadimitriou and Roughgarden，2008年）
逻辑（5.2）	
一阶逻辑可满足性	不可判定（图灵，1936年；丘奇，1936年）
命题逻辑可满足性	NP完全（库克，1971年）
模态逻辑有效性	PSPACE 完全（K系统、T系统、S4系统），NP完全（S5系统）（Ladner，1977年）
多智能体模态逻辑（MAML）	PSPACE 完全（Halpern 和摩西，1992年）
MAML + 共同知识	EXPTIME 完全（Halpern 和摩西，1992年）
时序逻辑有效性	PSPACE 完全（Sistla 和 Clarke，1985年；Spaan，1993年）
SAT-命题动态逻辑	EXPTIME 完全（Fischer 和 Ladner，1979年；Pratt，1980）
SAT-道义 STIT 逻辑	不可判定（Schwarzentruber 和 Semmling，2014年）
道义输入/输出逻辑	NP/co-NP-困难（Sun 和 Robaldo，2017年）
描述复杂性（5.2.1）	
FOL	对数时间层级 & AC^0 (Immerman，1998年)
最小不动点一阶逻辑	P类（Immerman，1982年；Vardi，1982年）
$SO\exists$、$SO\forall$ 和二阶逻辑	分别为 NP、co-NP 和多项式层次（Fagin，1974年）
带传递闭包的二阶逻辑	PSPACE（Immerman，1989年）
最小不动点存在二阶逻辑	EXPTIME（Abiteboul 等，1997年）
学习（6）	无免费午餐定理（Wolpert，1992年、1996年；Schaffer，1994年）
PAC 可学习性的样本复杂度	$O\left(\frac{VCD(\mathcal{H}) + \ln\frac{1}{\delta}}{\epsilon}\right)$ (Hanneke, 2016年)
适当可实现PAC	从P类到NP难（Pitt和Valiant，1988年）
非适当PAC	NP难（密码学假设）（Kearns和Valiant，1994年）
适当不可知PAC	NP难（Feldman等人，2012）

表3：所调查的复杂性结果汇总。

实现差异可以作负面解读：这表明无法建立通用基准来评估计算系统的伦理表现。更乐观地看，它也揭示了可以根据资源依赖性（例如，在时间或信息有限的情况下）建立领域特定或问题特定道德基准的空间。更直接地说，它指出了相关基准已然存在的领域：例如，在复杂的协调博弈中寻找道德上理想的均衡、在多臂老虎机情境中尽量减少遗憾、构建易处理的道德推理逻辑，或设计高效的贝叶斯推断算法。

这也表明，机器伦理学在道德层面仍有大量工作有待完成。几乎在每种情形下，最优性与可行性、性能与效率之间都存在令人不安的权衡，而这些权衡本身或许也需要规范性辩护。然而，尽管随着机器学习兴起，计算机科学或许已转向成为一门经验科学，我们的道德理论仍深深植根于关于是非的理论理想，而这些理想可能预设了能够不切实际地获取理性预言机。这凸显出一种尚未解决的张力：规范理论究竟应作为“理想善”的标准，还是作为行动指导启发式，以取得次优但可行的结果。例如，如果我们的道德理论假定了不切实际的计算，我们该如何为其在实践中的辩护奠定坚实基础？同样，如果这种张力仍未解决，我们就无法明确判定，某个有害行动究竟源于能力不足，还是道德过错。这也给道德理论带来了挑战：例如，一个依据 NT_1 评估自身行动的智能体，按照 NT_1，若遵循替代理论 NT_2所提供的行动决策，或许会处境更好。[134] 因此，关键问题是：我们应依据什么标准，来修订道德理论，使其能够作为决策程序，适用于智能体有限且因实现而异的资源？在机器语境中，我们认为，可以从计算复杂性的视角富有成效地探究这些开放性问题，因为它提供了分析手段，用以衡量资源如何关联于性能的形式化概念。更实际地说，它为探讨机器应当做什么提供了途径，依据的是机器究竟能够做什么，以及它们能够有效地做什么；这反过来也能揭示计算系统可以成功应用于作出胜任的伦理决策的领域。

7.2. 对人类道德的影响

更广泛地说，这些结果对我们理解人类道德有何影响？这最终取决于人们如何看待人类心智，尤其取决于人们是否认为人类心智在本质上具有计算性，以及这种计算性的程度。那些认为人类主要是文化、社会和精神性存在的人，可能会发现，计算视角对于理解人类道德完全无关紧要。另一些试图从认知能力角度理解人类行为的人，则可能会认为，假定人类认知至少具有某些特定特征和约束是有益的，因为这些特征和约束可以用来限制假设空间。简而言之：关于人类心智能够做什么、不能做什么的假设，应当有助于我们理解人类

[134]此类“自我消解”理论的典型例子是行为功利主义（（Parfit，1984），第9节和第17节），因为智能体在实践中不遵循行为功利主义的规定，反而会带来更多总体善。

心智。从后一种观点来看，转而接受某种关于人类心智的计算主义，便显得颇具吸引力，因为它提供了大量额外的科学工具；例如，运用计算架构来建模、检验和修正关于认知与行为的理论。如果采取这一步，计算复杂性就会成为不可或缺的工具，因为它有助于我们约束人类认知的可能计算理论空间。这一观点可概括为可处理认知论题，该论题指出，在某种合理的可处理性概念下，认知能力的计算模型必须是计算上易处理的（Van Rooij，2008年；Van Rooij等，2019年）。反之，如果某个计算理论蕴含难处理的计算，就表明该理论并不充分。该论题的一个形式化变体是*P*-认知论题，它断言认知功能受到多项式时间的约束。在认知科学中，Cherniak（1986年）、Tsotsos（1990年）、Levesque（1989年）和Frixione（2001年）明确提出将P-认知论题作为人类认知计算层面理论的指导原则，而大批认知心理学家也在隐含地将其作为约束因素。[135]此外，观察到许多NP-hard问题实际上可以针对输入参数的某一部分得到高效求解，这促成了这样一种论证：应当以固定参数可处理性论题取代P-认知论题（Van Rooij，2008年）。该论题指出，认知功能在总体输入规模上受到多项式时间的限制 n，同时允许其输入参数的某一部分使用超多项式时间（Downey和Fellows，2012）。一种相关的建模范式是资源理性认知的概念（Lieder和Griffiths，2020年），它从最佳利用有限计算资源的角度处理认知建模。

因此，即使人类心智究竟是一种什么样的计算机（或者说，能否用任何计算模型有意义地刻画人类心智）仍不清楚，可处理性考量——结合认知建模和实验数据——也可以作为一种假设，帮助我们划定认知理论的可行空间。自然，这也包括构成人类道德认知的各种功能。例如，如果我们认为人类会进行因果推理、策略推理或逻辑推理，从而表现出胜任的道德行为，那么可处理性考量就会直接约束构成道德行为基础的计算层问题空间。此外，如果我们有理由相信人类以特定方式进行这些道德推断——例如，运用某些贝叶斯推断、决策理论或逻辑推断技术——那么复杂性分析将有助于确定性能与可行性之间相关的权衡；这些权衡可能与本文综述的复杂性结果直接相关。

7.3. 道德可处理性论题

可处理性在道德行为与认知理论中的作用，可以表述为道德可处理性论题(MTT)。这是可处理认知论题的自然延伸，并指出：在某种合理的人类道德认知模型下，道德——包括道德行为、道德问题解决和道德认知——受到计算可处理性的约束（见图4）。[136] MTT是一种同时指向规范层面和

[135]关于可处理认知论题及其形式变体的详细论述，请参阅 Van Rooij（2008年）。[136]此处“合理”一词有意保持模糊，因为目前尚未就能反映人脑的具体计算模型达成共识。有关近期的一项提议，请参阅 Blum 和 Blum（2022）。

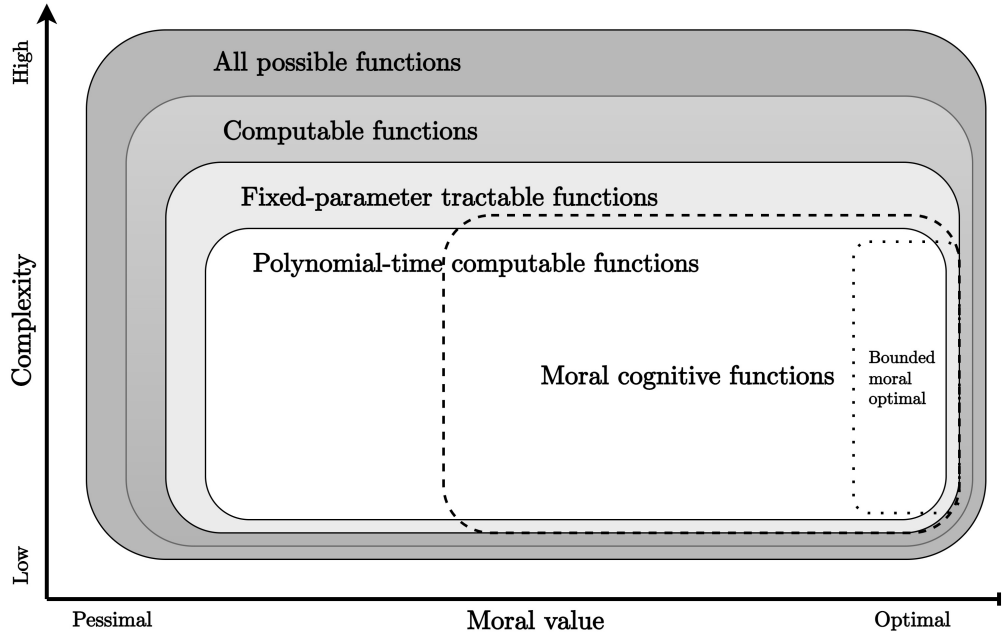

图4：道德可处理性论题（MTT）指出，可能的道德认知功能集合受到可处理性约束。更形式地说，这可以表述为一个函数子集：当输入参数足够小时，这些函数是固定参数可处理的；该子集还包含可在多项式时间内计算的函数集合。我们认为，MTT 可以发挥以下作用：(1) 作为规范理论行动指导的元伦理标准，即行动指导应当考虑智能体的资源而具备可行性；(2) 在智能体是否因认知约束失效而有不当行为尚不明确时，为规范判断和责任认定提供指引；(3) 作为研究人类道德认知与心理学的实验范式；(4) 缓解道德语境中可行性与性能之间的张力。

描述性方向：

(1) *MTT*作为规范理论的元伦理标准 — MTT提出了一项元伦理观点，强调道德理论所提出的计算问题，也应当相对于遵循该理论的智能体所拥有的资源而言是易处理的；也就是说，该理论在多大程度上旨在为智能体提供有意义的行动指导。如果做不到这一点，就可能表明该道德理论对智能体资源提出了不切实际的要求，因此应当修订该理论，以限制智能体可被要求解决的问题空间。因此，它有助于将那些不尊重智能体能力有限这一事实的规范性行动指导"自我挫败"理论排除在外。从实践角度看，它还能帮助我们识别这样的情境和语境：在其中，某种形式的行动指导有望比竞争性替代方案产生更好的结果。

(2) *MTT*作为规范判断与责任的指导原则 — 在智能体行为不道德的情境中，MTT可以帮助我们回答这样一个问题：该智能体的不道德行为是否源于道德能力和理性的缺失。例如，如果对于人类道德能力的某种合理理解而言，计算要求是公正的，那么这可能指向当下情境中缺少某种道德资源（例如，时间、记忆、知识、学习，或理性方面的要求）。如果无法确认存在此类缺失，也可以据此追究该智能体对其不道德行为的责任。

（3）*MTT*作为一种实验范式 ——在描述性方向上，MTT可以为研究人类道德的现有范式提供限定因素；从人类道德认知的计算建模，到道德心理学中的实验研究。更确切地说，MTT可用于识别支撑道德决策中认知过程的原则和算法，揭示可行性与性能之间相关的权衡，并进一步考察资源在特定道德语境中的作用。

（4）*MTT*作为解决可行性与最优性问题的方案 ——MTT的另一项特征是，它能够帮助解决道德语境中可行性与性能之间悬而未决的张力，即后者会受到前者的直接约束。例如，假设存在一个能够产生最优道德价值的不动点，那么MTT意味着存在一种行动指导理论，能够根据行动主体所拥有的资源，产生尽可能高的道德价值。[137] 同样的思想也可以用于道德共同体：根据道德共同体中智能体共同共享 的资源，制定能够产生最高道德繁荣的行动指导方案——例如共同福祉，或相互同意的道德价值观。换言之，可能存在一种规范性行动指导，能够实现对有限认知资源的道德最优利用；这就是资源理性道德理论（Lieder和Griffiths，2020年）。从博弈论视角来看，这可以被解释为智能体有限资源空间中的最优道德均衡点（例如最大共同收益）（图4）。

7.4. 局限性

总体而言，道德复杂性分析，尤其是 MTT，在解释可行性方面存在若干不足与局限。我们将围绕三个宽泛的问题，简要探讨其中一些问题：(i) 计算复杂性与

[137]更为温和地说，我们可以假设，道德可允许性存在一个合理的阈值，而非追求最优值；这一阈值的依据是某种道德价值概念。

在现实世界领域中开发和部署“具备道德能力”的人工智能系统有何关联，在多大程度上相关？(ii) 人类道德的哪些方面是复杂性分析未能涵盖的？(iii) 关于人类如何使用具有道德导向的人工智能系统（例如，旨在用于扩展或增强人类道德能力的人工智能系统），复杂性能够告诉我们什么？

(i) 需要强调的是，本文分析的计算问题对于人工智能系统在道德领域中的许多现实实现可能意义有限。例如，计算复杂性究竟能告诉我们什么——如果能告诉我们什么的话——关于自动驾驶、社会辅助机器人学和自然语言处理等领域中的道德行为？尽管自动驾驶车辆的总体目标可以概括为“从 A 驾驶到 B，确保安全”，但它包含大量更小的子任务和目标——例如路径规划、遵守交通规则、通过一个又一个路口——为此，它会运用一系列分布式能力——例如传感器、电机、地图和仿真——这些共同构成了具备能力的自动驾驶（Badue 等人，2021）。因此，除非基于特定汽车的能力，否则对自动驾驶汽车道德能力进行复杂性分析难以提供多少见解。正如第 6 节所述，大语言模型面临的一些技术伦理挑战，可能呈现出相反的问题：约束 生成文本这一普遍能力（例如，使其遵循人类原则与价值观）。也就是说，任务可能不是从尽可能大的搜索空间中学习最优解——例如，通过使用海量人类生成的文本进行训练——而是通过某种合理的归纳偏置来约束搜索，使之与人类价值观保持一致。同样，社会辅助机器人学中最受关注的普遍伦理问题——例如人机交互（HRI）中的欺骗、尊严、信任和认可——与本文讨论的因果、逻辑和策略推理的形式化概念相去甚远（Boada 等人，2021）。

核心在于，每个论域都有其独特的一系列技术与伦理挑战，需要结合该论域的特殊条件加以应对。反过来说，这些挑战可能与应用规范伦理学的计算关系不大，反而取决于某些人类实践所预设的特定规范性要求（Behdadi 和 Munthe，2020年）。然而，如果 人工智能系统的开发旨在使其行为符合规范伦理学，那么所综述的复杂性结果对于此类努力具有普遍意义，正如它们对于计算技术进步的若干重要范式仍然至关重要，例如概率推理、知识系统和学习。更浅显地说，即使目标只是为人工智能系统补充某些 进行某种 形式伦理决策所需的能力，这类系统的行为也会受到有限计算资源的制约。因此，资源理性复杂性分析有助于辨别哪些伦理问题能够由机器解决，哪些能够高效解决、哪些仍有改进空间，并指出相关资源与权衡。

（ii）另一个亟待解决的重要缺口，是人类道德与本文探讨的应用规范伦理学的计算形式之间的差异。必须着重强调，尽管本文综述的研究结果对计算系统有直接影响，但任何关于这些结果可能如何影响我们对人类道德的理解的推论，都建立在推测性假设之上。正如第 3 节所讨论的，伦理学是一个含义模糊且多面向的概念；人类实际上“解决”什么样的伦理问题、如何解决，以及他们会用什么来解决（例如情绪、推理），都是尚无定论的问题，可能的答案既多样又复杂。同样，尽管规范伦理学旨在

找出普遍适用的“善”与“恶”标准，但这不过是构成伦理生活的广阔图景中的一小部分。采取更为怀疑的观点，甚至可以主张，规范伦理学——按照当代英语分析哲学中的理论构想——与伦理生活毫无关联（Stocker，1977年）。因此，设想人类会用穷举搜索方法制定伦理行动计划（或制作最优沙拉）（4.1）、用（任意）大型贝叶斯网络进行因果推断（4.2）、在策略互动中计算平等主义纳什均衡（5.1.5），或针对可能世界检验行为规范遵循情况（5.2.2），或许显得有些奇怪。在日常伦理生活中，人类可能会运用广泛的资源——例如情绪与动机、内疚与羞耻、灵性与意识形态、批判性反思与心智理论——而这些都未纳入本分析。因此，从计算到算法再到实现，本文所呈现的分析不应阻碍研究者探索其他有望帮助理解人类道德复杂且多样形态的研究方向；例如，具身体验与切身感受、人际关系中彼此依存与脆弱性的共享，以及社会与文化中的规范和制度。

尽管如此，如果有人认为认知可处理性和 MTT 有任何价值，那么复杂度就能为我们探索规范伦理学、描述伦理学与应用伦理学之间广阔的交汇领域提供指引。即使仅仅承认计算限制对某些 形式的人类道德行为具有重要意义，也足以成为进一步研究这些限制的理由。此外，认为许多人类能够并且确实会——至少偶尔会——遵循义务论和后果主义决策程序，并按照美德伦理学的要求，培养使自己成为更好的道德主体的心理倾向，这并非不合理；尽管支撑这些过程的实际认知程序可能与本文所考察的有所不同。为此，若将这些规范理论视为与人类的认知能力完全无关，未免相当奇怪；例如，除非某些道德启发式方法（如原则或理论）适用于人类道德共同体——对人类而言可判定且易处理——否则很难解释它们在实践中的成功或广受欢迎（Alexander，2007年）。有时，甚至可以合理地认为，其中一些启发式方法的动机源于其计算效率；例如，遵循道德规则的计算效率（5.3）。因此，尽管我们应正视人类道德与计算道德之间的巨大鸿沟，但若排除这样一种可能性，那就不明智了：支撑前者的认知过程中存在或多或少严谨的模式，而这些模式反过来又能够借助后者接受形式化研究。

（iii）本文略去的另一个重要领域，是在人的道德行为中整合使用人工智能系统。也就是说，我们主要考察了计算系统孤立状态下的伦理行为，而没有纳入任何“人在回路”。因此，我们忽略了人工智能系统如何能够用于支持或增强人类伦理决策这一整合性前景。毕竟，机器是一大批计算方法的集合，旨在实现其人类用户的目标。由此，这或许指向一个同样可以通过计算复杂性加以阐明的研究领域：在资源受限的情况下，人工智能系统如何促进并支持人类实践中的道德繁荣（Vallor，2015年；Giubilini 和 Savulescu，2018年）。例如，许多对人类而言难处理的问题，对机器而言可能易处理；反过来，许多对机器而言难处理的问题，对机器而言也可能易处理。因此，核心问题是：基于资源约束，

如何开发和使用人工智能，才能拓展而非限制道德推理的空间（Vallor，2016年）？

7.5. 未来工作

除了已经描述的方向之外，还有许多值得进一步探究的领域，可以用来研究心智与机器的道德可处理性。首先应当强调，尽管本文讨论了许多与道德行为相关的复杂性结果，但也避开了数量更多的相关结果。对于所讨论的大多数计算问题，在不同条件和假设下，还有数百项相关结果值得考量。事实上，我们省略了复杂性理论中一些完整领域的结果——例如参数化复杂度、交流复杂度、证明复杂度和电路复杂度——这些结果可能为道德计算的局限性带来更多见解。在讨论过的问题中，我们尤其认为有两个领域值得进行更深入的研究：(i)（道德）机器学习的样本复杂度与遗憾最小化，以及 (ii) 算法博弈论（结合算法设计理论）。研究第一个领域的原因是，我们对道德语境下机器学习的复杂度仍然了解甚少，尤其是考虑到可解释性（Gunning 等人，2019年）、归纳（Sterkenburg 和 Gr¨unwald，2021年）以及“深度学习悖论”（Zhang 等人，2021年）等问题。另一个原因是，机器学习是推动现代人工智能发展的主要手段。更迫切的是，学习系统已经应用于广泛的人类实践之中，包括可能涉及显著道德决策的领域，如医疗保健、自动驾驶、法律、警务和教育。

之所以追求（ii），是因为博弈论为形式化研究互动提供了统一框架，进而使互动能够通过算法建模和分析加以处理。因此，我们认为，它有望在计算机科学、道德理论、进化生物学、行为经济学和社会科学等历来相互独立的领域之间催生富有成果的协同效应。例如，如果我们采纳这样一种观点：规范理论之间的共识多于分歧，[138] 那么，资源理性算法博弈论分析或许有助于确定，在什么条件下，某些理论比其他理论更具实践可行性。例如，要让某种道德启发式方法——如一种行动指导型规范理论——为主体共同体带来尽可能高的道德繁荣，需要具备怎样的认知能力和计算资源？哪些道德行为能够有效地计算或论证，哪些又只能通过学习获得？理想情况下，这类研究不仅能阐明计算主体要具备道德胜任力所需的具体资源，还能解释支撑最重要的伦理学理论的资源理性依据。

8. 结论

我们考察了一组数量庞大但远非详尽的复杂性结果，并讨论了它们对道德领域中的心智与机器的意义。归根结底，我们相信，这为

[138]例如，Parfit（2011年）认为，若相信结果主义者、契约主义者和康德主义者之间存在根本分歧，那就是错误的。他写道：“这些人只是在从不同方向攀登同一座山”（第385页）。

机器伦理学、道德哲学与道德认知心理学之间有趣的跨学科空间打开了大门；希望这不仅能为道德机器工程带来新的方向，也能推动我们理解复杂的道德科学。

附录

复杂性类

AC^0 — 一类判定问题：对于每种可能的输入规模，都存在一个对应的常深度无限扇入电路族可以求解这些问题，且电路门的数量受输入规模的某个多项式所界定。

LH — 一类判定问题，可由交替图灵机在时间 $O(\log n)$ 内通过有限次交替求解。详细介绍见 Immerman（1998）。

PL — 一类判定问题，可由受空间 $O(\log n)$ 限制的概率图灵机求解，错误概率为 $\epsilon < 1/2$（“概率对数空间”）。

P — 一类判定问题，可由受时间 $O(poly(n))$ 限制的确定性图灵机求解。

PP — 在错误概率为 $\epsilon < 1/2$ 的条件下，由受时间$O(poly(n))$限制的概率图灵机可解的判定问题类。

FTP — 由受时间$O(f(k)n^c)$限制的确定性图灵机可解的判定问题类，其中 f 是仅取决于参数 k 的函数，而 c 是常数。关于参数化复杂性的正式介绍，参见 Downey和Fellows（2012）。

NP — 由受时间 $O(poly(n))$ 限制的非确定性图灵机可解的判定问题类。或者说，这是这样一类判定问题：其中“是”实例可由确定性图灵机在多项式时间内验证。

co-NP— NP 的补集。这样一类判定问题：其中“否”实例可由确定性图灵机在多项式时间内验证。

#P — 函数问题 $f(x)$ 类，其中 f 是受时间 $O(poly(n))$ 限制的非确定性图灵机的接受路径数。通俗地说，它是与 NP 相关的计数问题集合；也就是说，NP 判定问题问“是否存在”，而 #P 函数问题问“有多少”。

NP^{PP} ——可由非确定性图灵机在时间 $O(poly(n))$ 限制下求解，并可访问针对 PP 类问题的预言机的判定问题类别。

TFNP ——由受限于时间 $O(poly(n))$ 的非确定性图灵机可求解的函数问题类别，其中保证存在解决方案（“全函数非确定性多项式”）。

PPAD ——TFNP 的子类，其中函数凭借有向图上的奇偶性论证而保证为全函数——也就是说，保证存在解决方案（“有向图上的多项式奇偶性论证”）。详细论述请参阅 Papadimitriou （ 1994 年）。

PH — 多项式层次中各类的并集。可以使用预言机机器递归地定义：给定 $\mathrm{P} = \Delta_0^{\mathrm{P}} = \Sigma_0^{\mathrm{P}} = \Pi_0^{\mathrm{P}}$，我们定义 $\mathrm{P}^{\Sigma_i^{\mathrm{P}}} = \Delta_{i+1}^{\mathrm{P}}$、$\mathrm{NP}^{\Sigma_i^{\mathrm{P}}} = \Sigma_{i+1}^{\mathrm{P}}$ 和 $\text{co-NP}^{\Sigma_i^{\mathrm{P}}} = \Pi_{i+1}^{\mathrm{P}}$，以表示该并集：

$$\mathrm{PH} = \bigcup_{i=0}^{\infty} \Delta_i^{\mathrm{P}} \cup \Sigma_i^{\mathrm{P}} \cup \Pi_i^{\mathrm{P}} \tag{8}$$

PSPACE — 由空间$O(poly(n))$受限的确定性图灵机可解的判定问题。

EXPTIME — 由时间$O(2^{poly(n)})$受限的确定性图灵机可解的判定问题。

NEXPTIME — 由时间 $O(2^{poly(n)})$受限的非确定性图灵机可解的判定问题。通常记作 NEXP。

EXPSPACE — 由空间 $O(2^{poly(n)})$受限的确定性图灵机可解的判定问题。

利益冲突声明

作者声明，本研究是在不存在任何可被视为潜在利益冲突的商业或财务关系的情况下开展的。

作者贡献

本文所述工作由 JS 独立完成。

资助

本研究部分获得瓦伦堡人工智能、自主系统与软件计划——人文与社会（WASP-HS）的资助，该计划由玛丽安娜和马库斯·瓦伦堡基金会以及马库斯和阿玛莉亚·瓦伦堡基金会资助。

致谢

作者特别感谢 Per Austrin（瑞典皇家理工学院）和 Felix Lindner（Universit¨at Ulm）就计算复杂性提出的宝贵意见。作者也感谢隆德大学哲学与认知科学系的同事们，他们对论文的早期版本提出了有益的评论。作者尤其要感谢 Niklas Dahl 解答了有关逻辑和语义学的无数查询，感谢 Karl Samson 和 Gustav Stenseke Arup 对……提供的意见

合法性与道德之间的关系，也感谢 Christian Balkenius、Bj¨orn Petersson、Ylva von Gerber、Sandra Lofs Midelf、Jiwon Kim、Alexander Velichkov、Trond Arild Tjøstheim 和 Alfred Stenseke 提供的反馈。最后，作者感谢实践哲学高级研讨会（隆德大学）、哲学博士研讨会（隆德大学）、WASP-HS 2022年冬季会议、Georg-August-Universit¨at 的哲学研讨会，以及 Institut f¨ur K¨unstliche Intelligenz（Universit¨at Ulm）研讨会的与会者；论文的早期部分曾在这些场合宣读。

参考文献

Aaronson, S., 2013. 为什么哲学家应该关注计算复杂性。《可计算性：图灵、哥德尔、丘奇及其他》261, 327. Abdelbar, A.M., Hedetniemi, S.M., 1998. 信念网络地图的近似是NP难问题及其他定理。《人工智能》102, 21–38. Abel, D., MacGlashan, J., Littman, M.L., 2016. 将强化学习作为伦理决策框架，载于：AAAI人工智能、伦理与社会研讨会，亚利桑那州凤凰城。第02页。 Abiteboul, S., Vardi, M.Y., Vianu, V., 1997. 不动点逻辑、关系机与计算复杂性。《美国计算机协会会刊》（JACM）44, 30–56. Adam, S.P., Alexandropoulos, S.A.N., Pardalos, P.M., Vrahatis, M.N., 2019. 无免费午餐定理：综述。《近似与优化》，57–82. Albrecht, S.V., Stone, P., 2018. 对其他智能体建模的自主智能体：全面综述与开放问题。《人工智能》258, 66–95. Alexander, J.M., 2007. 道德的结构演化。Allen, C., Smit, I., Wallach, W., 2005. 人工道德：自上而下、自下而上及混合方法。《伦理学与信息技术》7, 149–155. Amaldi, E., Kann, V., 1998. 关于线性系统中非零变量或未满足关系最小化问题的可近似性。《理论计算机科学》209, 237–260. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Man´e, D., 2016. 人工智能安全中的具体问题。arXiv预印本 arXiv:1606.06565。 Anderson, M., Anderson, S.L., 2008. Ethel：迈向有原则的伦理型老年照护系统，载于：AAAI秋季研讨会：老年照护中的人工智能——解决老问题的新方案，AAAI出版社，弗吉尼亚州阿灵顿。第4–11页。 Anderson, M., Anderson, S.L., 2011. 机器伦理学。剑桥大学出版社。 Angluin, D., Laird, P., 1988. 从带噪样例中学习。《机器学习》2, 343–370. Annas, J., 2011. 智慧美德。牛津大学出版社。 安斯康姆, G.E.M., 1958. 现代道德哲学。《哲学》33, 1–19.

Applebaum, B., Barak, B., Xiao, D., 2008. 基于最坏情况假设的学习下界，载于：2008年第49届IEEE计算机科学基础年会，IEEE。第211–220页。 Arkin, R.C., 2007. 致命行为治理：将伦理嵌入混合审议式／混合机器人架构。报告。报告编号GIT-GVU-07-11，美国佐治亚州亚特兰大：佐治亚理工学院GVU……。 Arkoudas, K., Bringsjord, S., Bello, P., 2005. 通过机械化道义逻辑实现伦理机器人，载于：AAAI机器伦理秋季研讨会，AAAI出版社，美国加利福尼亚州门洛帕克。第17–23页。 Armstrong, S., 2015. 面向人工智能体的动机化价值选择，载于：AAAI人工智能与伦理研讨会，AAAI出版社，美国加利福尼亚州帕洛阿尔托。 Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., 等，2017. 深度网络记忆现象的深入考察，载于：国际机器学习会议，PMLR。第233–242页。 Asimov, I., 1942年. Runaround。《惊奇科幻》29, 94–103.̊Astr¨om, K.J., 1965年. 不完全状态信息下马尔可夫过程的最优控制。《数学分析与应用杂志》10, 174–205. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E., 1995. 赌局被操纵的赌场：对抗性多臂老虎机问题，载于：IEEE第36届计算机科学基础年会论文集，IEEE。第322–331页。 Aumann, R.J., 1974年. 随机化策略中的主观性与相关性。《数学经济学杂志》1, 67–96. Aumann, R.J., 1987年. 相关均衡作为贝叶斯理性的体现。《计量经济学会会刊》，1–18. Aumann, R.J., 2016年. 16. 一般合作型n人博弈中的可接受点，载于：《博弈论贡献》（AM-40），第四卷。普林斯顿大学出版社，第287–324页。 Austrin, P., Braverman, M., Chlamt´aˇc, E., 2013年. 纳什均衡NP完全变体的不可近似性。《计算理论》9, 117–142. Axelrod, R., Hamilton, W.D., 1981年. 合作的演化。《科学》211, 1390–1396. Azad-Manjiri, M., 2014年. 一种基于C4.5决策树算法构建道德智能体的新架构。《国际信息技术与计算机科学杂志》（IJITCS）6, 50–57. B¨ackstr¨om, C., Nebel, B., 1995. SAS规划的复杂性结果。《计算智能》11, 625–655.

Badue, C., Guidolini, R., Carneiro, R.V., Azevedo, P., Cardoso, V.B., Forechi, A., Jesus, L., Berriel, R., Paixao, T.M., Mutz, F., 等，2021年。《自动驾驶汽车：综述》。《Expert Systems with Applications》165，113816。Baker, C.L., 特南鲍姆，J.B., Saxe, R.R., 2007年。目标推断即逆向规划，载于：《认知科学学会年会论文集》。Balbiani, P., Herzig, A., Troquard, N., 2008年。审议型STIT理论的替代公理化与复杂度。《哲学逻辑杂志》37，387–406。Bales, R.E., 1971年。行为功利主义：对使行为正确的特征的说明，还是一种决策程序？American Philosophical Quarterly 8，257–265。Barsalou, L.W., 1999年。知觉符号系统。《Behavioral and Brain Sciences》22，577–660。Bauer, W.A., 2020年。有德性的人工道德主体与功利主义人工道德主体。《AI & SOCIETY》35，263–271。Bazerman, M.H., Tenbrunsel, A.E., 2011年。《盲点》，载于：《盲点》。普林斯顿大学出版社。Beall, J.C., 2007年。《说谎者的复仇：关于悖论的新论文》。牛津大学出版社。Behdadi, D., Munthe, C., 2020年。人工道德能动性的规范性方法。《Minds and Machines》30，195–218。Ben-David, S., Eiron, N., Long, P.M., 2003年。论近似最大化一致性的难度。《计算机与系统科学杂志》66，496–514。Bennett, J.F., 1976年。语言行为。边沁，J., 1789年。《道德与立法原理导论》。Garden City：Doubleday（1961年）。Berberich, N., Diepold, K., 2018年。有德性的机器——新技术的旧伦理？arXiv预印本 arXiv:1806.10322。Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A., 2021年。刑事司法风险评估中的公平性：现状。《Sociological Methods & Research》50，3–44。Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., 2019年。大规模深度强化学习下的Dota 2。arXiv预印本 arXiv:1912.06680。Bernstein, D.S., Givan, R., Immerman, N., Zilberstein, S., 2002。《马尔可夫决策过程去中心化控制的复杂度》。《运筹学数学》27，819–840。Bernstein, E., Vazirani, U., 1997年。量子复杂性理论。《SIAM Journal on Computing》26，1411–1473。

Bicchieri, C., 2005年。《社会语法：社会规范的本质与动态》。剑桥大学出版社。Binder, K., Heermann, D., Roelofs, L., Mallinckrodt, A.J., McKay, S., 1993年。统计物理学中的蒙特卡罗模拟。《Computers in Physics》7，156–157。Blackburn, S., 1992年。无论顺境还是逆境，载于：《亚里士多德学会论文集》，第284–299页。Blackburn, S., 1998年。《支配性激情》。牛津大学出版社，牛津。Block, N., 2019年。无报告范式有何问题，以及如何加以改进。《认知科学趋势》23，1003–1013。Blum, L., Blum, M., 2022年。从理论计算机科学视角提出的意识理论：来自意识图灵机的启示。《美国国家科学院院刊》119，e2115934119。Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K., 1989年。可学习性与Vapnik–Chervonenkis维数。《美国计算机协会会刊》36，929–965。Boada, J.P., Maestre, B.R., Gen´ıs, C.T., 2021年。社会辅助机器人学的伦理问题：批判性文献综述。《社会中的技术》67，101726。Bodlaender, H.L., 1994年。树宽导览。《控制论学报》11，1。Bouneffouf, D., Rish, I., 2019年。多臂老虎机与上下文老虎机实际应用综述。arXiv预印本 arXiv:1904.10040。Brandenburger, A., Dekel, E., 1993年。信念层级与共同知识。《经济理论杂志》59，189–198。Brandom, R., 1994年。《明示其义：推理、表征与话语承诺》。哈佛大学出版社。Brandom, R., 2006。关于心智、意义与理性的康德式教诲。《南方哲学杂志》44，49–71。Brandt, R.B., 1979年。《善与正当理论》。Brentano, F., 1874年。《经验立场下的心理学》。Bringsjord, S., Taylor, J., 2012。机器人伦理学的神命令进路。《机器人伦理学：机器人技术的伦理与社会影响》，85–108。Broome, J., 1987年。功利主义与期望效用。《哲学杂志》84，405–422。Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 等，2020年。语言模型是小样本学习者。《神经信息处理系统进展》33，1877–1901。

Bro˙zek, B., Janik, B., 2019年。人工智能能成为道德主体吗？《心理学新思想》54，101–106。Brundage, M., 2014年。机器伦理学的局限性与风险。《实验与理论人工智能杂志》26，355–372。Bubeck, S., Wang, T., Viswanathan, N., 2013年。多臂老虎机中的多重识别，载于：国际机器学习会议，PMLR，第258–265页。Bylander, T., 1991年。规划的复杂性结果，载于：IJCAI，第274–279页。Bylander, T., 1994年。命题STRIPS规划的计算复杂性。《人工智能》69，165–204。Cai, Y., Papadimitriou, C., 2014年。同时贝叶斯拍卖与计算复杂性，载于：第十五届ACM经济学与计算会议论文集，第895–910页。Camerer, C.F., Ho, T.H., Chong, J.K., 2004。博弈的认知层级模型。《经济学季刊》119，861–898。Campbell, M.S., Marsland, T.A., 1983年。极小极大树搜索算法的比较。《人工智能》20，347–367。de Campos, C.P., 2020年。贝叶斯网络中MAP复杂性几乎没有新进展，载于：国际概率图模型会议，PMLR，第149–160页。Capraro, V., Rand, D.G., 2018年。做正确的事：实验证据表明，驱动人类亲社会性的，是人们对道德行为的偏好，而非公平或效率本身。即将发表于《判断与决策》。Carnap, R., 1947年。意义与必然性：语义学与模态逻辑研究。Casebeer, W.D., 2003年。道德认知及其神经组成要素。《自然·综述：神经科学》4，840–846。Casella, G., Berger, R.L., 2021年。《统计推断》。圣智学习。Cassandra, A.R., Kaelbling, L.P., Littman, M.L., 1994年。在部分可观测随机领域中采取最优行动，载于：AAAI，第1023–1028页。Cervantes, J.A., L´opez, S., Rodr´ıguez, L.F., Cervantes, S., Cervantes, F., Ramos, F., 2020年。人工道德主体：现状综述。《科学与工程伦理》26，501–532。Cesa-Bianchi, N., Lugosi, G., 2006年。《预测、学习与博弈》。剑桥大学出版社。Chalmers, D.J., 1997年。《意识心灵：探寻基础理论》。牛津平装书。

Chatterjee, K., Chmelik, M., Tracol, M., 2016年。对于具有 ω-正则目标的部分可观测马尔可夫决策过程，哪些问题是可判定的？《计算机与系统科学杂志》82，878–911。Chen, X., Deng, X., Teng, S.H., 2009。确定计算双人纳什均衡的复杂度。《美国计算机协会会刊》56，1–57。Cherniak, C., 1986年。《最低限度理性》。麻省理工学院出版社。Church, A., 1936年。关于判定问题的一则注记。《符号逻辑杂志》1，40–41。Cloos, C., 2005年。Utilibot项目：一种基于功利主义的自主移动机器人，载于：《机器伦理学：2005年AAAI秋季研讨会论文集》，AAA出版社，门洛帕克，加利福尼亚州，第38–45页。Cobham, A., 1965年。函数的内在计算难度。Cofino, A.S., Cano, R., Sordo, C., Gutierrez, J.M., 2002年。用于概率天气预报的贝叶斯网络，载于：第15届欧洲人工智能会议（ECAI），Citeseer。Coleman, K.G., 2001年。Android arete：迈向计算代理的美德伦理学。《伦理学与信息技术》3，247–265。Conitzer, V., Sandholm, T., 2008年。关于纳什均衡的新复杂性结果。《博弈与经济行为》63，621–641。Conway, P., Gawronski, B., 2013年。道义论与功利主义倾向对道德决策的影响：一种过程分离方法。《人格与社会心理学杂志》104，216。Cook, S.A., 1971年。定理证明过程的复杂度，载于：第三届ACM计算理论年会论文集，第151–158页。Cooper, G.F., 1990年。利用贝叶斯信念网络进行概率推理的计算复杂性。《人工智能》42，393–405。Courville, A.C., Daw, N.D., Touretzky, D.S., 2006年。变化世界中的条件作用贝叶斯理论。《认知科学趋势》10，294–300。Crawford, V.P., Sobel, J., 1982年。策略性信息传递。《计量经济学：计量经济学会杂志》，1431–1451。Crisp, R., Slote, M.A., 1997年。《美德伦理学》。牛津大学出版社，牛津。Cummings, R., Ligett, K., Radhakrishnan, J., Roth, A., Wu, Z.S., 2016年。协调复杂度：以少量信息协调庞大群体，载于：2016年ACM理论计算机科学创新会议论文集，第281–290页。

Dagum, P., Luby, M., 1993年。贝叶斯信念网络中的概率推理近似是NP难问题。《人工智能》60，141–153。Dagum, P., Luby, M., 1997年。贝叶斯推断的最优近似算法。《人工智能》93，1–27。Dahl, N., 2022。一项关于意义理论的不动点问题。Synthese 200，1–15。Daniely, A., Linial, N., Shalev-Shwartz, S., 2014年。从平均情形复杂度到非适当学习复杂度，载于：第四十六届ACM计算理论研讨会论文集，第441–448页。Daskalakis, C., Mehta, A., Papadimitriou, C., 2006。一则关于近似纳什均衡的短文，载于：互联网与网络经济国际研讨会，施普林格，第297–306页。De Giacomo, G., Massacci, F., 2000年。将演绎与模型检测结合于逆向命题动态逻辑的语义树和算法中。《信息与计算》162，117–137。Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., 等，2022。通过深度强化学习控制托卡马克等离子体。《自然》602，414–419。Dehghani, M., Tomai, E., Forbus, K.D., Klenk, M., 2008a。一种道德决策的综合推理方法，载于：AAAI，第1280–1286页。Dehghani, M., Tomai, E., Klenk, M., 2008b。一种道德决策的综合推理方法，载于：第二十三届人工智能促进协会人工智能会议论文集，AAAI出版社，芝加哥，伊利诺伊州，第1280–1286页。Doshi, P., Gmytrasiewicz, P.J., 2009。用于近似交互式部分可观测马尔可夫决策过程的蒙特卡洛采样方法。《人工智能研究杂志》34，297–337。Downey, R.G., Fellows, M.R., 2012。参数化复杂性。施普林格科学与商业媒体。Ehrenfeucht, A., Haussler, D., Kearns, M., Valiant, L., 1989年。学习所需样例数量的一般下界。《信息与计算》82，247–261。Fagin, R., 1974年。广义一阶谱与多项式时间可识别集合。《计算复杂性》7，43–73。Fehr, E., Fischbacher, U., 2004。社会规范与人类合作。《认知科学趋势》8，185–190。URL: https://www.sciencedirect.com/science/article/pii/S1364661304000506，doi: https://doi.org/10.1016/j.tics.2004.02.007。Fehr, E., G¨achter, S., 2000年。公共物品实验中的合作与惩罚。《美国经济评论》90，980–994。

Feigenbaum, J., Shenker, S., 2004。分布式算法机制设计：近期成果与未来方向，载于：《理论计算机科学当前趋势：新世纪的挑战》第1卷《算法与复杂性》；第2卷《形式模型与语义》，World Scientific，第403–434页。Feldman, V., Guruswami, V., Raghavendra, P., Wu, Y., 2012。半空间对单项式的不可知学习是困难的。SIAM Journal on Computing 41，1558–1590。FeldmanHall, O., Mobbs, D., 2015。一种用于道德决策的神经网络。FeldmanHall, O., & Mobbs, D.(2015) 一种用于道德决策的神经网络。载于 AW Toga、MD Lieberman（编），《脑成像：百科全书式参考》，爱思唯尔：牛津。Fellows, M.R., 2002。参数化复杂性：核心思想及其与实际计算的联系，载于：《实验算法学》。施普林格，第51–77页。Fischer, M.J., Ladner, R.E., 1979年。正则程序的命题动态逻辑。《计算机与系统科学杂志》18，194–211。Floridi, L., Sanders, J.W., 2004。论人工智能体的道德性。Minds and Machines 14，349–379。Foot, P., 1967年。堕胎问题与双重效果学说。《牛津评论》5。Foster, D.P., Young, H.P., 2001。论预测理性代理行为的不可能性。《美国国家科学院院刊》98，12848–12853。Friston, K., 2010年。自由能原理：一种统一脑理论？《自然评论：神经科学》11，127–138。Frixione, M., 2001。可处理能力。Minds and Machines 11，379–397。Furbach, U., Schon, C., Stolzenburg, F., 2014。道义逻辑中的自动推理，载于：人工智能多学科趋势国际研讨会，施普林格，第57–68页。Gabbay, D., Horty, J., Parent, X., van der Meyden, R., van der Torre, L., 2013。《道义逻辑与规范系统手册》。Gabriel, I., 2020。人工智能、价值观与对齐。Minds and Machines 30，411–437。Garcıa, J., Fern´andez, F., 2015年。安全强化学习综合综述。《机器学习研究杂志》16，1437–1480。Garey, M.R., Johnson, D.S., 1979年。《计算机与不可解性》。第174卷。旧金山：Freeman。Gauthier, D., 1987年。《协商道德》。Clarendon出版社。

Geertz, C., et al., 1973年。《文化的解释》，第5019卷。Basic Books。Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013年。《贝叶斯数据分析》。Gilboa, I., Zemel, E., 1989年。纳什均衡与相关均衡：一些复杂性考量。《博弈与经济行为》1，80–93。Gill, J., 1977年。概率图灵机的计算复杂性。《SIAM Journal on Computing》6，675–695。Giraud-Carrier, C., Provost, F., 2005年。迈向元学习的辩护：无免费午餐定理是否构成阻碍？载于：《ICML-2005元学习研讨会论文集》，第12–19页。Giubilini, A., Savulescu, J., 2018年。人工道德顾问：当“理想观察者”遇上人工智能。《哲学与技术》31，169–188。Gmytrasiewicz, P.J., Doshi, P., 2005年。多智能体环境中的序贯规划框架。《人工智能研究杂志》24，49–79。G¨odel, K., 1930年。《论逻辑演算的完备性》。博士论文。博士论文，维也纳大学。¨G¨odel, K., 1931年。《论〈数学原理〉及相关系统中形式上不可判定的命题 I》。《数学与物理月刊》38，173–198。Goodie, A.S., Doshi, P., Young, D.L., 2012年。竞争性博弈中心智理论推理的层次。《行为决策杂志》25，95–108。Governatori, G., Olivieri, F., Rotolo, A., Scannapieco, S., 2013年。计算可撤销逻辑中的强许可与弱许可。《哲学逻辑杂志》42，799–829。Govindarajulu, N.S., Bringsjord, S., 2017年。关于自动化双重效果原则，载于：第26届国际人工智能联合会议论文集，第4722–4730页。Govindarajulu, N.S., Bringsjord, S., Ghosh, R., Sarathy, V., 2019年。迈向有德机器的工程学，载于：2019年AAAI/ACM人工智能、伦理与社会会议论文集，第29–35页。Greene, J.D., 2007年。为什么腹内侧前额叶皮层患者更偏向功利主义？双加工理论对道德判断的解释。《认知科学趋势》11，322–323。Greene, J.D., 2014年。超越即射即判式道德观：认知（神经）科学为何与伦理学相关。《伦理学》124，695–726。Greene, J.D., Morelli, S.A., Lowenberg, K., Nystrom, L.E., Cohen, J.D., 2008年。认知负荷会选择性地干扰功利主义道德判断。《认知》107，1144–1154。

Grice, H.P., 1975年。逻辑与会话，载于：《言语行为》。Brill，第41–58页。Griffiths, T.L., Tenenbaum, J.B., 2005年。因果归纳中的结构与强度。《认知心理学》51，334–384。Guha, S., Munagala, K., Shi, P., 2010年。躁动多臂老虎机问题的近似算法。《美国计算机协会会刊》（JACM）58，1–50。Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z., 2019年。可解释人工智能。《科学·机器人学》4，eaay7120。Hagendorff, T., Danks, D., 2022。《构建具有道德导向的人工智能系统的伦理与方法论挑战》。《人工智能与伦理》，1–14。Haigh, T., Priestley, M., Rope, C., 2014年。洛斯阿拉莫斯押注ENIAC：1947—1948年的核蒙特卡洛模拟。《IEEE计算史年刊》36，42–63。Hajek, B., 2015年。《工程师的随机过程》。剑桥大学出版社。Halpern, J.Y., Moses, Y., 1992年。知识与信念模态逻辑的完备性与复杂性指南。《人工智能》54，319–379。Hanneke, S., 2016年。PAC学习的最优样本复杂度。《机器学习研究期刊》17，1319–1333。Hansen, J., 2008年。优先级条件命令：问题与新提议。《自主智能体与多智能体系统》17，11–35。Hare, R.M., 1952年。《道德的语言》。Hare, R.M., 1981年。《道德思维：层次、方法与目的》。牛津：Clarendon出版社；纽约：牛津大学出版社。Harnad, S., 1990年。符号奠基问题。《物理学D：非线性现象》42，335–346。Harsanyi, J.C., 1967年。不完全信息博弈：由“贝叶斯”博弈参与者进行的博弈，I–III，第一部分：基本模型。《管理科学》14，159–182。Hart, S., Mas-Colell, A., 2000年。一种导向相关均衡的简单自适应程序。《计量经济学》68，1127–1150。Hart, S., Schmeidler, D., 1989年。相关均衡的存在性。《运筹学数学》14，18–25。H˚astad, J., 2001年。一些最优不可近似性结果。《美国计算机协会会刊》（JACM）48，798–859。Hazan, E., Krauthgamer, R., 2011年。近似最佳纳什均衡有多难？《SIAM Journal on Computing》40，79–91。

Hedden, T., Zhang, J., 2002年。《你认为我认为你认为是什么？：矩阵博弈中的策略推理》。Cognition 85, 1–36. Hellstr¨om, T., 2013年。《军事机器人的道德责任》。《伦理学与信息技术》15, 99–107. Herken, R., 1995。《通用图灵机：半世纪回顾》。施普林格出版社。Herzig, A., Schwarzentruber, F., 2008年。《个体与群体能动性逻辑的性质》。《模态逻辑进展》7, 133–149. Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., 等，2018年。《从示范中进行深度Q学习》，载于：AAAI人工智能会议论文集。Heuer, L., Orland, A., 2019年。《囚徒困境中的合作：纯策略与混合策略的实验比较》。《皇家学会开放科学》6, 182142. Hew, P.C., 2014年。《以可预见的技术而言，人工道德主体无法实现》。《伦理学与信息技术》16, 197–206. Himma, K.E., 2009。《人工能动性、意识与道德能动性标准：人工智能体必须具备哪些性质才能成为道德主体？》。《伦理学与信息技术》11, 19–29. 霍布斯, T., 1651。《利维坦》。Hofstadter, D., 2002。《直视艾美的双眼——并做到》。《创造力、认知与知识：一种互动》, 67. Hohwy, J., 2013。《预测心智》。牛津大学出版社。Holt, C.A., Roth, A.E., 2004。《纳什均衡：一种观点》。《美国国家科学院院刊》101, 3999–4002. Hooker, B., 2016。《规则后果主义》，载于：Zalta, E.N.（编），《斯坦福哲学百科全书》。2016年冬季版。斯坦福大学形而上学研究实验室。Hopkins, M., Kane, D.M., Lovett, S., Mahajan, G., 2022。《可实现学习就是你所需要的一切》，载于：学习理论会议，PMLR。第3015–3069页。Horty, J.F., 2001。《能动性与道义逻辑》。牛津大学出版社。Horty, J.F., 2012。《作为默认规则的理由》。牛津大学出版社（美国）。Howard, D., Muntean, I., 2017。《人工道德认知：道德功能主义与自主道德能动性》。施普林格。第121–159页。休谟, D., 1739。《人性论》。Courier Corporation，2003年。

Hummert, S., Bohl, K., Basanta, D., Deutsch, A., Werner, S., Theißen, G., Schroeter, A., Schuster, S., 2014年。《进化博弈论：作为博弈参与者的细胞》。《Molecular BioSystems》10, 3044–3065. Hursthouse, R., 1999。《论美德伦理学》。牛津大学出版社。Hurwicz, L., Reiter, S., 2006。《经济机制设计》。剑桥大学出版社。Huttegger, S.M., 2007年。《进化与意义的解释》。《科学哲学》74, 1–27. Immerman, N., 1982。《可在多项式时间内计算的关系查询》，载于：第十四届ACM计算理论研讨会论文集，第147–152页。Immerman, N., 1989。《描述复杂性与计算复杂性》，载于：《计算复杂性理论》，J. Hartmanis 编，《应用数学研讨会论文集》，第75–91页。Immerman, N., 1998。《描述复杂性》。施普林格科学与商业媒体。Johnson, V.R., 1990。《1789年〈人与公民权利宣言〉、恐怖统治与巴黎革命法庭》。《BC国际与比较法评论》13, 1. Jørgensen, J., 1937年。《命令句与逻辑》。《Erkenntnis》7, 288–296. Kahneman, D., 2011年。《思考，快与慢》。麦克米伦。Kakade, S.M., 2003。《强化学习的样本复杂度》。伦敦大学，伦敦大学学院（英国）。康德, I., 1785。《道德形而上学奠基》。劳特利奇（2013）。卡普, R.M., 1972年。《组合问题之间的可归约性》，载于：《计算机计算复杂性》。施普林格，第85–103页。Kawaguchi, K., Kaelbling, L.P., Bengio, Y., 2017年。《深度学习中的泛化》。arXiv预印本 arXiv:1710.05468。Kaye, P., Laflamme, R., Mosca, M., 2006。《量子计算导论》。牛津大学出版社。Kearns, M., Valiant, L., 1994年。《学习布尔公式与有限自动机的密码学限制》。《美国计算机协会会刊》41, 67–95. Kearns, M.J., Schapire, R.E., Sellie, L.M., 1992年。《迈向高效的不可知学习》，载于：第五届计算学习理论年度研讨会论文集，第341–352页。Keller, G.B., Mrsic-Flogel, T.D., 2018年。《预测处理：一种典型的皮层计算》。《神经元》100, 424–435. Kemp, C., Tenenbaum, J.B., 2008年。《结构形式的发现》。《美国国家科学院院刊》105, 10687–10692.

Khachiyan, L.G., 1979年。《线性规划中的多项式算法》，载于：《苏联科学院通报》，俄罗斯科学院。第1093–1096页。Kifer, D., Machanavajjhala, A., 2011年。《数据隐私领域不存在免费午餐定理》，载于：2011年 ACM SIGMOD 数据管理国际会议论文集，第193–204页。Koenig, S., Simmons, R.G., 1993年。《实时强化学习的复杂性分析》，载于：AAAI，第99–107页。Kohlberg, L., Hersh, R.H., 1977年。《道德发展：理论综述》。《理论与实践》16，53–59。K¨ording, K.P., Wolpert, D.M., 2006。《感觉运动控制中的贝叶斯决策理论》。《认知科学趋势》10，319–326。Koutsoupias, E., Papadimitriou, C., 2009。《最坏情况均衡》。《计算机科学评论》3，65–69。Kreps, D.M., Milgrom, P., Roberts, J., Wilson, R., 1982年。《有限次重复囚徒困境中的理性合作》。《经济理论杂志》27，245–252。Kripke, S., 1976年。《真理论纲要》。《哲学杂志》72，690–716。Kripke, S.A., 1963。《模态逻辑的语义分析 I：正规模态命题演算》。《数学逻辑季刊》9，67–96。Kwisthout, J., 2011年。《贝叶斯网络中的最大概率解释：复杂度与可处理性》。《国际近似推理杂志》52，1452–1469。Kwisthout, J., Wareham, T., Van Rooij, I., 2011年。《贝叶斯不可处理性并非近似方法可以治愈的顽疾》。《认知科学》35，779–784。L Griffiths, T., Kemp, C., B Tenenbaum, J., 2008。《认知的贝叶斯模型》。Ladner, R.E., 1977年。《模态命题逻辑系统中可证性的计算复杂度》。《SIAM Journal on Computing》6，467–480。Langford, J., Zhang, T., 2007年。《上下文多臂老虎机的分阶段贪心算法》。《神经信息处理系统进展》20，96–1。Lattimore, T., Hutter, M., 2013年。《监督学习中的无免费午餐定理与奥卡姆剃刀》，载于：《算法概率及其相关主题》。《贝叶斯预测与人工智能》。施普林格，第223–235页。Lau, H.C., 2007年。《意识的高阶贝叶斯决策理论》。《脑研究进展》168，35–48。Leben, D., 2018年。《机器人伦理学：如何设计道德算法》。劳特利奇。Lee, J., Bahri, Y., Novak, R., Schoenholz, S.S., Pennington, J., Sohl-Dickstein, J., 2017年。《深度神经网络即高斯过程》。arXiv预印本 arXiv:1711.00165。

Lenman, J., 2000年。《结果主义与茫然无措》。《哲学与公共事务》29，342–370。网址： http://www.jstor.org/stable/2672830。Levesque, H.J., 1989年。《逻辑与推理的复杂度》，载于：《哲学逻辑与人工智能》。施普林格，第73–107页。Lewis, D., 1969年。《惯例》。剑桥：哈佛大学出版社。Lewis, D., 1975年。《语言与语言》。Lewis, D., 1979年。《语言游戏中的记分》，载于：《来自不同视角的语义学》。施普林格，第172–187页。Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., 等，2022。《语言模型的整体评估》。arXiv预印本 arXiv:2211.09110。Lieder, F., Griffiths, T.L., 2020年。《资源理性分析：将人类认知理解为对有限计算资源的最优利用》。《行为与脑科学》43。Lindner, F., Mattm¨uller, R., Nebel, B., 2020年。《行动计划道德可允许性的评估》。《人工智能》287，103350。网址： https://www.sciencedirect.com/science/article/pii/S0004370219301043，doi:https://doi.org/10.1016/j.artint.2020.103350。Lipton, R.J., Markakis, E., Mehta, A., 2003年。《运用简单策略进行大型博弈》，载于：第4届 ACM电子商务会议论文集，第36–41页。Littman, M.L., 1996年。《序贯决策算法》。布朗大学。Littman, M.L., Goldsmith, J., Mundhenk, M., 1998年。《概率规划的计算复杂度》。《人工智能研究杂志》9，1–36。Locatelli, A., Gutzeit, M., Carpentier, A., 2016年。《阈值多臂老虎机问题的最优算法》，载于：国际机器学习会议，PMLR。第1690–1698页。Lorenz, E.N., 1963。《确定性非周期流》。《大气科学杂志》20，130–141。Lucas, P.J., de Bruijn, N.C., Schurink, K., Hoepelman, A., 2000年。《一种管理重症监护病房传染病的概率与决策理论方法》。《医学中的人工智能》19，251–279。Luce, R.D., Raiffa, H., 1989年。《博弈与决策：导论与批判性综述》。Courier Corporation。Mabaso, B.A., 2021年。《计算理性智能体能够成为道德智能体》。《伦理学与信息技术》23，137–145。

Madani, O., Hanks, S., Condon, A., 2003年。《概率规划及相关随机优化问题的不可判定性》。人工智能 147, 5–34。Makinson, D., Van Der Torre, L., 2000年。《输入/输出逻辑》。哲学逻辑杂志 c

29, 383–408。Malle, B., Scheutz, M., Austerweil, J., 2017年a。《人类 n

与机器人智能体的社会规范与道德规范网络》，载于：《机器人世界》。智能系统、控制与自动化 n:

科学与工程。施普林格，Cham。第84卷，第3–17页。Malle, B.F., Scheutz, M., Austerweil, J.L., 2017b。《人类与机器人智能体的社会规范与道德规范网络》，载于：《机器人世界》。施普林格，第3–17页。Marr, D., 1977年。《人工智能——个人观点》。《人工智能》9, 37–48。Marr, D., 1981年。《视觉：对人类表征的计算研究，以及 d

视觉信息的处理。美国加利福尼亚州旧金山：W. H. Freeman。Mertens, J.F., Zamir, S., 1985年。针对不完 e

全信息博弈的贝叶斯分析公式。《国际博弈论杂志》14，1–29。Meyer, J.J.C., 等，1988年。道义逻辑的一种不同方法：将道义逻辑视为 a

动态逻辑的一种变体。《圣母大学形式逻辑期刊》29，109–136。Mill, J.S., 1859年。《论自由》。牛津大学出版社，美国（1998年）。Mill, J.S., 1861年。《功利主义》。纽约：牛津大学出版社（1998年）。Moka-Mubelo, W., 2017年。《法律与道德》，载于：《在人类 n

权利话语》。施普林格，第51–88页。Mundhenk, M., Goldsmith, J., Lusena, C., Allender, E., 2000年。有限时域的复杂度 n

马尔可夫决策过程问题。《美国计算机协会会刊》（JACM）47，681–720。N. J. A. Sloane, R.G.W.，2022。在线整数数列百科全书中的条目 a007526 s.

网址： https://oeis.org/A007526。纳维森，J.，2001。《自由意志主义理念》。Broadview出版社。纳什，J.，1951。《非合作博弈》。《数学年刊》，286–295。Nash, J.F. 等，1950。《n人博弈中的均衡点》。美国国家科学院院刊 l

36，48–49。v. Neumann, J.，1928。《关于博弈论》。《数学年刊》10 0,

295–320。Newen, A.、De Bruin, L.、Gallagher, S.，2018。《牛津4E认知手册》。牛津 d

大学出版社。

Neyman, A., 1985年。有界复杂度为有限次重复囚徒困境中的合作提供了正当理由。《经济学快报》19, 227–229。Neyshabur, B., Bhojanapalli, S., McAllester, D., Srebro, N., 2017年。探索深度学习中的泛化。《神经信息处理系统进展》30。Ng, A.Y., Russell, S.J., 2000年。逆向强化学习算法，载于：国际机器学习大会，p. 2。Niedermeier, R., 2006年。《固定参数算法导论》。第31卷。牛津大学出版社。Nievergelt, J., Gasser, R., M¨aser, F., Wirth, C., 1995。大海捞针：穷举搜索能否克服组合混沌？施普林格·柏林·海德堡，柏林，海德堡。第254–274页。URL: https://doi.org/10.1007/BFb0015248, doi:10.1007/BFb0015248。Nisan, N., Ronen, A., 1999年。算法机制设计，载于：第三十一届年度ACM计算理论研讨会论文集，第129–140页。Nowak, M.A., 2006。合作演化的五条规则。《科学》314, 1560–1563。Nussbaum, M.C., 1988。非相对美德：一种亚里士多德式方法。《中西部哲学研究》13, 32–53。Oaksford, M., Chater, N., 2001。人类推理的概率方法。《认知科学趋势》5, 349–357。Papadimitriou, C.H., 1994。奇偶性论证及其他低效存在性证明的复杂度。《计算机与系统科学杂志》48, 498–532。Papadimitriou, C.H., Roughgarden, T., 2008。计算多人博弈中的相关均衡。《美国计算机协会会刊》55, 1–29。Papadimitriou, C.H., Tsitsiklis, J.N., 1987。马尔可夫决策过程的复杂度。《运筹学数学》12, 441–450。Papadimitriou, C.H., Tsitsiklis, J.N., 1994。最优排队网络控制的复杂度，载于：IEEE第九届年度复杂性理论结构会议论文集，IEEE。第318–322页。Parberry, I., Garey, M.R., Meyer, A., 1994。电路复杂度与神经网络。麻省理工学院出版社。Parfit, D., 1984。理由与人格。牛津大学出版社。Parfit, D., 2011。《论重要之事》。第1卷。牛津大学出版社。Park, J.D., Darwiche, A., 2004。最大后验解释的复杂性结果与近似策略。《人工智能研究杂志》21, 101–133。

Pearl, J., 1985. 贝叶斯网络：一种用于证据推理的自激活记忆模型，载于：《认知科学学会第7届会议论文集》，加州大学欧文分校，美国加利福尼亚州，第15–17页。Pearl, J., 1987. 运用因果模型的随机模拟进行证据推理。《人工智能》32，245–257。Pearl, J., 2022. 贝叶斯牧师论推理引擎：一种分布式层级方法，载于：《概率与因果推断：Judea Pearl文集》，第129–138页。Pereira, L.M., Saptawijaya, A., 2009. 用前瞻逻辑建模道德。《基于推理的智能系统国际期刊》1，209–221。Pitt, L., Valiant, L.G., 1988. 从实例中学习的计算限制。《美国计算机协会会刊》（JACM）35，965–984。Pnueli, A., 1977. 程序的时序逻辑，载于：《第18届计算机科学基础年会》（sfcs 1977），IEEE，第46–57页。Pontier, M., Hoorn, J., 2012. 迈向行为比人类更合乎伦理的机器，载于：认知科学学会年会论文集。Popper, K.R., 1962. 《科学哲学：猜想与反驳》。科学知识的增长……第140卷。Basic Books，纽约，1962年。Pratt, V.R., 1976. 关于弗洛伊德—霍尔逻辑的语义考量，载于：《第17届计算机科学基础年会》（sfcs 1976），IEEE，第109–121页。Pratt, V.R., 1980. 一种近乎最优的行动推理方法。《计算机与系统科学杂志》20，231–254。Purves, D., Jenkins, R., Strawser, B.J., 2015. 自主机器、道德判断与基于正当理由的行动。《伦理理论与道德实践》18，851–872。Putnam, H., 2004. 《事实/价值二分的瓦解及其他论文》。哈佛大学出版社。Radner, R., 1986. 有限理性能够化解囚徒困境吗？《纪念Gerard Debreu的论文集》，387–399。Rathnasabapathy, B., Doshi, P., Gmytrasiewicz, P., 2006. 利用行为等价求解交互式部分可观测马尔可夫决策过程的精确解，载于：《第五届自主代理与多智能体系统国际联合会议论文集》，第1025–1032页。Rawls, J., 1980. 道德理论中的康德式建构主义。《哲学杂志》77，515–572。Rawls, J.B., 1971. 《正义论》。

Reinikainen, J., 2005. 黄金法则与普遍化要求。《价值探究杂志》39，155。Rest, J.R., Narvaez, D., Thoma, S.J., Bebeau, M.J., 1999. Dit2：设计并测试一种修订后的道德判断测量工具。《教育心理学杂志》91，644。Reynolds, C., 2005. 行动评价的计算复杂性，载于：《第6届国际计算机伦理：哲学探究会议》（特温特大学，恩斯赫德，荷兰，2005年），Citeseer。Rosser, B., 1936. 哥德尔和丘奇若干定理的扩展。《符号逻辑杂志》1，87–91。Roth, D., 1996. 近似推理的困难性。《人工智能》82，273–302。Roughgarden, T., 2005. 自私路由与无政府代价。麻省理工学院出版社。Roughgarden, T., Tardos, ´E., 2002. 自私路由有多糟糕？《美国计算机协会会刊》（JACM）49，236–259。Rubinstein, A., 1998. 有限理性建模。麻省理工学院出版社。Russell, S.J., Subramanian, D., 1994. 可证明的有界最优智能体。《人工智能研究杂志》2，575–609。Scanlon, T.M., 2000.《我们彼此负有什么义务》。哈佛大学出版社。Scarselli, F., Tsoi, A.C., 1998. 使用前馈神经网络实现通用逼近：现有方法综述及一些新结果。《神经网络》11，15–37。Schaffer, C., 1994. 泛化性能的守恒定律，载于：《1994年机器学习会议论文集》。爱思唯尔，第259–265页。Schelling, T.C., 1960.《冲突的策略：作者新序》。哈佛大学出版社。Schiffer, S.R., 1972.《意义》。Schnoebelen, P., 2002. 时序逻辑模型检验的复杂度。《模态逻辑进展》4，35。Schroeder, M., 2010.《支持：评估表达主义的语义学纲领》。牛津大学出版社。Schurz, G., 2017. 无免费午餐定理、归纳怀疑论与元归纳的最优性。《科学哲学》84，825–839。Schwarzentruber, F., Semmling, C., 2014. STIT具有危险的不可判定性，载于：ECAI。Searle, J.R., 1980. 心智、大脑与程序。《行为与脑科学》3，417–424。

Searle, J.R., 1992年。《心灵的重新发现》。麻省理工学院出版社。Sen, A., 1979年。功利主义与福利主义。《哲学杂志》76, 463–489。Serfozo, R.F., 1979年。连续时间与离散时间马尔可夫决策过程之间的一种等价关系。《运筹学》27, 616–620。Sergot, M., 1998年。规范立场。《规范、逻辑与信息系统》49, 289–308。Shim, J., Arkin, R., Pettinatti, M., 2017年。用于患者与照护者关系中机器人调解者的介入式伦理监管器：实现与评估，载于：2017年IEEE机器人与自动化国际会议（ICRA），IEEE，纽约，美国。第2936–2942页。Shimony, S.E., 1994年。为信念网络寻找地图是NP难问题。《人工智能》68, 399–410。Sidgwick, H., 1907年。《伦理学方法》。Good Press（2019）。Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T.等，2018年。一种通过自我对弈精通国际象棋、将棋和围棋的通用强化学习算法。《科学》362, 1140–1144。Silver, D., Singh, S., Precup, D., Sutton, R.S., 2021年。奖励已足够。《人工智能》299, 103535。Simon, H.A., 1955年。理性选择的行为模型。《经济学季刊》69, 99–118。Simon, H.A., 1990年。有限理性，载于：《效用与概率》。施普林格，第15–18页。Singer, M.G., 2002年。理性道德的理想：哲学论集。牛津大学出版社。Singer, P., 2011年。《实践伦理学》。剑桥大学出版社。Sinnott-Armstrong, W., 2021年。结果主义，载于：Zalta, E.N.（编），《斯坦福哲学百科全书》。2021年秋季版。形而上学研究实验室，斯坦福大学。Sistla, A.P., Clarke, E.M., 1985年。命题线性时序逻辑的复杂度。《美国计算机协会会刊》（JACM）32, 733–749。Skyrms, B., 2004年。《猎鹿博弈与社会结构的进化》。剑桥大学出版社。Skyrms, B., 2010年。《信号：进化、学习与信息》。牛津大学出版社。Slivkins, A., 2019年。《多臂老虎机导论》。arXiv预印本 arXiv:1904.07272。Smart, J.J.C., 1956年。极端功利主义与受限功利主义。《哲学季刊》（1950-）6, 344–354。

Smith, J.M., Price, G.R., 1973年。动物冲突的逻辑。《自然》246, 15–18。doi:10.1038/246015a0。Sobel, D.M., Kirkham, N.Z., 2006年。Blicket因果线索与婴儿：幼儿和婴儿因果推理能力的发展。《发展心理学》42, 1103。Spaan, E., 1993年。命题时态逻辑的复杂度，载于：《钻石与默认规则》。施普林格，第287–307页。Spaan, E., 2016年。模态逻辑的复杂度。博士论文。阿姆斯特丹大学。Stace, W.T., 1937年。道德概念。Stenseke, J., 2021年。人工德性智能体：从理论到机器实现。AI & SOCIETY doi:10.1007/s00146-021-01325-7。Stenseke, J., 2022a。多智能体公地悲剧中的人工德性智能体。AI & SOCIETY，1–18。Stenseke, J., 2022b。道德机器语境中的跨学科混淆与化解。《科学与工程伦理》28, 1–17。Stenseke, J., 2023年。规范伦理学在道德机器中的运用与滥用，载于：《社会制度中的社会机器人》。IOS出版社，第155–164页。Stenseke, J., Balkenius, C., 2022年。评估伦理算法的时间效率，载于：《CEUR研讨会论文集》，CEUR-WS。Sterkenburg, T.F., Gr¨unwald, P.D., 2021年。监督学习的无免费午餐定理。Synthese 199, 9979–10015。Stevenson, C.L., 1937年。伦理术语的情感意义。《心灵》46, 14–31。Steyvers, M., Tenenbaum, J.B., Wagenmakers, E.J., Blum, B., 2003年。根据观测和干预推断因果网络。《认知科学》27, 453–489。Stocker, M., 1977年。现代伦理理论的精神分裂。《哲学杂志》73, 453–466。Sun, X., Robaldo, L., 2017年。输入／输出逻辑的复杂度。《应用逻辑杂志》25, 69–88。Taddeo, M., Floridi, L., 2005年。解决符号奠基问题：对十五年研究的批判性回顾。《实验与理论人工智能杂志》17, 419–445。Tarski, A., 1944年。真理的语义观：兼论语义学基础。《哲学与现象学研究》4, 341–376。Tenbrunsel, A.E., Messick, D.M., 2004年。伦理淡化：自我欺骗在不道德行为中的作用。《社会正义研究》17, 223–236。

Tenenbaum, J.B., Griffiths, T.L., Kemp, C., 2006。基于理论的贝叶斯归纳学习与推理模型。《认知科学趋势》10，309–318。Thornton, S.M., Pan, S., Erlien, S.M., Gerdes, J.C., 2016。将伦理考量纳入自动驾驶车辆控制。IEEE《智能交通系统汇刊》18，1429–1439。Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A., 2020。机器伦理学的实现：一项综述。《ACM计算调查》（CSUR）53，1–38。Tr´emoli`ere, B., Neys, W.D., Bonnefon, J.F., 2012。死亡显著性与道德：思考死亡会使人们变得不那么功利主义。Cognition 124，379–384。网址： https://www.sciencedirect.com/science/article/pii/S0010027712001035，doi:https://doi.org/10.1016/j.cognition.2012.05.011。Tsotsos, J.K., 1990。在复杂度层级上分析视觉。Behavioral and Brain Sciences 13，423–445。Tufi¸s, M., Ganascia, J.G., 2015。在BDI智能体模型中植入规范，载于：《机器人伦理系统构建手册》。施普林格，第119–133页。Turing, A.M., 1936年。论可计算数及其在判定问题中的应用。《数学杂志》58，5。Ullmann-Margalit, E., 2015。规范的涌现。牛津大学出版社。Urbano, A., Vila, J. E., 2002。计算复杂性与交流：双人博弈中的协调。《计量经济学》70，1893–1927。Valiant, L.G., 1984。可学习理论。《ACM通讯》27，1134–1142。Vallor, S., 2015。新机器时代的道德技能退化与提升：对品格模糊未来的思考。《哲学与技术》28，107–124。Vallor, S., 2016。技术与美德：一部关于值得期待的未来的哲学指南。牛津大学出版社。Van Der Meyden, R., 1996。许可的动态逻辑。《逻辑与计算杂志》6，465–479。Van Rooij, I., 2008。可处理认知论题。《认知科学》32，939–984。Van Rooij, I., Blokpoel, M., Kwisthout, J., Wareham, T., 2019。认知与不可处理性：经典与参数化复杂性分析指南。剑桥大学出版社。Vanmarcke, E., 2010。随机场：分析与综合。World Scientific。Vapnik, V., 1999。统计学习理论的本质。施普林格科学与商业媒体。Vapnik, V., Chervonenkis, A., 1974。模式识别理论。

Vapnik, V.N., Chervonenkis, A.Y., 2015。论事件相对频率向其概率的一致收敛，载于：复杂性度量。施普林格，第11–30页。Vardi, M.Y., 1982。关系查询语言的复杂性，载于：第十四届ACM计算理论年度研讨会论文集，第137–146页。Vardi, M.Y., 1997。为什么模态逻辑具有如此强的可判定性？技术报告。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, ´L., Polosukhin, I., 2017。注意力就是一切。《神经信息处理系统进展》30。Vollmer, H., Schnoor, I., Schnoor, H., Schneider, T., Bauland, M., 2009。线性时态逻辑中广义可满足性的复杂性。《计算机科学中的逻辑方法》5。Von Luxburg, U., Sch¨olkopf, B., 2011。统计学习理论：模型、概念与结果，载于：《逻辑史手册》。爱思唯尔。第10卷，第651–706页。Von Neumann, J., Morgenstern, O., 1947。博弈论与经济行为，第2次修订版。Von Wright, G.H., 1951年。道义逻辑。《心灵》60，1–15。Vul, E., Goodman, N., Griffiths, T.L., Tenenbaum, J.B., 2014。一次就够？从极少样本中作出最优决策。《认知科学》38，599–637。Wallach, W., Allen, C., 2008。道德机器：教机器人明辨是非。牛津大学出版社。Wattles, J., 1996。黄金法则。牛津大学出版社。de Weerd, H., Verbrugge, R., Verheij, B., 2013。了解她知道你知道什么能带来多大帮助？一项基于智能体的仿真研究。人工智能199，67–92。de Weerd, H., Verbrugge, R., Verheij, B., 2017。与其他心智谈判：递归心智理论在不完全信息谈判中的作用。《自主智能体与多智能体系统》31，250–287。Whitehead, S.D., 1991。强化学习中合作机制的复杂性分析，载于：AAAI，第607–613页。Whittle, P., 1988。无休止多臂老虎机：变化世界中的活动分配。《应用概率杂志》25，287–298。Wiegel, V., van den Berg, J., 2009。结合道德理论、模态逻辑与多智能体系统，创建行为良好的人工智能体。《国际社会机器人学杂志》1，233–242。Williams, B., 2006。《伦理学与哲学的界限》。劳特利奇。Williamson, D.P., Sh moys, D.B., 2011。近似算法的设计。剑桥大学出版社。

Wittgenstein, L., 2010年。《哲学研究》。约翰·威立父子公司。Wolpert, D.H., 1992年。样本内测试与泛化误差之间的联系。《复杂系统》6, 47。Wolpert, D.H., 1996年。学习算法之间不存在先验区分。《神经计算》8, 1341–1390。Wolpert, D.H., 2002年。监督学习无免费午餐定理。《软计算与工业》，25–42。Wolpert, D.H., Macready, W.G., 1997年。优化无免费午餐定理。《IEEE进化计算汇刊》1, 67–82。Xu, M., 1998年。审议型STIT公理。《哲学逻辑杂志》27, 505–552。Yi, S.K.M., Steyvers, M., Lee, M.D., Dry, M.J., 2012。《组合问题中的群体智慧》。《认知科学》36, 452–470。Yoshida, W., Dolan, R.J., Friston, K.J., 2008年。心智博弈论。《PLoS计算生物学》4, e1000254。Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021年。理解深度学习（仍然）需要重新思考泛化。《ACM通讯》64, 107–115。Zhang, N.L., Poole, D., 1996年。在贝叶斯网络推理中利用因果独立性。《人工智能研究杂志》5, 301–328。Zinkevich, M., Johanson, M., Bowling, M., Piccione, C., 2007年。不完全信息博弈中的遗憾最小化。《神经信息处理系统进展》第20卷。