社会技术系统宏观伦理原则：分类学和未来方向

JESSICA WOODGATE 和 NIRAV AJMERI,英国布里斯托大学

人工智能 (AI) 的迅速采用需要仔细分析其伦理含义。在解决 AI 的伦理问题时,重要的是要检查整个范围内的伦理相关特征,而不是单独查看个体代理。通过将视角转移到嵌入代理的系统,可以了解更广泛的伦理特征,这被封装在社会技术系统 (STS) 的宏观伦理中。从宏观伦理的角度来看,系统治理 这是参与者试图促进反映其价值观的结果和规范的地方是关键。然而,当 STS 的利益相关者具有不同的价值偏好,或者当 STS 中的规范发生冲突时,STS 中可能会出现多用户社会困境。以令人满意的方式解决这些困境的能力可以通过在推理中纳入规范的伦理原则来帮助。规范的伦理原则被理解为从哲学理论中推断出的可操作的规则。因此,伦理原则的分类有利于实践者在推理中利用它们。

这项工作制定了规范伦理原则的分类法,可在STS 治理中加以实施。我们确定了一系列道德原则,在分类树上有 25 个节点。我们描述了之前实施每项原则的方式,并解释了可能出现的潜在困难。我们进一步建议如何将原则的操作化应用于 STS 的宏观伦理。我们设想这种分类法将促进方法论的发展,以将道德原则纳入管理公平 STS 的推理能力。

ACM 参考格式:
杰西卡·伍德盖特和尼拉夫·阿杰梅里。 2023. 社会技术系统宏观伦理学原则:分类学和未来方向。 1, 1（2023 年 2 月）,34 页。 https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 简介

人工智能系统的快速发展需要了解它们的伦理影响[23]。在 AI 的发展中,最近代理（执行动作以实现目标的行为实体,这些目标是使用 AI 做出的决策， [79])研究从强调单一代理到多代理系统（MAS:多个技术代理部署到一个公共环境中， [80]）。研究中对 MAS 的日益关注需要仔细分析其伦理意义[18, 31]。在追求更符合道德的 MAS 的发展过程中,将视角转向社会技术系统 (STS) 很重要,因为它将人为因素纳入了道德推理[76, 87]。在 STS 中,人类和代理作为道德二重奏一起工作,代理代表他们的人类对手行事。在 STS 的背景下,采用宏观（相对于微观)伦理学的观点也很重要[18]。这是因为关注STS 中单个代理人决策的微观伦理学可能过于狭隘而无法考虑整体

作者地址:Jessica Woodgate,yp19484@bristol.ac.uk； Nirav Ajmeri,nirav.ajmeri@bristol.ac.uk,布里斯托大学,英国布里斯托尔,BS8 1UB。

XXXX-XXXX/2023/2-ART 15.00 美元
https://doi.org/10.1145/nnnnnnn.nnnnnnn

一系列与道德相关的特征。因此,宏观伦理通过考虑 STS 的治理采用更广泛的观点。

在 STS 的宏观伦理学视角下,利益相关者通过促进某些符合其价值观（对我们生活重要的东西， [88])的结果和规范（预期行为规则， [74])来治理系统。重视在 STS 治理中纳入规范和价值观是关键,因为伦理应被理解为结合背景的反思性发展过程[57、67、73、102] 。价值观是背景[65]的一个重要方面,因为它们反映了利益相关者的偏好[3, 32]。规范对于上下文也很重要,因为它们可以帮助确保系统的行为与人类价值观一致[72、76、90] 。这是因为可以在代理社会中使用规范来鼓励参与的代理人以可接受的方式行事[74]。因此,从宏观伦理的角度来看,价值观和规范对于伦理推理[2, 30, 91, 100]和 STS 的关键特征至关重要。

然而,用户（这里,用户是利益相关者的同义词)可能有不同的价值偏好,或者他们的价值观可能与规范相冲突[24]。因此,在做出涉及多个用户的决策时出现了挑战[56、61]。这些场景被称为多用户社交困境,可能发生在平凡的环境中。这种普通设置的一个例子可能是智能家居代理决定何时打开暖气,同时考虑现有用户的偏好和其他上下文特征。如果利益相关者有不同的价值偏好,例如一些重视舒适度（因此暗示应该打开暖气)和其他人重视省钱（因此暗示应该关闭暖气） ,则可能会出现两难境地。

在推理中纳入规范的伦理原则可能有助于以公平的总体目标以令人满意的方式解决这些困境[99]。规范伦理学是研究通过使用原则和准则,或对是非标准的理性和系统研究来确定行为的伦理性的实践手段[76]。伦理原则意味着某些逻辑命题必须为真,才能使给定的行动计划符合伦理[54]。因此,伦理原则的应用可能有助于系统地思考困境并促进令人满意的结果[21]。这些原则有助于指导规范性判断,理解不同的观点,并确定具体行动方案的道德允许性[16、64、69、86]。

规范的伦理原则以前已被用于计算机科学中的各种不同应用。对于二进制机器学习算法,Binns [11]和 Leben [59]等作品应用伦理原则来改进公平性考虑。在代理人的决策中， Cointe 等人。 [20]实施道德原则,使代理人能够在特定情况下做出道德判断。伦理原则也可用于改善系统分析中的公平性考虑[21, 86]。通过研究文献如何实施这些原则,有可能在 STS 治理中实施规范的伦理理论。这可能有助于令人满意地解决价值和规范冲突的多用户社会困境。由于需要通过欣赏各种不同的方法来培养道德思维,考虑到每种方法的优点和局限性[15, 83],我们设想道德原则的分类法,包括它们以前是如何实施的,将有助于提高 STS 治理公平性的总体目标。

1.1 道德原则分类的动机因此,这项工作的动机源于需要改进STS

中的道德考虑。在 STS 治理中实施规范的伦理原则可能有助于解决这些问题[99]。伦理原则意味着某些逻辑命题必须为真,才能使给定的行动计划符合伦理[54]。因此,伦理原则的应用可能有助于系统地思考困境并促进令人满意的结果[21]。

这些原则有助于指导规范性判断,理解不同的观点,并确定具体行动方案的道德允许性[16、64、69、86] 。需要通过欣赏各种不同的方法来培养道德思维,同时考虑到每种方法的优点和局限性[15, 83]。因此,我们设想道德原则的分类将有助于这种道德思考。

1.2 相关研究的差距在人工智能伦理的

背景下,有两种类型的原则被提及:(1) 从规范伦理学中推断出来的原则,例如 Leben [59] 中发现的道义论和后果论,以及 (2)改编自其他学科,如医学和生物伦理学,如 Floridi 和Cowls [39]、 Jobin 等人建议的那些学科。 [48], Fjeld 等人。 [37],和 Cheng 等人。 [17]包括慈善、非恶意、自治、正义、公平、非歧视、透明度、责任、隐私、问责制、安全和安保、可解释性、人类对技术的控制以及人类价值观的提升。

为了确保术语的清晰度,我们将规范伦理学中的原则称为伦理原则,以及 Floridi 和 Cowls [39]以及 Jobin 等人强调的原则。 [48]作为 AI 基石。因此,此处定义的伦理原则是既定的哲学理论,可以在推理能力中加以运用,因为它们暗示了某些逻辑命题,这些命题必须为真才能使给定的行动计划符合伦理,Kim 等人。 [54]旨。 C8 这些原则大致分为义务论原则(那些要求遵守规则、法律和规范的原则,Hagendorff 等人[44]解释)和目的论原则(那些从好的或可取的事物中衍生出责任或道德义务的事物,作为结束实现,大英百科全书[14]讨论)。另一方面,AI Keystones 是道德原则在其应用中可能针对的更普遍的主题。为了说明这种区别,例如,平等主义的原则支持人类在某种基本意义上是平等的观念,正如 Binns [11] 所解释的那样。为了实现公平的人工智能基石,可以通过加大努力避免不平等来实施平等主义。这可以采取规则的形式,即机会必须对所有申请人平等开放,Lee 等人。 [60] 建议。

AI Keystones主题,例如慈善、非恶意、自治、正义、公平、非歧视、透明度、责任、隐私、问责制、安全和安保、解释能力、人类对技术的控制以及应该支持设计的人类价值观的提升人工智能技术。

伦理原则从道义论和后果论等哲学理论中推断出的可操作规则。

现有的分类法和调查存在于 AI 基石的相关但不同的领域,例如 Jobin 等人。 [48]、 Floridi 和 Cowls [39]以及 Khan 等人。 [53],然而,不是在这里定义的道德原则。 Tolmeijer 等人的工作。 [96]概述了机器伦理的实施,为实施伦理和评估系统的技术和非技术方面提供了有用的指导。然而,作者并没有捕捉到我们捕捉到的所有道德原则。此外,Tolmeijer 等人。考虑更大范围的机器伦理,而不是我们旨在解决的 AI 伦理和 MAS。同样,Yu 等人。 [101]确定了道德原则的高层次概述,但未能认识到在我们的工作中发现的范围,并且没有在相同的深度水平上考虑它们。 Dignum [29]、 Leben [59]、 Robbins和 Wallace [83]对规范伦理进行了总结,然而,为了实现更广泛的适用性,这些作品可能受益于正式的分类法,包括计算机科学中出现的其他伦理原则。

1.3 组织第 2 节简要

解释了方法。这可能有助于未来的研究,通过复制此处使用的方法来扩展伦理原则的分类。第 3 节探讨了我们从客观Qp 中得出的发现,以及迄今为止在计算机科学文献中提出的伦理原则。第 4 节检查目标Qo,研究道德原则之前是如何实施的,以及寻求实施原则的从业者应该采取哪些步骤。第 6 节侧重于目标Qg,解释了我们在计算机科学和人工智能中实施伦理原则方面发现的差距。第 7 节总结了我们的要点。

2 重现性方法简介

从软件工程研究中汲取灵感,例如 Lo 等人。 [66],为了分类法的可重复性和可扩展性,我们遵循 Kitchenham 的[55]指导方针进行系统的文献综述,以制定我们的道德原则分类法。

2.1 目标我们的主要目

标是调查当前对人工智能和计算机科学伦理原则的理解,以及这些原则是如何实施的。具体来说,我们解决以下问题： Qp （原则）。迄今为止,计算机科学文献中提出了哪些伦理原则?

这个问题的目的是帮助识别目前在人工智能和计算机科学领域的文献中使用的原则。由于哲学话语的复杂性,我们遵循 Tolmeijer 等人的方法。 [96]提供关于每个原则如何在文献中定义的简要概述,而不是试图介绍道德哲学。

Qo （运作化）。伦理原则是如何在人工智能和计算机中实施的
科学研究?

这个问题着眼于确定的原则,以检查它们如何在人工智能和计算机科学中得到应用。 Leben [59]和 Tolmeijer 等人的作品。 [96]就如何实施某些道德原则提供了一些指导,但是,他们遗漏了一些原则。

Qg （间隙）。人工智能和计算机科学的伦理研究存在哪些差距,特别是与 STS 的操作原则相关的差距?

这个问题有助于分析在STS 范围内实施原则以指导未来研究方面存在的差距。

2.2 来源选择和策略在确定了我们的目标和问题之后,

我们形成了通过识别关键词和资源来搜索原始研究的策略。我们选择 Google Scholar 和 University of Bristol Online Library 作为搜索资源。它们都是大型数据库,可以链接到各种其他研究来源以及关于该主题的已发表论文。我们使用所选关键字的各种组合搜索了所选资源,这些关键字可以在第 2.2.1 节中找到。

我们首先检查了每个资源中最多前 5 页的结果,然后通过对标题应用包含和排除标准来缩小搜索范围。这将搜索指定为一小部分摘要被阅读的作品。然后更严格地应用纳入和排除标准,从而确定主要研究。从最初搜索中收集的研究作品中,符合标准的相关引文被用来扩大搜索范围,从而可以从更广泛的来源收集材料。

图 1 简要概述了我们的方法。

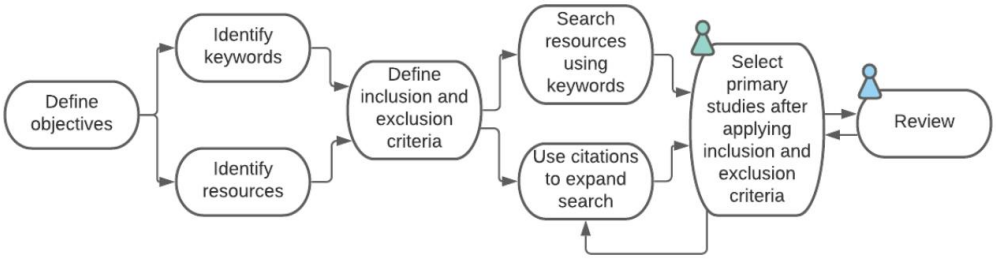

图 1. 简要方法

2.2.1 搜索字符串定义。我们的搜索字符串包含两个主要部分。第一个组成部分与 AI 和各种相关术语相关,而第二个组成部分与规范伦理相关。使用的搜索字符串是（ AI OR Agent OR ML OR Multiple-User OR Multiagent ）
AND（“结果主义”或“道义论”或“平等主义”或“平等”或“伦理学”或“功利主义”）。

2.2.2 纳入和排除标准。首先,作品来自一系列知名期刊和会议,这些期刊和会议是从最初搜索中找到的文献中确定的。特别包括这些资源可确保包括专题作品,但它也带来了可能遗漏不在列表中的资源的威胁。我们通过遵循主要研究的相关引用来扩大范围来降低风险,但承认时间限制仍然存在。我们排除了有关元伦理学（例如道德判断的意义)和计算机科学以外的应用伦理学（例如生物学伦理学)的著作。

其次,我们包括与个人或团体公平相关的作品。我们排除了有关特定 ML 方法论公平性的工作,因为这超出了本项目的范围。第三,我们包括与多用户社会困境相关的作品,以检查道德原则如何在这些环境中运作。我们排除了关于伦理原则如何影响其他非社会困境的研究。第四,我们包括规范伦理与多用户 AI 或 MAS 研究的交叉点,而我们排除了该领域的非伦理研究（例如关于技术实施） 。第五,我们包括了关于规范伦理原则和人工智能的研究,但我们排除了仅关于人工智能基石的研究。这是因为,虽然 AI keystones 包含有关道德实施的重要信息,但它不在本次审查的范围内。第六,我们包括关于与道德原则相关的偏见的研究,因为这与道德原则如何影响公平有关,但是我们排除了关于不谈论道德原则的偏见的研究。

2.3 相关著作

我们在 01-Jun-21 进行了初步搜索。该搜索在 Google学术搜索中产生了 374 万个结果,在布里斯托大学在线图书馆中产生了 998,613 个结果。查看结果的前 5 页,我们应用了纳入和排除标准,从每个资源中得出大约 10-20 项研究。对这些作品进行更仔细的检查可以识别出相关的引用,并将其纳入我们的评论中。这些作品的选择受到了二级研究人员的批评,这有助于确定进一步的相关研究。这导致 56 篇论文被纳入审查。我们在 22 年 5 月 23 日进行了第二次检索,结果又有 7篇论文被纳入审查范围。

表 1. 根据提取的原则对研究进行分类:框架

类型 / 话题	框架（概念化）	框架（应用程序）	算法	观点或评论
道义论	[1, 11, 13, 20, 43, 59, 76, 86, 98]	[6, 9, 10, 25, 46, 62, 64, 83]	[46, 84]	[44, 52, 96, 101]
平均主义	[9, 11, 19, 34, 38, 40, 59, 75, 82, 89]	[33]	–	[60]
比例主义	[35, 50, 59]	[33]	–	–
康德的	[1, 4, 45, 51, 54, 98]	[10, 62, 83, 95]	[84]	[58, 96]
美德	[1, 5, 13, 20, 43, 45, 46, 76, 86, 98]	[42, 46, 83]	[46, 84]	[44, 52, 96, 97, 101]
结果论	[1, 13, 20, 22, 43, 45, 59, 86, 92, 93]	[9, 10, 62]	[84]	[36, 96, 101]
功利主义	[1, 4, 5, 9, 13, 46 , 54, 59, 70, 75, 76, 98]	[2, 7, 10, 25, 62, 64, 83, 95]	[46, 84, 90]	[36, 52, 58, 101]
最大最小值	[59, 81]	[2, 10]	[28, 94]	[60]
无嫉妒	[12]	–	[94]	[60]
双重原则影响	–	[10, 41, 64, 71]	–	[26]
迪斯主义对等影响	[11]	–	[78]	–
不要伤害	[27]	[64]	–	–

3 伦理原则的分类

为了解决Qp （原则），人工智能和伦理原则的研究根据它们对规范伦理学的定义（道义论、平等主义、比例主义、康德主义、美德、结果主义、功利主义、最大化、无嫉妒、 Doctrine of Double Effect、Doctrine of Disparate Impact 和 Do No Harm），以及五种类型的研究（框架（概念化）、框架（应用）、算法和观点或评论），基于论文的结构和贡献。这显示在表 1 中。

就论文结构而言，我们发现绝大多数作品都集中在概念框架上，这些概念框架提出了关于如何实施这些原则的理论思想，例如 Leben [59]和 Wallach 等人。 [98]。一些论文在计算上应用了这样的框架，例如 Limarga 等人。 [62]和林德等人。 [64]。一些论文提出了将伦理原则机械化的算法，例如 Sun 等人。 [94]和戴安娜等人。 [28]。最后，我们发现也有针对相关领域的观点或评论论文，例如 Yu 等人。 [101] 和李等人。 [60]。

关于规范伦理原则，有两个主要的理论分支：道义论和目的论。道义论理论围绕规则、权利和义务展开[76, 98]。另一方面，目的论伦理学从作为要达到的目的的善或可取的东西中推导出责任或道德义务， [14]。目的论伦理学可进一步分为结果主义、利己主义和美德伦理学。图 2 以树形结构显示了文献中确定的原则分类，标出了它们之间的关系。

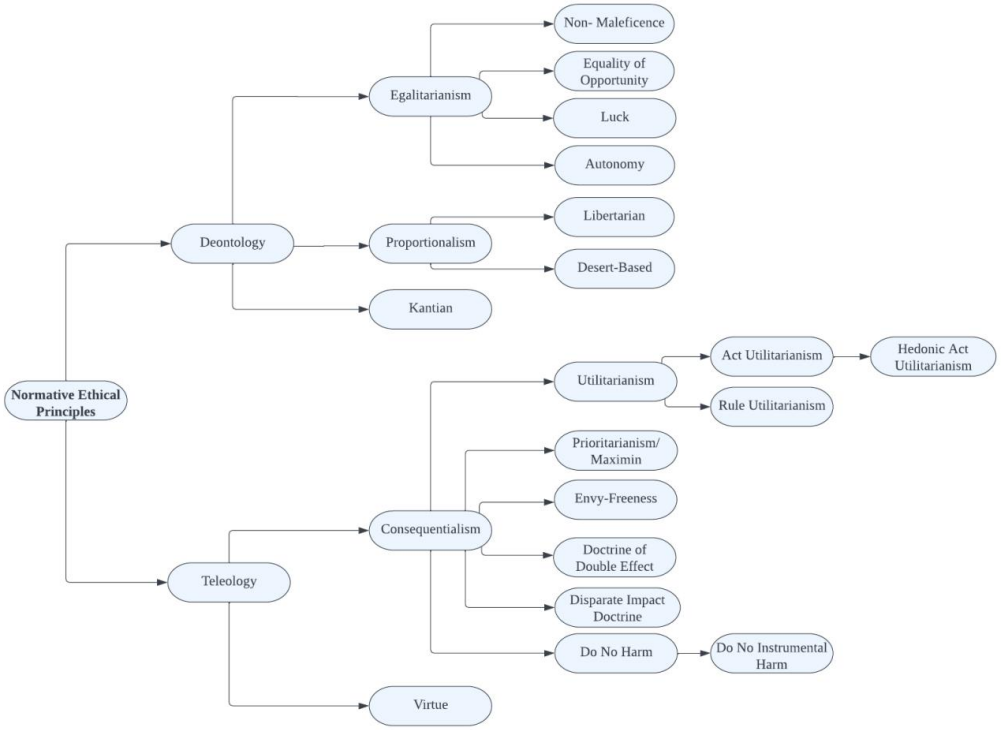

图 2. 伦理原则分类

我们发现,某些原则（例如功利主义)比其他原则（例如不伤害)被讨论得更多,如表 1 所示。我们还发现,有大量研究引用“道义论”和“后果论”作为广义术语,但没有具体说明他们指的是什么类型的道义论或结果论,例如 Cointe等人。 [20],格林等人。 [43],以及安德森和安德森[6]。这些工作可能会受益于更明确地说明他们正在使用的道德原则,以便进行更精确的操作。

3.1 道义论道义论需

要遵守规则、法律和规范[44、76] ,并尊重源于义务和权利的相关义务和许可[20] [4、52、83、84、86、96、98、101] .对于道义论理论,行为的可允许性在于行为本身的内在特征。当且仅当行为本身在道德上是善的且与结果无关时,该行为才是允许的[13, 62, 64]。为了实施道义论理论,可以使用基于规则的方法来确定适当的行动:Limarga 等。 [62]使用谓词对规则进行编码,然后对不同类型的动作进行推理； Berreby 等人。 [10]首先收集上下文信息以模拟行动的结果,然后使用道义规范评估该结果的伦理考虑； Tolmeijer 等人。 [96]认为道义论可以通过输入动作（在心理状态和后果方面） ,使用规则和职责作为决策标准,然后通过动作与规则的契合程度来机械化动作来实现。

道义论所应用的不同背景也值得考虑。
Binns [11]使用道义论在不兼容的公平性度量之间进行选择,而 Leben [59]将其应用于评估二元分类算法的分布。有些人还建议仅在特定情况下使用道义论:Dehghani 等。 [25]选择在具有“神圣价值”的情况下实施道义论,用它来选择不违反神圣价值的行动选择。

但是,在应用道义论时可能会出现一些问题。一个普遍的担忧是,由于道义论方法侧重于行为的内在本质,因此它们未能考虑到最可能的后果。这使得逻辑难以充分捕捉复杂的伦理见解[1, 86]。此外,基于权利的伦理围绕基于受决策影响的人的权利的决策,但在权利未受到侵犯但仍存在某种道德困境的情况下,这可能不太有用。在实施方面,当规则出现例外情况时,可能会出现问题。规则应该被严格遵守,这意味着对于每一个例外,它们都必须被修改,这可能会使它们相当长。因此,确定正确的细节级别对于确保机器的可解释性非常重要[96]。最后,规则之间可能存在冲突。可以通过对规则进行排序或权衡来解决冲突,但这会导致确定重要性顺序的困难。

3.1.1 平均主义。平等主义源于这样一种观念,即人类在某种基本意义上是平等的;因此,应努力避免和纠正某些形式的不平等[11]。
文学通过以不同方式促进平等来实现平等主义:Murukannaiah 等。 [75]建议尽量减少利益相关者之间在满足其偏好方面的差异;德沃克等人。 [33]对在特定属性方面相似的个体进行相似分类; Leben [59]赋予人口中的每个成员平等的权利(因此平等的份额)。然而,如果不可能在整个人口的所有指标上实现平等,他们建议采用一种分布,以尽量减少与某些公平标准(例如人口规模)的距离。 Lee 等人提出的平等主义的应用。 [60]是评估各种算法公平性指标,例如阳性预测奇偶校验或相等赔率。这可能有助于模型开发人员决定哪些层次的不平等应该(不)影响模型的预测。

在考虑平等主义时,承认某些困难很重要。例如,关于是否应该在不同的社会背景下应用单一的平等主义演算,或者是否存在内部“正义领域”,不同的公平指标可能适用,在这些领域之间重新分配可能不合适,存在着一个突出的争论[11].平等主义的特定措施可能适用于不同的情况,例如,在被允许投票参加政治选举之前普遍实施测试可能会导致社会经济背景较低的人被排除在民主之外。然而,如果每个人都有平等的机会参加考试,那么对工作职位进行考试似乎是合适的,因为个人的才能和能力各不相同。因此,人们应该仔细评估用于实施平等主义的指标。

平均主义的亚型见表 2。

3.1.2 比例主义。比例主义推断根据每个人对生产的贡献按比例调整每个人的权利。贡献可能包括投入生产的每个人口成员的资源、部署这些资源的实际工作量,以及投入这些资源的运气。 Leben [59]通过构建根据贡献评估权利分配的效用函数来实施比例主义。公平标准通过将贡献总量除以每个人的贡献量来建立理想的权利分配。

表 2. 平均主义的亚型

原则	描述	困难
非恶意	在损害而非利益上强加平等主义 [59]。	允许结果出现任意大的不平等,并假设“富裕”和“贫困”之间存在可疑的区别[59]。因此很难定义什么是伤害以及什么是伤害 一个好处是。
等于机会_	由于个人的出生情况或随机选择而导致的负面属性应该 不反对他们,但个人仍应为自己负责 行动[33, 40]。可以检查每个组是否同样有可能在给定该组的基本比率的情况下被预测为理想的结果[11],或者根据相关的优点定义确保所有机会应平等地向所有申请人开放[60]。	从理论上讲,即使只有一小部分人口具有获得机会的现实前景,也可以完全满足这一要求 [38]。
运气	消除未经选择的不平等;没有人应该因为运气不好而最终变得更糟,相反,人们应该根据自己的选择获得差异化的经济利益 [34, 60]。	通常很难定义个人真正控制范围内的内容[11]。理想的解决方案是允许因人们的自由选择和知情风险承担而导致的不平等,而忽略那些完全靠运气造成的不平等。
自治自治平等已被提议为包括全方位的个人自由[60]。必须有最低水平的自主权、选择的多样性和质量以及决策能力 [38]。		然而,当权力和信息严重不对称时,理性决策者的自主性就无法作为一种伦理 目标 [38]。

因此,最好的分配是与所有个人的公平标准的距离最小的分配。

比例主义的一个挑战是,在某些情况下,团体或个人可能没有赋予生产贡献,但仍应被授予权利分配。例如,一个因残疾而无法做出贡献的群体仍应享有公平的权利分配。然而,考虑到运气的影响,这可能会有所缓解。对于比例主义的子类型,请参见表 3。

3.1.3康德。康德[51]伦理学认为,伦理原则源自行动的逻辑结构,首先将自由行动(行为人有理由的行为)与单纯的行为区分开来[54]。康德的绝对命令以所有道德义务为基础[98],因为它无条件地适用于理性主体(绝对),并且是可以遵循但可能不是(命令)[49] 的命令。因此,绝对命令意味着理性主体必须相信他们行动的理由与假设相一致,即适用这些理由的所有理性主体都可以从事相同的行动(也称为自然的普遍法则)[1, 58, 83]。例如,“不要杀人”是一个绝对命令:它是绝对的,因为如果所有理性代理人都被杀死,就不会有理性代理人留下;这是一个命令,因为理性代理人有能力杀人但不应该杀人。派生自绝对命令是手段-目的

表 3. 比例主义的子类型

原则	描述	困难
自由主义自由主义	强调每个人的自由的重要性,只要不伤害其他任何人[60]。权利根据每个人在同意时的总贡献进行相应分配。每个组有权获得至少与初始贡献一样公平的成功率 [59]。	这种方法的一个困难在于它没有针对可能仍然值得缓解的先前存在的不平等现象。例如,由于代际财富不平等等他们无法控制的因素,某些人可能无法做出与其他人一样多的贡献。另一方面,生来富贵的人,按照这种做法,会得到更多的权利,这似乎是不公平的。
沙漠 基于	贡献是根据个人来定义的 努力,运气的影响,权利是相应分配的。这是因为人口中某种特征的先前流行可能是不公正环境的结果[59]。这可以通过在评估沙漠的度量空间中为每个个体分配一些距离,然后通过度量空间中每个组的个体之间的平均距离来评估模型的公平性来实现[33]。	这个原则的一个弱点是运气是一个抽象的概念,很难定义,并且可能因上下文而异。因此,这使得评估应该减轻哪些特征具有挑战性。

原则（也称为人性公式）。这表示将他人视为达到目的的手段是不道德的[1, 58]。永远不可能将对待他人作为达到某种目的的手段普遍化;这样做会与绝对命令相矛盾。这是因为我们有能力进行理性的自我导向行为。

通过实施规则,该原则已在以前的文献中得到实施。
利马加等人。 [62]通过强加两条规则来实施绝对命令:首先,由于它是普遍的,代理人在采用遵循的原则（或判断行为是其职责)时,必须模拟一个每个人都遵守的世界该原则并考虑该世界的理想。其次,由于行为在本质上在道德上是允许的、禁止的或强制的,因此代理人必须纯粹因为这是一个人的职责而履行其职责,而不是作为达到目的的手段或通过雇用另一个人作为达到目的的手段。 Berreby 等人。 [10]实施手段-目的原则的规则是,如果一项行动涉及并影响至少一个人,但这种影响不是行动的目的,则该行动是不允许的。 Svegliato 等人。 [95]将道德原则（以规则的形式)与决策模块分离;对于康德主义,他们使用道德规则,即政策应该无矛盾地普遍适用于利益相关者。艾伦等人。 [4]建议可以将绝对命令作为评估其他规则的更高原则来实施。

例如,在决定是否应用平等主义（确保平等分配)时,代理人可以通过检查它是否符合绝对命令来评估这是否是正确的事情;如果所有代理人都适用该原则是合理的。

绝对命令的一个困难是它可能过于宽容;它可以通过允许任何可以具有普遍性准则的行动来允许直觉上的坏事[1]。一个常见的例子是让谋杀进入你的房子,因为你不能撒谎,并说那个人

他们要杀的是不存在的。手段-目的原则也可能过于严格,正如严格解释的那样,它禁止一个人在未经他们明确同意的情况下影响另一个人的任何行为。

3.2 目的论伦理原则本节检查审查中确定的每

项目的论原则(道德源于要达到的目的的善或可取[14]),包括有关这些原则之前如何实施的详细信息以及可能出现的困难.

3.2.1 美德伦理。根据美德伦理学,道德源于个人的内在性格,而不是个人行为的正确或错误[1, 13, 52, 76, 83, 98, 101]。正确的行为是由具有美德品格的人执行的。按照这一理论,人们不应该问自己应该做什么,而应该问自己应该成为什么样的人[1, 5, 84, 86]。一个人拥有的品质应该是第一重要的,其次是行动。道德美德可以通过习惯和实践来学习和发展。美德的稳定性(如果一个人有美德,就不能表现得好像没有美德一样)意味着美德伦理可能是一种将道德灌输给机器的有用方法 [98]。

美德伦理在文学中有多种不同的实施方式。
罗宾斯和华莱士[83]认为,要实施这一原则,可以通过应用“良性”特征以道德方式解决问题。正如 Vanhé和 Borit [97]所建议的那样,这可以通过以系统设计者为目标并通过教育帮助设计者培养美德来实现。其他工作侧重于将美德直接应用到机器中;根据Tolmeijer 等人的说法。 [96],在机器中实施美德伦理的输入将是代理的属性,决策标准将基于美德,这将通过美德的实例化来机械化。 Govindarajulu 等人举例说明了这一点。 [41]谁定义美德是通过在观察有德行的人时体验钦佩的情绪,然后复制那些人的特征来学习的。作者通过使用计算形式逻辑将情绪(特别是钦佩的情绪)形式化、表示(美德)特征并建立学习特征的过程来实例化这一点。格林等人。 [43]认为,基于美德的系统必须在给定情况下理解需要采取一种行动而不是另一种行动的所有特征。美德伦理也可以与其他方法一起使用; Hagendorff [44]认为道义论方法应该通过观察价值观和性格倾向与美德伦理学相结合。

美德伦理学的一个问题是它采用的整体观点使其更难应用于个别情况[86]。进一步的挑战涉及相互冲突的美德和美德的具体化[96]。仅仅通过观察一个行为或一系列似乎暗示美德的行为是不可能判断机器或人是否有美德的 需要弄清楚背后的原因。因此,这使得很难将美德构建到机器中,因为对于美德的实际含义存在高度抽象。此外,美德的概念会随着时间和文化的不同而发生巨大变化。随着美德的改变,现在在机器中实例化的美德可能会在未来导致不公平的结果。

3.2.2 结果论。在后果论方法中,正确的行为是通过其影响来确定的[13, 20, 59, 101]。因此,只能通过考虑其后果来判断一个行为的道德有效性[62, 84, 86]。这样做的一个优势是,它可以通过检查这些利益和危害的分配方式,用于评估具有复杂结果的决策,其中一些利益和一些利益受到损害。因此,它可以解释许多困扰道义论理论的道德直觉,因为后果论者可以说,最好的结果是收益大于成本的结果 [92]。

因此,结果主义原则可以通过分析不同行动的后果来实施。这与道义论不同,道义论认为“心理状态”对于确定行为的道德性非常重要,但对于结果主义可以在很大程度上被忽视， Tolmeijer 等人。 [96]争论。后果论由 Limarga 等人实施。 [62],谁根据每个动作的最坏后果分配一个权重。动作是实现目标的一系列动作的一部分,它们的权重累积为一个总值。然后优化该值以选择具有最佳整体结果的序列。 Suikkanen [93] 同样建议根据代理人的选择结果的总价值对代理人的选择进行排序。一个选项是正确的当且仅当没有其他选项具有更高的评价等级。 Tolmeijer 等人。 [96]认为结果主义原则的输入将是行动（及其后果）,而决策标准将是相对幸福感。这将通过选择具有最大效用的结果来机械化。对于二元分类算法,Leben [59]建议通过查看如何根据相对社会成本将权重分配给每个组结果来实施结果论。

然而,在实践中,为每个结果分配权重对于所有结果来说可能是不现实的[59]。可能会有很高的计算成本,因为后果论系统需要一台机器来表示它可用的所有动作[43]。解决的一个相关问题在于难以估计长期或不确定的后果以及确定应为谁考虑后果[36, 86]。结果主义之外也可能存在道德约束,即使某些行为有最好的结果,也会禁止某些行为,因此导致结果主义理论不完整[93]。对结果主义的另一种常见批评涉及决定什么是有价值的或本质上好的:无论是快乐、偏好满足、一个人基本能力的完善,还是一些完全不同的客观商品（例如知识、美等)[13、96] .对于结果主义的子类型,请参见表 4。

表 4. 结果论的子类型。

原则 描述功利主义当且仅当它使所		困难这可能导
有受影响的人的总净预期效用最大化时,它才是合乎道德的[52、54、58、64、76、84、98] 。	最终的目的是尽可能远离痛苦并享受尽可能丰富的生活[4, 70]。因此,可以训练代理人做出能够为最大数量的人带来最大幸福的判断[58],例如,通过为用于最终评估的每个动作分配一个值[62];选择效用最高的选项[9, 25]。	致少数人为了更大的利益而受到不公平的对待[2, 7]。计算每一个行动的效用可能是不可能的,而且该理论无法解释权利和义务的概念[1]。也很难量化效用[36]。为了缓解这些问题,功利主义可能是道德行为的额外必要条件,而不是唯一的道德原则 [54]。
	另一种方法将道德原则与决策模块分离,有一个单独的道德规则来评估建议的政策[95]。要根据与道德价值观的一致程度来选择规范,递归效用函数可以识别每个价值观的偏好效用。然后通过将支持规范的每个值的效用相加来计算规范的价值支持。这可可以说是通过确保选择促进最大效用的规范来强化功利主义,其中效用被理解为（价值)偏好满足[90]。	
	为了证明二元分类算法中公平性指标的设计选择是合理的,一个函数可以对每个潜在分布及其影响（效用函数/幸福结果的度量)进行建模,然后对聚合效用运行选择过程以最大化总和 [59].	
（享乐主义）实用主义主义	通过其后果关注行为的道德[10, 96]； Hedonic Act Utili tarianism 需要计算获得最大净快乐的最佳行动[13]。利用这一点的机器可以权衡与其后果相对应的行动,然后相应地对它们进行排序;如果存在另一个权重更大的动作,则该动作不太可取[10]。或者,可以输入受影响的人数,以及每个人对每个可能的行为的快乐/不快乐的强度。该算法然后计算强度、持续时间和概率的乘积,以获得每个人的净快乐。为每个备选动作执行此计算[7]。	对享乐行为功利主义的批评是 很难定义快乐;对一个人来说愉快的事情对另一个人来说可能并不愉快。这使得很难 确定具有最大净请求的行动 当然。

原则	描述通过首先根	困难
规则实用程序 塔里安主义	据效用原则评估道德规则来对行为进行道德评估 决定是否（一套）道德规则会导致最好的总体结果,因为假设所有/大多数代理人都遵循它。这可以使用谓词来实现,该谓词包含属于特定规则的操作的所有有效权重,然后通过谓词 [10] 对这些权重求和。	有时一条规则可能会导致不直观的结果出现,因此应该被打破。这使得规则功利主义看起来更像是行为功利主义,在这种情况下,正确的做法是通过每个行为的后果来评估的 行动。
优先的 主义/马克西姆	通过寻求改善社会中最坏情况下的体验来最大化最小效用[60, 81],例如旨在改善任何用户的最小体验/最坏情况结果的代理[2]。对于二元分类算法的公平性指标的选择,可以构建一个对每个潜在分布及其影响进行建模的函数,然后在聚合效用上运行一个选择程序[59]。公平性可以通过所有组的最坏情况结果来衡量,而不是组结果之间的差异[28],或者通过最小化所有分配中分配的最大成本[94]。	虽然总效用可能会增加,但并不一定会减轻歧视的影响[94]。它仍然允许群体之间存在差异。因此,尽管整体体验得到改善,但特权最高的群体可能仍比特权最低的群体享有更多特权。
嫉妒 游离度	在 Envy-Free 分配中,没有代理人嫉妒另一个代理人[94]。因此,当群体或个人之间的嫉妒水平最低时,公平就存在了。资源可能分配不均,但只要代理不互相嫉妒,这就被认为是公平的[12]。	可以说,重要的可能不是相对于其他人的条件,而是人们是否有足够的能力拥有令人满意的生活前景[60]。此外,当项目不可分割时,无法保证存在 Envy Free 分配,例如需要分配给多个代理的家务[94]。
学说 双重效果	故意造成伤害是错误的,即使它会带来好处[26, 71]。如果一个行为本身在道德上是好的/中性的,一个行为是允许的,一些积极的结果是有意的,没有消极的结果是达到目标的手段,并且积极的结果足以超过消极的结果 [10, 41, 64]。	双重效应原则的一个问题是,只要不是故意的,它仍然允许不良行为发生,这可能会产生一些道德上可疑的结果。
不同的 影响 教义	为了群体公平,任何群体在算法 [78] 提供的解决方案中都必须具有近似相等或成比例的代表性。还提出了“不同的虐待”的概念,它考虑了组间假阳性率的差异[11]。因此,它强调了确保影响在相关群体之间按比例分配的重要性。	这可能会导致个人受到有利于群体的不公平对待。
不要 (伴奏) 伤害	不应造成任何伤害,因此任何造成伤害的行为都是不道德的[27, 64]。 Do No Instrumental Harm 允许将伤害作为副作用,而不是作为实现目标的手段。	然而,有时在某些情况下造成伤害是不可避免的。在这种情况下,单凭这一原则将无法提供明确的伦理指导。

3.3 其他原则除了此处列出
的原则外,还有文献中提到的其他原则。
由于各种原因,这些未包括在分类法中,如此处将解释的那样。

3.3.1 利己主义。利己主义是在不考虑他人的情况下为自己达到最大的可能结果[58, 83]。这一原则在文献中很少提及,这可能是因为如果将其灌输到 AI 代理中,可能会导致不道德的结果。如果代理人主要关心自己而不考虑其他人,那么公平似乎不太可能成为他们的道德目标。这是因为公平的目的是为了他人和自己的福祉,而利己主义则完全以自我为中心。

3.3.2 特殊主义。特殊主义强调规范价值没有唯一的来源,也没有单一的、普遍适用的道德评估程序[96]。规则或先例可以指导评估实践,但它们被认为过于粗糙,无法公正对待许多个别情况。因此,某个特征是否具有道德相关性以及它发挥什么作用将对情况的其他特征敏感,因此应根据具体情况进行伦理评估。特殊主义的输入可能包括情况(背景、特征、意图和后果),决策标准基于经验法则和先例,因为所有情况都是独一无二的。决定一项行动的机制将取决于它在多大程度上符合规则或先例。确定的一些挑战是没有独特和普遍的逻辑,因此每种情况都需要进行独特的评估。缺乏通用逻辑是不将其包含在该分类法中的部分原因:它没有给出明确的指导。然而,特殊主义可能与伦理原则的实施方式有关,因为它强调将情境纳入道德推理过程。它本身并不是一个可操作的伦理原则,但也许更像是一个元原则,可以用于其他伦理原则的应用。

3.3.3 关怀与责任伦理。关怀和责任的伦理与考虑你与他人相互联系的感受有关[68, 83]。为了符合道德规范,人们应该考虑其他人和您所处的情况。根据您的经验,您应该以一种有教养和负责任的方式行事。这是应用道德原则的关键指导因素,因为它增强了考虑自己以外的其他人的重要性。这为实现公平的目标提供了很好的支持,但它本身并不是一个原则,因为它表示某个动作,这就是为什么它没有被包含在分类法中。

3.3.4 其他文化。最后,在西方伦理史之外的文化中提出了各种各样的原则。世界各地的社会都建立了道德框架,包括儒家、神道教和印度教思想,以及犹太教、基督教和伊斯兰教等宗教框架[45]。跨文化存在多种道德框架,在这些框架内存在显着差异。可以说,道德和文化是密不可分的,要理解一个,你必须看看另一个。因此,必须在其文化背景下考虑伦理。这些原则未包含在分类法中的原因不是因为它们不重要,而是因为它们需要自己的完整分类法。未来工作的一个重要方向是将本项目中使用的方法专门应用于非西方伦理原则,目标是形成此类原则的分类法。这对于帮助开发人员构建跨文化伦理技术至关重要。

4 先前的道德原则实施情况

我们对审查中确定的论文进行了迭代,以对之前的Qo （操作化)伦理原则操作进行分析。我们发现了多种用于道德原则的技术实施的技术,总结在表 5、表 6 和表 7 中。我们发现以前的文献以自上而下、自下而上或混合架构将原则整合到推理能力中,总结在表 8 中。我们发现从业者应该具体说明他们正在实施的原则,之前的文献表明多元主义可能有助于做出这一决定。我们还发现,抽象地,操作化属于为义务论原则应用规则、为美德伦理发展美德或为结果主义原则评估后果的类别。

4.1 选择技术实施我们发现在以前的文献中,各种不同

的技术实施已被用于编码伦理原则。扩展 Tolmeijer 等人解释的分类。 [96],将原理编码为计算机可以理解的格式的方法包括逻辑推理、概率推理、学习、优化和基于案例的推理[85]。这在表 5、表 6 和表 7 中进行了总结。

表 5. 编码原则的技术实现概述,改编自 [96]

实施类型		描述	例子
	演绎的逻辑	知识被表示为命题和规则,从中可以推导出新的命题	康德伦理学、美德伦理学、功利主义 [83]
合乎逻辑推理	非单调的逻辑	允许修改结论 发生冲突时的解决方案（例如来自新信息）	康德伦理学、Maximin、Act Utilitarianism、Rule Utilitarianism、Doctrine of Double Effect [10]
	外展逻辑	给定前提得出最可能的命题	康德伦理学,功利主义 [62] 双重效应学说 [71]
	道义逻辑用于表达与义务、许可、禁止和可选性相关的规范性命题		双重效应学说 [41] 美德伦理 [42] 道义论、功利主义、双重效应论、不（在工具上)伤害 [64]
	基于规则的系统项目	系统必须遵守一组规则（上面提到的许多逻辑类型通常作为基于规则的系统实现）	马克西姆 [2] 非恶意、自治、功利主义 [9] 道义论、美德伦理学、结果论 [20] 道义论、功利主义 [25] 不要伤害 [27] 自由主义 [35] 康德伦理学、美德伦理学、功利主义 [83]

实施类型		描述	例子
	事件演算不同事件触发不同	行为类型	康德伦理学,行为功利主义，规则功利主义、Maximin、双重效应学说 [10] 双重效应学说 [41] 美德伦理 [42] 康德伦理学,功利主义 [62]
	知识表示和本体	关注系统如何通过提高数据质量来利用知识（而不是单独使用算法）	道义论、美德伦理学、结果论 [20] 道义论、功利主义 [25] 康德伦理学、美德伦理学、功利主义 [83]
	归纳逻辑前提是从示例中归纳或学习的（而不是预先定义的）		道义论 [6]
概率的推理	贝叶斯方法	使用贝叶斯规则通过使用先验知识计算事件的可能性	功利主义 [8]
	马尔可夫模型假设未来事件仅取决于当前（而非之前)事件		双重效应学说 [41] 道义论,康德,美德伦理学，结果主义伦理学 [84]
	统计推断	从数据中的概率分布预测未来事件发生的可能性	康德伦理学,功利主义 [95] 机会均等,比例主义 [33]
学习	决策树	通过将决策空间探索为搜索树并计算预期效用来解决分类问题的监督学习方法	非恶意、自治、功利主义 [9]
	加强学习	对行为的奖励或惩罚教会系统学习什么	马克西姆 [2] 道义论、康德伦理学、美德伦理学,后果论 [84]
	逆加固学习	系统通过观察行为学习奖励函数	道义论 [77]
	神经网络作品	经过案例训练后,可以根据相关特征对新案例进行分类	道义论、美德伦理学、享乐主义行为功利主义 [46] 美德伦理 [47]
	进化的计算	模型以迭代方式发展;当有不同的com时使用 道德机器的宠物模型	美德伦理 [47]

实施类型	描述不同的	例子
优化	动作根据预先确定的公式分配不同的值,并选择最佳值（例如最高值）	非恶意行为,美德伦理， 功利主义，（享乐)法 功利主义 [7] 功利主义 [8] 非恶意、自治、功利主义 [9] 马克西姆 [28] 机会均等,比例主义 [33] 平均主义、比例主义、 功利主义,Maximin [59] 马克西姆 [78] 平均主义[90] 无嫉妒 [94]
基于案例推理	根据先前案例的集合评估新情况	道义论、功利主义 [25] 道义论 [69]

4.2 澄清架构为了设计道德敏感的系统,从

业者必须决定整合道德原则的架构[98]。这些属于三种广泛的方法:自上而下的伦理理论强加;自下而上的系统构建,其目标可能会明确指定,也可能不会明确指定;一种结合了自上而下和自下而上特征的混合方法。表 8 总结了我们关于如何根据不同架构实施道德原则的发现。

4.2.1 自下而上的方法。自下而上的方法被理解为机器通过观察实际情况下的人类行为来学习做出道德决策,而无需教授任何正式规则或道德哲学[36]。 Tolmeijer 等人建议的自下而上技术。 [96]包括使用人工神经网络、强化学习和进化计算。

Noothigattu 等人就是一个例子。 [77],他们使用逆向强化学习通过从观察到的行为中学习策略来使代理与人类价值观保持一致。在未来的工作中,逆向强化学习可用于使政策与道德原则保持一致,其方式与Noothigattu 等人的方式类似。 [77]使政策与人类价值观保持一致。这可以通过将政策与原则同化来提高可解释性,这些原则本质上暗示可以推理的逻辑命题[54]。然而,自下而上方法的挑战在于机器学习错误规则或无法可靠地推断出未反映在训练数据中的情况的风险。

4.2.2 自上而下的方法。自上而下的方法将道德直接植入机器[54],而不是要求机器从经验中学习（如自下而上的方法）。自上而下的方法是基于规则的:道德被理解为通过识别应该遵循的规则来调查正确的行为,以便执行道德上正确的（或至少是允许的)行为[63]。我们发现许多作品使用自上而下的方法将道德原则整合到机器的推理能力中。 Dehghani 等人。 [25]通过结合定性建模、第一性原理逻辑推理和类比推理来实施道义论和功利主义原则。戴安娜等人。 [28]使用 oracle-efficient 学习实施 minimax 原则（最小化最大损失,改编自 Maximin - 最大化最小值）

表 6. 原则的技术实现

实施类型		道德原则				
		道义论平均主义	比例主义	康德主义		美德
合乎逻辑推理	演绎的逻辑	–	–	–	[83]	[83]
	非单调的逻辑	–	–	–	[10, 62]	–
	外展逻辑	–	–	–	–	–
	道义的逻辑	[64]	–	–	–	[42]
	基于规则的系统	[20, 25]	[9]	[35]	[83]	[20, 83]
	事件计算	–	–	–	[10, 62]	[42]
	卢斯知识代表和本体论	[20, 25]	–	–	[83]	[20, 83]
	感应式逻辑	[6]	–	–	–	–
概率的推理	贝叶斯方法	–	–	–	–	–
	马尔可夫模型	[84]	–	–	[84, 95]	[84]
	统计推断	–	[33]	[33]	–	–
学习	决定树	–	[9]	–	–	–
	加强学习	[84]	–	–	[84]	[84]
	逆向重新执法	[77]	–	–	–	–
	学习神经网络作品	[46]	–	–	–	[46, 47]
	进化的计算	–	–	–	–	[47]
优化		–	[7, 9, 33, 59, 90]	[33, 59]	–	[7]
基于案例推理		[25, 69]	–	–	–	–

表 7. 原则的技术实现

实施类型		道德原则						
		结果主义	功用 万物主义	极小嫉妒	游离度	教义 斗的 祝福 影响	教义 的迪斯 帕拉特 影响	不做 伤害
合乎逻辑 推理	演绎的 逻辑	–	[83]	–	–	–	–	
	非 单调的 逻辑	–	[10, 62]	[10]	–	[10]	–	–
	外展 逻辑	–	–	–	–	[71]	–	–
	道义的 逻辑	–	[64]	–	–	[41, 64] –		[64]
	基于规则 系统	[20]	[9, 25, 83] [2]		–	–	–	[27]
	事件计算 卢斯	–	[10, 62]	[10]	–	[10, 42] –		–
	知识 代表和	[20]	[25, 83]	–	–	–	–	–
	本体论 感应式 逻辑	–	–	–	–	–	–	–
概率的 推理	贝叶斯 方法	–	[8]	–	–	–	–	–
	马尔可夫 楷模	[84]	[95]	–	–	[41]	–	–
	统计输入 参考	–	–	–	–	–	–	–
学习	决定 树	–	[9]	–	–	–	–	–
	加固[84] 学习		–	[2]	–	–	–	–
	逆向重新 执法	–	–	–	–	–	–	–
	学习 神经网络 作品	–	[46]	–	–	–	–	–
	进化的 计算	–	–	–	–	–	–	–
优化		–	[7–9, 59] [28, 59, 78]		[94]	–	[11, 78] –	
案件 基于 推理		–	[25]	–	–	–	–	–

算法。他们应用 minimax 来分析群体结果之间差异的公平性考虑。同样考虑到公平性,Sun 等人。 [94]将 Envy-Freeness 形式化为检查不同公平分配之间权衡的规则。 Tolmeijer 等人。 [96]发现原则可以通过逻辑或基于案例的推理作为规则实施,使用领域知识来推理作为输入给出的情况。然而,由于人类知识往往不是非常结构化的,因此在使用之前需要对其进行解释。因此,自上而下方法的一个困难在于,人类对哲学规则的理解需要以机器可以理解的方式进行编码,这可能意味着信息丢失或被歪曲。

4.2.3 混合方法。混合方法体现了自上而下和自下而上方法的各个方面。由于自上而下和自下而上的方法各自体现了道德敏感性的不同方面,因此将两者结合在混合方法中可能会更好地实施道德原则[4]。混合示例包括 Berreby 等人。 [10],他们通过自下而上的上下文信息观察来补充自上而下的规则实施,允许代理人代表和推理各种道义论和后果论理论。他们提出了一个基于事件演算修改版本的模块化逻辑框架,并在答案集编程中实现。利马加等人。 [62]在事件集演算中使用非单调推理实施原则,这允许在发生冲突时修改规则。罗德里格斯索托等人。 [84]提出了一种方法,首先将道德行为描述为道德奖励,然后使用多目标强化学习将这种奖励嵌入到代理的学习环境中。他们的算法构建了一个环境,在这个环境中,代理人在追求其个人目标的同时以合乎道德的方式行事符合最大利益。为了考虑多个目标,他们将环境建模为多目标马尔可夫决策过程,允许代理人同时考虑个人和道德目标。按照自上而下的方法,道德原则沿着规范(行为是好还是坏)和评估维度(它有多好)形式化。然后以自下而上的方式将原则用作奖励函数。混合方法的一个好处是它们结合了道德推理和经验观察,从而可以考虑到背景。

表 8. 实施原则的架构

自下而上的道德原则		自顶向下	杂交种
道义论	逆 加强 学习 [77] 归纳逻辑 [6]	基于规则的系统、知识表示和 本体,基于案例 推理 [25] 道义逻辑 [64] 基于案例的推理 [69]	神经网络 [46] 规则库系统、知识表示和 本体 [20] 马尔可夫模型,强化学习 [84]
平均主义	–	统计推断,优化[33] 优化 [7, 59, 90]	基于规则的系统、决策树、优化 [9]
比例主义 –		统计推断,优化[33] 基于规则的系统 [35] 优化 [59]	–

自下而上的道德原则		自顶向下	杂交种
康德的	–	演绎逻辑,规则 基于系统,知识 表示和本体论 [83] 马尔可夫模型 [95]	非单调推理和事件演算[10, 62] 马尔可夫模型,强化学习 [84]
美德	进化计算[47]	演绎逻辑,规则 基于系统,知识 表示和本体论 [83]	神经网络 [46] 道义逻辑,事件演算[42] 规则库系统、知识表示和 本体 [20] 马尔可夫模型,强化学习 [84]
结果主义		–	规则库系统、知识表示和 本体 [20] 马尔可夫模型,强化学习 [84]
功利主义	基于规则的系统、知识表示和 本体 [25]	演绎逻辑,规则 基于系统,知识 表示和本体论 [83] 道义逻辑 [64] 优化 [7, 59]	非单调推理和事件演算[10, 62] 神经网络 [46] 基于规则的系统、决策树、优化 [9] 贝叶斯方法,优化[8]
最大最小值	–	优化 [28, 59, 78] 基于规则的系统,强化学习 [2]	非单调推理与事件演算 [10]
无嫉妒	–	优化 [94]	
斗学 蓝色效果	–	归纳逻辑 [71] 道义逻辑 [64] 道义逻辑、事件演算、马尔可夫模型 [41]	非单调推理与事件演算 [10]
迪斯主义 对等影响	–	优化 [11, 78]	–
不要伤害	–	道义逻辑 [64] 基于规则的系统 [27]	–

4.3 明确伦理原则从业者应明确将实施哪些伦理

原则。我们建议的分类法可以帮助实现这一点,该分类法包含AI 和计算机科学文献中的广泛伦理原则,有 25 个节点（图 2）。清楚正在使用的原则将有助于设计人员进一步指定他们的应用程序需要哪些输入,这反过来将提高道德推理能力和决策制定的可解释性 [59]。

在具体说明伦理原则时,应用多元主义理论可能会有用。多元主义指出,没有一种方法是最好的[83],因为人类道德是复杂的,不能被单一的经典伦理理论所涵盖。上下文和各种推理技术可用于在适当的原则之间进行选择。Tolmeijer 等人。 [96] 提倡根据这种方法进行进一步的研究,建议开发多理论模型,机器可以根据情况的类型互换应用不同的理论。在 Svegliato 等人的著作中可以找到多元主义的一个例子。 [95],他们提出了一个框架,在该框架中,道德合规与任务完成脱钩,以避免无法反映利益相关者价值观的意外情况。他们建议,这可以通过以具有代表道德原则的额外道德约束的形式实施多元主义方法来实现。这允许考虑其道德背景来评估决策模块的政策的道德性,从而为不同的道德原则作为道德规则来实施留出空间。多元主义是一种适用于伦理的有用方法,并且可能有助于帮助开发人员决定哪些伦理原则是合适的。

4.4 使用规则、后果或美德我们发现以前的工作以三种主

要方式实现了原则,具体取决于所使用原则的类型。通过规则的应用,然后根据特定规则的符合程度来选择行动,道义论原则已经得到实施。

美德伦理通过美德特征的发展而得以实施。结果主义原则已经通过评估后果然后根据其产生的后果选择行动来实施。

4.4.1 应用规则。对于义务论原则,一些方法建议通过对可能的行动应用一组规则来确定哪些行动是令人满意的来实施原则,例如 Abney [1]、Greene 等人。 [43],和 Berreby 等人。 [10].这方面的例子是应用从平等主义原则中提取的利益相关者偏好满意度差异应最小化的规则[75]。另一个例子是应用利益相关者应根据他们对生产的贡献按比例对待的规则[59]。然而,由于伦理的抽象性,很难找到合适的方法将伦理原则编码为具体规则[96]。这些困难的一方面可能在于决定规则是否应该被解释为严格的或可废除的。例如,康德主义[51]的一个重要组成部分是行为的原因必须对所有代理人都是普遍的,因此也许这条规则应该是严格的。但是,可以说这可能会允许根据其他原则[1] 采取不良行为,这表明它应该是可废除的。因此,创建将道德原则编码为规则的系统方法,包括理解规则应该是严格的还是可废除的,以便在 STS 的背景下使用可能是未来研究的方向。

4.4.2 培养美德。对于美德伦理,伦理源于个人的内在性格[1, 13, 52, 76, 98, 101]。根据这一理论解决问题,应运用良性特征[83]。因此,该理论可以通过美德的实例化来实施

[96]。 Govindarajulu 等人举例说明了这一点。 [41],他们将美德理解为通过观察有德行的人时体验钦佩的情感而学习,然后复制这些人的特征。这是通过使用计算形式逻辑来形式化情绪（特别是钦佩的情绪）、表示特征（在本例中是良性的)并建立学习特征的过程来实例化的。为了形式化美德,作者使用道义认知事件演算,这是一种计算形式逻辑。更具体地说,它是一个量化的多运算符模态逻辑（可以将句子作为参数并允许可能的状态[85]） ,其中包括事件演算（不同的事件触发不同类型的行为） ,这是使用的一阶演算随着时间的推移和变化进行推理。通过以这种方式形式化情绪（钦佩） ,代理人将钦佩与他人的行为联系起来。特征被形式化为一种行为的一系列实例化。如果对这些特征感到足够的钦佩,代理人就会学习这些特征,从而实例化美德。因此,为了将美德伦理实施到机器中,需要发展美德特征或特质。然而,美德伦理可能难以应用于个别情况[86],因此跨时间和文化应用美德会带来挑战[96]。因此,未来的研究可以将美德伦理学在 STS 不同背景下的适用性纳入其中。

4.4.3 评估后果。结果主义原则可以通过评估不同行动的后果来实施[62]。这可以通过根据代理人的后果有多少总福利对代理人的选择进行排名来完成[93]。 Dehghani 等人。 [25]通过选择具有最高效用的选择,用功利主义的原则来指定这一点。 Ajmeri 等人没有选择福利最高的结果。 [2]选择通过改善行动后果的最低限度经验来实施优先主义原则。使用后果的另一种方式是实施无嫉妒原则,其中 Sun 等人。 [94]以最低水平的团体或个人之间的嫉妒促进结果。其他原则,例如不同影响原则,着眼于后果中群体的代表性,并假定令人满意的结果将受到平等或成比例的对待[78]。然而,在预测一个动作可能产生的所有可能性时会出现一些问题,因为这在计算上可能具有挑战性,需要复杂的计算,而这些计算可能是错误的[43]。因此,在用于管理 STS 的推理能力的背景下模拟结果可能是未来研究的方向。在非确定性（动作不会产生某种影响)和概率性（使用概率表示不确定性)的环境中,无法模拟动作的所有可能后果是有局限性的。

因此,在 STS 中模拟后果可以从研究确定性环境开始,在这种环境中,每个动作的下一个状态都是已知的,并且后果是可获得的。然后可以将由此得出的发现扩展到更广泛的环境类型范围。

5 在社会技术系统中实施伦理原则

根据调查结果,我们提出了伦理原则在社会技术系统中的运作方式。利益相关者通过促进符合其价值观的规范来管理社会技术系统[18]。但是,可能会出现价值观和规范发生冲突的困境。将规范的伦理原则纳入价值和规范的推理可能有助于发展社会技术系统的公平治理[99]。在表 9 和表 10 中,对于我们分类法中的每个原则（图 2） ,我们建议潜在输入、可应用于这些输入以实施原则的规则,以及将规则应用于输入的潜在输出。

表 9. STS 中的操作道义原则。

原则	输入	规则	产出
	利益相关者的价值观和 规范		
平均主义	+ 利益相关者权利	给予利益相关者同等的权重	关于价值观的决定和 促进利益相关者平等权利的规范
非恶意 平均主义	+危害的定义和量化	在利益相关者之间平均分配损害	关于伤害平均分配的价值观和规范的决定
机会均等	+ 相关机会	确保机会平等地提供给所有利益相关者	关于促进平等获得机会的价值观和规范的决定
运气平均主义+运气的定义和量化		减轻运气的影响（例如,根据利益相关者的运气给他们一个权重)并确保收益 结果分发 利益相关者拥有 选择	关于减轻运气影响的价值观和规范的决定
自治平等主义	+自主性的定义和量化	确保最低 达到自主水平	关于价值观的决定和 促进自主的规范
比例主义	+ 利益相关者权利 + 利益相关者的贡献 化生产	根据贡献 利益相关者 调整权重 生产	关于价值观的决定和 根据利益相关者的贡献反映其权利的规范
自由主义比例主义	+ 利益相关者权利 + 利益相关者的总贡献 同意	根据同意 利益相关者 时的总贡献调整权重	关于价值观的决定和 反映利益相关者权利的规范,代表他们在同意时的贡献
基于沙漠的比例主义	+ 利益相关者权利 +运气的定义和量化	根据利益相关者的贡献调整权重,考虑运气的影响（即权重=贡献-运气）	关于反映利益相关者权利的价值观和规范的决定,代表他们的贡献减去运气的影响
康德主义	+ 行动的理由	确保没有人被当作一种手段来对待 结束	关于价值观的决定和 考虑原因的规范 采取行动确保没有人被视为达到目的的手段

表 10. 在 STS 中实施目的论原则。

原则	输入	规则	产出
	利益相关者的价值观和 规范		
美德	+ 利益相关者的特质优先考虑具有良好特质的利益相关者的偏好		良性利益相关者决定 关于价值观和规范,因为问题是通过应用美德特征来解决的
功利主义	+ conse 的效用 序列	效用总和最大化	关于导致具有最高效用的结果的价值观和规范的决定
享乐法实用主义 万物主义	+ 对每个结果的净愉悦 行动	最大化净快乐的行动	关于价值观和规范的决定会带来最大的快乐
规则功利主义	+道德规则的后果	选择规则 导致最好的整体 结果	根据具有最佳结果的规则对价值观和规范进行优先排序
最大最小值	+ conse 的效用 序列	最大化最小值 公用事业	关于价值观的决定和 具有改善最坏情况体验的后果的规范
无嫉妒	+利益相关者嫉妒的量化 结果	将导致结果的价值观和规范的嫉妒程度降到最低	嫉妒程度最低的顺序
双重原则 影响	+ 行动的后果	只选择有积极意义的行动 序列	关于价值观的决定和 导致积极后果的规范;如果无法预见,负面后果可能是允许的
不同的影响 教义	+ 对利益相关者的影响	确保影响平均或按比例分配	关于价值和规范的决定,其中影响在利益相关者之间平均或按比例分配 结果
不要 (伴奏) 伤害	+ 伤害的定义和量化 结果	不造成伤害（不 工具性伤害允许伤害作为意外的副作用）	关于其后果不会造成伤害的价值观和规范的决定

6在计算机科学和计算机科学中实施伦理原则方面的差距
人工智能

我们现在检查计算机科学和人工智能文献中伦理和公平研究中存在的差距,特别是与在多智能体系统中实施这些原则有关的差距。

6.1 扩展分类法主要差距包括缺乏对较少使用

原则的研究。我们建议未来的研究可以包括这些不太常见的原则,或者纳入更广泛的原则。

这包括研究西方学说以外的其他文化的原则。这不仅允许具有更好道德推理能力的代理,而且还有助于AI 代理的可解释性。在查看代理人做出特定决定的原因时,可以参考他们在解释中使用的确切原则。这将有助于技术的可及性和公平性,因为它可以更好地适用于来自不同背景的利益相关者群体。

6.2 在 STS 中实施伦理原则大多数已确定的研究并未明

确与 STS 相关。 Tolmeijer 等人。 [96]研究伦理原则如何与机器伦理相关,但不考虑伦理原则与 STS 背景下的价值观和规范的关系。阿杰梅里等人。 [2]在利用价值观和规范进行伦理推理的背景下广泛参考平等主义和功利主义的原则,但是这项研究可能会受益于对其他伦理原则的考虑,以实现更广泛的适用性。因此,未来的工作可以使这些作者建议的方法适应 STS 的背景。

6.3 解决道德困境最后,调查结果表明,每一个已确

定的道德原则都存在困难。

这意味着对于每项原则,在某些情况下都会导致不公平的结果。因此,伦理困境是指应用伦理原则导致不公平结果、无法支持一种行为优于另一种行为或与另一种伦理原则相冲突的情景。当一项原则不能支持一项行动优于另一项行动时,Azad-Manjiri [9]通过检查之前如何做出类似决定来解决困境。如果之前没有做出类似的决定,则随机选择一个动作。然而,依靠随机选择可能不会导致最合乎道德的行动。或者,可以通过使用多元主义方法来缓解道德困境,在这种方法中,可以权衡各种原则以找到最公平的答案。为此,使用特殊主义（将相关背景因素纳入道德推理以确定某个特征是否在道德上相关， [96]）,可以帮助确定哪种原则在该环境中最合适。因此,在如何解决实施特定伦理原则时可能出现的困难方面存在差距,这可能会在未来的工作中通过使用多元主义和特殊主义方法来解决。

7 结论

为了更好地解决对道德人工智能的追求,研究必须以人为本[31]。将视角转向检查系统治理的 STS 宏观伦理对于更好地实现这一目标至关重要[18]。在 STS 的治理中,利益相关者试图使规范与他们的价值观保持一致。
然而,当利益相关者有不同的偏好时,决策过程中可能会出现困境[75]。
以令人满意的方式解决这些困境,促进更高的公平、道德目标

原则可以帮助确定行为的道德允许性[64, 69]。在本次调查中,我们确定了以前在计算机科学中应用过的各种伦理原则（第 3 节）。我们还确定了在 AI 中实施伦理原则的关键方面,包括选择技术实施、阐明架构、指定伦理原则以及使用规则、后果或美德（第 4 节）。我们发现以前的文献没有在 STS 中实施伦理原则,并建议如何在抽象层面上做到这一点（第 5 节）。突出未来研究方向的主要差距包括扩大分类、在 STS 中实施原则,以及解决原则冲突或导致不公平结果的道德困境（第 6 节）。我们设想,通过将道德原则纳入治理中使用的推理能力,本次调查的结果将有助于开发更具道德规范的 STS 。

参考

[1] 基思·阿布尼。 2011. 机器人、伦理理论和元伦理学:困惑者指南。麻省理工学院出版社,剑桥,35-52。

[2] Nirav Ajmeri、Hui Guo、Pradeep K. Murukannaiah 和 Munindar P. Singh。 2020. Elessar:规范感知代理中的道德规范。在第 19 届国际自治代理和多代理系统会议 (AAMAS) 的会议记录中。IFAAMAS,奥克兰,16-24。 https://doi.org/10.5555/3398761.3398769

[3] Nirav Ajmeri 和 Pradeep Murukannaiah。 2021. 社会技术系统中的伦理学。在 BIAS - 布里斯托互动人工智能暑期学校。布里斯托尔大学,布里斯托尔。

[4]科林·艾伦、伊娃·斯密特和温德尔·瓦拉赫。 2005. 人工道德:自上而下、自下而上和混合方法。道德与信息技术 7, 3（2005 年 9 月）,149–155。 https://doi.org/10.1007/s10676-006-0004-4 [5]迈克尔·安德森和苏珊·利·安德森。 2007. 机器伦理:创造一个有道德的智能代理。 AI Magazine 28, 4（2007 年 12 月）,15。https ://doi.org/10.1609/aimag.v28i4.2065 [6] Michael Anderson 和 Susan Leigh Anderson。 2014. GenEth:通用道德困境分析器。在全国人工智能会议论文集中,卷。 1. 人工智能促进协会,魁北克,253-261。

[7] 迈克尔·安德森、苏珊·利·安德森和克里斯·阿曼。 2004. 迈向机器伦理。在 AAAI-04 车间关于代理组织:理论与实践。 AAAI,圣何塞,1–7。

[8]斯图尔特·阿姆斯特朗。 2015. 人工代理的动机价值选择.. 在 AAAI AI 和 AI 研讨会论文集中伦理。奥斯汀。

[9]梅萨姆·阿扎德-曼吉里。 2014. 基于 C4.5 决策树算法制作道德代理的新架构。国际信息技术与计算机科学杂志 6, 5 (2014), 50–57。 https://doi.org/10.5815/ijitcs。2014.05.07

[10] Fiona Berreby、Gauvain Bourgne 和 Jean-Gabriel Ganascia。 2017. 用于表示和应用道德原则的声明性模块化框架。在第 16 届自治代理和多代理系统 (AAMAS) 会议记录中。自治代理和多代理系统国际基金会,圣保罗,96-104。 https: //doi.org/10.5555/3091125.3091145 [11]鲁本宾斯。 2018. 机器学习中的公平性:政治哲学的教训。在第一届公平、问责制和透明度会议论文集（机器学习研究论文集,第 81 卷)中,Sorelle Friedler 和 Christo Wilson（编）。 PMLR,纽约,149-159。 https://proceedings.mlr.press/v81/binns18a.html [12] Niclas Boehmer 和 Rolf Niedermeier。 2021. 扩大计算社会选择的研究议程:多重偏好概况和多重解决方案。在第 20 届国际自治代理和多代理系统会议(AAMAS) 的会议记录中。自主代理和多代理系统国际基金会,伦敦,1-5。 https://doi.org/10.5555/3463952.3463954 [13]大卫布林克。 2007. 结果论的一些形式和局限性。牛津伦理理论手册 1, 1（ 2007 年 6 月）,381–423。 https://doi.org/10.1093/oxfordhb/9780195325911.003.0015 [14] 大英百科全书。 2021. 目的伦理学。 https://www.britannica.com/topic/teleological-ethics。访问时间:2021-09-23。

[15] Emanuelle Burton、Judy Goldsmith、Sven Koenig、Benjamin Kuipers、Nicholas Mattei 和 Toby Walsh。 2017.人工智能课程中的伦理考量。人工智能杂志 38, 2（2017 年 7 月）,22-34。 https://doi.org/10.1609/aimag。v38i2.2731

[16]坎苏坎卡． 2020. 实施人工智能伦理原则。公社。 ACM 63、12（2020 年 11 月）,18-21。 https: //doi.org/10.1145/3430368 [17] Lu Cheng, K. Varshney, and Huan Liu. 2021. 社会责任人工智能算法:问题、目的和挑战。JAIR 71（2021 年 8 月）,1137–1181。 https://doi.org/10.1613/jair.1.128142

[18] Amit Chopra 和 Munindar Singh。 2018. 社会技术系统和伦理在大。在 2018 年AAAI/ACM 人工智能、伦理和社会会议 (AIES) 会议记录中。计算机协会,新奥尔良,48-53。 https://doi.org/10.1145/3278721.3278740

[19]杰拉尔德·艾伦·科恩。 1989. 关于平等正义的货币。伦理学 99, 4 (1989), 906–944。 https://doi.org/10。2307/2381239

[20] Nicolas Cointe、Grégory Bonnet 和 Olivier Boissier。 2016. 多代理系统中代理行为的伦理判断。在 2016 年自治代理和多代理系统国际会议论文集中。 IFAAMAS,新加坡,1106–1114。

[21] Vincent Conitzer、Walter Sinnott-Armstrong、JS Borg、Yuan Deng 和 Max Kramer。 2017.人工智能的道德决策框架。在第 31 届 AAAI 人工智能会议 (AAAI) 的会议记录中。
AAAI,檀香山,4831–4835。

[22]戴维·卡米斯基。 1990. 康德后果论。伦理学 100, 3 (1990), 586–615。 https://doi.org/10.2307/2381810 [23] Mehdi Dastani 和 Vahid Yazdanpanah。 2022. 人工智能系统的责任。人工智能与社会 1, 1435-5655（2022 年 6 月）， 1-10。 https://doi.org/10.1007/s00146-022-01481-4 [24] Francien Dechesne、Gennaro Di Tosto、Virginia Dignum 和 Frank Dignum。 2013. 这里禁止吸烟:多代理系统中的价值观、规范和文化。人工智能与法律 21, 1（2013 年 3 月 1 日）,79-107。 https: //doi.org/10.1007/s10506-012-9128-5

[25] Morteza Dehghani、Emmett Tomai 和 Matthew Klenk。 2008. 道德决策的综合推理方法。机器伦理 3（2008 年 1 月）,1280–1286。 https://doi.org/10.1017/CBO9780511978036.024 [26]邓波尔。 2015. 机器伦理:机器人的困境。自然 523（2015 年 7 月）,24-26。问题 7558。https ://doi.org/10。1038/523024a

[27]路易丝·丹尼斯、迈克尔·费舍尔、玛丽亚·斯拉夫科维克和马特·韦伯斯特。 2016. 自治系统中道德选择的形式验证。机器人与自治系统 77 (2016),1-14。 https://doi.org/10.1016/j.robot.2015.11.012 [28]艾米丽·戴安娜、韦斯利·吉尔、迈克尔·卡恩斯、克里希纳拉姆·肯塔帕迪和亚伦·罗斯。 2021. 用于（宽松)Minimax 公平性的收敛算法。 CoRR abs/2011.03108 (2021), 1–22。 arXiv:2011.03108 [cs.LG] https://arxiv.org/abs/ 2011.03108

[29]弗吉尼亚·迪格努姆。 2019. 道德决策。斯普林格,查姆,35-46。 https://doi.org/10.1007/978-3-030-30371-6_3 [30] Virginia Dignum、Matteo Baldoni、Cristina Baroglio、Maurizio Caon、Raja Chatila、Louise Dennis、Gonzalo Génova、 Galit Haim、Malte S. Kließ、Maite Lopez-Sanchez、Roberto Micalizio、Juan Pavón、Marija Slavkovik、Matthijs Smakman、 Marlies van Steenbergen、Stefano Tedeschi、Leon van der Toree、Serena Villata 和 Tristan de Wildt。 2018. 设计伦理:必要性还是诅咒？。在 2018 年 AAAI/ACM 人工智能、伦理和社会会议论文集（美国路易斯安那州新奥尔良)(AIES 18)。美国纽约州计算机协会,60––66。 https://doi.org/10.1145/3278721.3278745

[31]弗吉尼亚·迪格努姆和弗兰克·迪格努姆。 2020. 特工已死。特工万岁！在第 19 届国际自治代理和多代理系统会议 (AAMAS) 的会议记录中。自治代理和多代理系统国际基金会(AAMAS),奥克兰,1701–1705。 https://doi.org/10.5555/3398761.3398957 [32] Veljko Dubljevi 、Sean Douglas、Jovan Milojevich、Nirav Ajmeri、William A. Bauer、George F. List 和 Munin dar P. Singh。 2021. 自动驾驶汽车的道德和社会后果。 CoRR abs/2101.11775（2021 年 1 月）,1-8。 arXiv:2101.11775 https://arxiv.org/abs/2101.11775 [33] Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。 2012. 通过意识实现公平。

在第三届理论计算机科学会议 (ITCS) 创新会议记录中。 ACM,剑桥,214-226。

[34]罗纳德·德沃金。 1981. 什么是平等?第 1 部分:福利平等。哲学与公共事务 10, 3 (1981), 185–246。 https://doi.org/10.2307/2264894

[35] Amitai Etzioni 和 Oren Etzioni。 2016. 人工智能辅助道德。道德与信息技术 18（2016 年 6 月）,149-156。第 2 期。https ://doi.org/10.1007/s10676-016-9400-6

[36] Amitai Etzioni 和 Oren Etzioni。 2017. 将道德融入人工智能。伦理学杂志 21（12 月 2017）,403-418。第 4 期。https://doi.org/10.1007/s10892-017-9252-2

[37] Jessica Fjeld、Nele Achten、Hannah Hilligoss、Adam Nagy 和 Madhulika Srikumar。 2020. 有原则的人工智能:将基于道德和权利的方法的共识映射到 AI 原则。伯克曼克莱因中心研究出版物第 2020-1 号。伯克曼克莱因中心,剑桥,1-39。 https://doi.org/10.2139/ssrn.3518482

[38] 马克弗勒贝。 2008. 公平、责任和福利。牛津大学出版社,牛津。

[39] Luciano Floridi 和 Josh Cowls。 2019. 社会人工智能五项原则的统一框架。哈佛数据科学评论 1、1（2019 年 7 月 1 日）,1。https: //doi.org/10.1162/99608f92.8cd550d1 https://hdsr.mitpress.mit.edu/pub/l0jsh9d1。

[40] Sorelle A. Friedler、Carlos Scheidegger 和 Suresh Venkatasubramanian。 2021. 公平的 (Im) 可能性:不同的价值体系需要不同的公平决策机制。 ACM 通讯(CACM) 64, 4（2021 年 3 月）,136–143。 https://doi.org/10.1145/3433949

[41] Naveen Sundar Govindarajulu 和 Selmer Bringsjord。 2017. 关于双重效应学说的自动化。在第二十六届国际人工智能联合会议论文集,IJCAI-17。 IJCAI,墨尔本， 4722–4730。 https://doi.org/10.24963/ijcai.2017/658 [42] Naveen Sundar Govindarajulu、Selmer Bringsjord、Rikhiya Ghosh 和 Vasanth
Sarathy。 2019. 迈向良性机器工程。在 2019 年 AAAI/ACM 人工智能、伦理和社会会议 (AIES 19) 的会议记录中。美国檀香山计算机协会，29 -35。 https://doi.org/10.1145/3306618.3314256 [43] Joshua Greene、Francesca Rossi、John Tasioulas、Kristen Brent Venable 和 Brian Williams。 2016. 在集体决策支持系统中嵌入道德原则。在第 13 届 AAAI 人工智能会议
(AAAI) 的会议记录中。 AAAI 出版社,雪鸟,4147–4151。 https://doi.org/10.5555/3016387.3016503 [44]蒂洛·哈根多夫。 2020. 人工智能伦理的伦理:指南评估。思维与机器 30（2020 年 3 月），

99–120。问题 1. https://doi.org/10.1007/s11023-020-09517-8
[45] Alexa Hagerty 和 Igor Rubinov。 2019. 全球人工智能伦理:人工智能的社会影响和伦理意义回顾。 CoRR abs/1907.07892（2019 年 7 月），1-27。 arXiv:1907.07892 http://arxiv.org/abs/1907.07892 [46] Ali Reza Honarvar 和 Nasser Ghasem-
Aghaee。 2009. 一种用于创建道德人工代理的人工神经网络方法。 2009 年 IEEE 国际机器人与自动化计算智能研讨会 - (CIRA)。 IEEE,大田,290–295。 https://doi.org/10.1109/CIRA.2009.5423190

[47] Muntean Ioan 和 Don Howard。 2017. 人工道德认知:道德功能主义和自主道德能动性。在哲学和计算:认识论、心灵哲学、逻辑和伦理学论文集,Thomas Powers （主编）。施普林格,在线。

[48] Anna Jobin、Marcello Ienca 和 Effy Vayena。 2019. 人工智能伦理准则的全球格局。自然机器
情报 1, 9（2019 年 9 月），389–399。 https://doi.org/10.1038/s42256-019-0088-2
[49]罗伯特·约翰逊和亚当·库尔顿。 2022. 康德的道德哲学。在《斯坦福哲学百科全书》（ 2022 年秋季版)中,Edward N. Zalta 和 Uri Nodelman（编）。斯坦福大学形而上学研究实验室,斯坦福大学。
[50] 雪莉·卡根。 1998. 平等与沙漠。牛津大学出版社,牛津,第 30 章,298–314。
[51]伊曼纽尔·康德。 2011. 伊曼纽尔康德:道德形而上学的基础:德英版。剑桥大学出版社,剑桥。 https://doi.org/10.1017/CBO9780511973741 [52] Emre Kazim 和 Adriano Koshiyama。 2020. 人工智能
伦理的高级概述。 SSRN 1、1（2020 年 5 月），1-18。
https://doi.org/10.2139/ssrn.3609292
[53] Arif Ali Khan、Sher Badshah、Peng Liang、Bilal Khan、Muhammad Waseem、Mahmood Niazi 和 Muhammad Azeem Akbar。 2021. 人工智能伦理:原则和挑战的系统文献综述。 CoRR abs/2109.07906 (2021), 1–17。 arXiv:2109.07906 https://arxiv.org/abs/2109.07906 [54] Tae Wan Kim、John Hooker 和 Thomas
Donaldson。 2021. 认真对待原则:价值的混合方法
结盟。 JAIR 70（2021 年 2 月），871–890。
[55] Barbara Kitchenham 和 Stuart Charters。 2007. 在软件工程中执行系统文献综述的指南。技术报告。基尔大学和杜伦大学联合报告。 https://www.elsevier.com/__ data/promis_misc/525444systematicreviewsguide.pdf [56] Nadin Kökciyan、Nefise Yaglikci 和 Pinar Yolum。 2017. 解决在线社交网络隐私纠
纷的论证方法。 ACM 跨。互联网技术。 17、3、第27条（2017年6月），22页。 https: //doi.org/10.1145/3003434 [57] Nadin Kökciyan 和 P nar Yolum。 2020. TURP:管理信任以规范物联网隐私。 IEEE

互联网计算 24, 6 (2020), 9–16。
[58] Shailendra Kumar 和 Sanghamitra Choudhury。 2022. 规范伦理、人权和人工智能。 AI & Ethics 2（2022 年 5 月），1-10。 https://doi.org/10.1007/s43681-022-00170-8 [59]德里克·莱本。 2020. 评估机
器学习公平性的规范原则。在 AAAI/ACM人工智能、伦理和社会 (AIES) 会议记录中。计算机协会,纽约,86-92。 https: //doi.org/10.1145/3375627.3375808 [60] Michelle Lee、Luciano Floridi 和 Jat Singh。 2021. 将算法公平性之外的权衡形式化:伦理哲学和福利经济学的教训。人工智能与伦理 1,
1 (06 2021), 529–544。 https://doi.org/10.1007/s43681-021- 00067-y

[61] Beishui Liao、Marija Slavkovik 和 Leendert van der Torre。 2019. Building Jiminy Cricket:利益相关者之间道德协议的架构。在 AAAI/ACM 人工智能、伦理和社会 (AIES) 会议记录中。 ACM,火奴鲁鲁,147-153。 https://doi.org/10.1145/3306618.3314257

[62] Raynaldio Limarga、Maurice Pagnucco、Yang Song 和 Abhaya Nayak。 2020.情境演算的机器伦理非单调推理。在 AI 2020 中:人工智能的进展。施普林格国际出版社,堪培拉,203–215。 https://doi.org/10.1007/978-3-030-64984-5_16

[63] Patrick Lin、Keith Abney 和 George Bekey。 2011. 机器人伦理:描绘机械化世界的问题。人工智能175, 5 (2011), 942–949。 https://doi.org/10.1016/j.artint.2010.11.026特别评论问题。

[64] Felix Lindner、Robert Mattmüller 和 Bernhard Nebel。 2019. 行动计划的道德许可。 AAAI 人工智能会议论文集33, 01（2019 年 7 月）,7635–7642。 https://doi.org/10.1609/aaai.v33i01.33017635 [65] Enrico Liscio、Michiel van der Meer、Luciano C. Siebert、Catholijn Jonker、Niek Mouter 和 Pradeep K. Murukannaiah。 2021. 轴:识别和评估特定上下文的值。在第 20 届国际自治代理和多代理系统会议 (AAMAS) 的会议记录中。自治代理和多代理系统国际基金会,伦敦,799-808。 https://doi.org/10.5555/3463952.3464048

[66] Sin Kit Lo、Qinghua Lu、Chen Wang、Hye-Young Paik 和 Liming Zhu。 2021. 联邦机器学习的系统文献综述:从软件工程的角度来看。 ACM 计算机。生存。 54, 5, 第 95 条（ 2021 年 5 月）,39 页。 https://doi.org/10.1145/3450288 [67] Ángeles Manjarrés、Celia Fernádez-Aller、Maite López-Sánchez、Juan Antonio Rodríguez-Aguilar 和 Manuel Sierra Castañer。 2021. 人工智能创造一个公平、公正、公平的世界。 IEEE 技术与社会杂志 40, 1 (2021), 19–24。 https://doi.org/10.1109/MTS.2021.3056292

[68]克尔斯滕·马丁。 2022. 护理伦理作为人工智能的道德基础。奥尔巴赫出版社,博卡拉顿,1-6。 https: //doi.org/10.1201/9781003278290 [69] Bruce M. McLaren。 2003. 扩展定义道德原则和案例:人工智能模型。人工智能 150, 1 (2003), 145–181。 https://doi.org/10.1016/S0004-3702(03)00135-8人工智能与法律。

[70] 约翰·米尔。 1863. 功利主义。朗文斯、格林公司。

[71]路易斯·莫尼兹·佩雷拉和阿里·萨普塔维贾亚。 2007. 用前瞻性逻辑建模道德。国际期刊基于推理的智能系统 1 (2007), 1–13。 https://doi.org/10.1504/IJRIS.2009.028020

[72] Nieves Montes 和 Carles Sierra。 2021. 参数规范系统的价值引导综合。在第 20 届国际自治代理和多代理系统会议 (AAMAS)的会议记录中。自治代理和多代理系统国际基金会,伦敦,907–915。

[73]杰西卡·莫利、阿纳特·埃尔哈拉尔、弗朗西斯卡·加西亚、利比·金赛、雅各布·莫坎德和卢西亚诺·弗洛里迪。 2021. 道德即服务:人工智能道德的务实运作。思维与机器 31（2021 年 6 月）,239–256。 https: //doi.org/10.1007/s11023-021-09563-w

[74] Andreasa Morris-Martin、Marina De Vos 和 Julian Padget。 2019. 多代理系统中的规范出现:一篇观点论文。自治代理和多代理系统 (JAAMAS) 33, 6 (2019), 706–749。

[75] Pradeep K. Murukannaiah、Nirav Ajmeri、Catholijn M. Jonker 和 Munindar P. Singh。 2020.道德多代理系统的新基础。在第 19 届国际自治代理和多代理系统会议(AAMAS) 的会议记录中。 IFAAMAS,奥克兰,1706-1710 年。 https://doi.org/10.5555/3398761.3398958 蓝天创意轨道。

[76] Pradeep K. Murukannaiah 和 Munindar P. Singh。 2020. 从机器伦理到互联网伦理:拓宽视野。 IEEE 互联网计算 24, 3（2020 年 5 月）,51-57。 https://doi.org/10.1109/MIC.2020.2989935 [77] R. Noothigattu、D. Bouneffouf、N. Mattei、R. Chandra、P. Madan、KR Varshney、M. Campbell、M. Singh 和 F.罗西。 2019. 使用强化学习和政策编排教授 AI 代理人的道德价值观。 IBM研究与开发杂志 63, 4/5 (2019), 2:1–2:9。 https://doi.org/10.1147/JRD.2019.2940428

[78] Deval Patel、Arindam Khan 和 Anand Louis。 2020. 背包问题的群体公平。 CoRR abs/2006.07832 （2020 年 6 月）,1-36。 arXiv:2006.07832 https://arxiv.org/abs/2006.07832

[79]普里亚·佩丹卡。 2021. 人工智能中的智能代理。 https://www.educba.com/intelligent-agent-in-ai/。访问时间:2021-12-21。

[80] Roxana Rădulescu、Patrick Mannion、Diederik M. Roijers 和 Ann Nowé。 2019. 多目标多代理决策:基于效用的分析和调查。 CoRR abs/1909.02964 (2019),1-48。 arXiv:1909.02964 HTTP: //arxiv.org/abs/1909.02964 [81] 约翰罗尔斯。 1967. 分配正义。哲学、政治与社会 1 (1967), 58–82。

[82]约翰·罗尔斯。 1985. 作为公平的正义:政治而非形而上学。哲学与公共事务 14, 3（1985 年夏季刊）， 223–251。 https://www.jstor.org/stable/2265349 [83]拉塞尔·罗宾斯和威廉·华莱士。 2007. 道德问题解决的决策支持:多代理方法。 决策支持系统 43, 4 (2007), 1571–1587。 https://doi.org/10.1016/j.dss.2006.03.003特刊集群。

[84] Manel Rodriguez-Soto、Marc Serramia、Maite Lopez-Sanchez 和 Juan Antonio Rodriguez-Aguilar。 2022.通过多目标强化学习灌输道德价值取向。伦理与信息技术 24, 1 （2022 年 1 月）,9。https ://doi.org/10.1007/s10676-022-09635-0

[85] Stuart J. Russell 和 Peter Norvig。 2016. 人工智能:一种现代方法。培生教育有限公司。

[86] Jeffrey Saltz、Michael Skirpan、Casey Fiesler、Micha Gorelick、Tom Yeh、Robert Heckman、Neil Dewar 和 Nathan Beard。 2019. 将道德融入机器学习课程。 ACM 跨。电脑。教育。 19、4（2019 年 8 月）,1-26。 https://doi.org/10.1145/3341164

[87]康拉德·桑德森、陆庆华、大卫·道格拉斯、徐希伟、朱黎明和乔恩·惠特尔。 2022. 迈向实施负责任的人工智能。 https://doi.org/10.48550/ARXIV.2205.04358

[88]沙洛姆 H 施瓦茨。 2012. 施瓦茨基本价值观理论概述。在线阅读心理学和文化 2, 1 (2012), 2307–0919。

[89] Amartya Sen. 1992. 重新审视不平等。克拉伦登出版社,牛津。
[90] Marc Serramia、Maite López-Sánchez、Juan A. Rodríguez-Aguilar、Manel Rodríguez、Michael Wooldridge、Javier Morales 和 Carlos Ansótegui。 2018. 规范决策中的道德价值观。在第 17 届自治代理和多代理系统 (AAMAS)会议记录中。 IFAAMAS,斯德哥尔摩，1294–1302。 https://doi.org/10.5555/3237383.3237891
[91] Munindar P. Singh。 2013. 规范作为管理社会技术系统的基础。 ACM智能交易系统与技术 (TIST) 5, 1, 第 21 条（2013 年 12 月）,23 页。
[92]沃尔特·辛诺特-阿姆斯特朗。 2021. 结果论。在斯坦福哲学百科全书（2021 年秋季版)中，爱德华 N. Zalta（主编）。斯坦福大学形而上学研究实验室,斯坦福大学。
[93]尤西·苏卡宁。 2017. 结果论、约束和相对好:对马克·施罗德的回复。杂志伦理与社会哲学 3, 1 (2017), 1–9。 https://doi.org/10.26556/jesp.v3i1.124。
[94]孙安康,陈博,杜安春。 2021. 分配不可分割的家务的公平标准与效率之间的联系。 CoRR abs/2101.07435（2021 年 1 月）,1-32。 arXiv:2101.07435 https://arxiv.org/abs/2101.07435 [95] Justin Svegliato、Samer Nashed 和 Shlomo Zilberstein。 2020. 道德自治系统的综合方法。人工智能和应用前沿 325 (2020), 2941 – 2942。https: //doi.org/10.3233/FAIA200464 [96] Suzanne Tolmeijer、Markus Kneer、Cristina Sarasua、Markus Christen 和 Abraham Bernstein。 2021.机器伦理的实施:一项调查。 ACM 计算机。生存。 53, 6, 第 132 条（2021 年 12 月）,38 页。 https://doi.org/10.1145/ 3419633

[97] Loïs Vanhée 和 Melania Borit。 2022. 观点:设计师的道德 - 如何培养符合道德的人工设计师智力。 JAIR 73 (2022), 619–631。
[98]温德尔·瓦拉赫、科林·艾伦和伊娃·斯密特。 2008. 机器道德:自下而上和自上而下的人类道德能力建模方法。人工智能与社会 22, 4 (2008), 565–582。 https://doi.org/10.1007/s00146-007-0099-0 [99]杰西卡·伍德盖特和尼拉夫·阿杰梅里。 2022. 治理公平社会技术系统的宏观伦理。在第 21 届自治代理和多代理系统 (AAMAS) 国际会议论文集中。 IFAAMAS,在线， 1824-1828 年。 https://doi.org/10.5555/3535850.3536118蓝天创意轨道。

[100] Vahid Yazdanpanah、Enrico Gerding、Sebastian Stein、Mehdi Dastani、Catholijn M. Jonker 和 Timothy Norman。 2021. 可信自治系统的责任研究。在第 20 届自治代理和多代理系统国际会议（03/05/21 - 07/05/21）。自治代理和多代理系统国际基金会(AAMAS),在线,57–62。 https://eprints.soton.ac.uk/447511/ [101] Han Yu、Zhiqi Shen、Chunyan Miao、Cyril Leung、Victor R. Lesser 和 Qiang Yang。 2018. 将道德规范融入人工智能。在第 27 届国际人工智能联合会议论文集中,IJCAI。 IJCAI,斯德哥尔摩，5527–5533。
[102]朱黎明、徐希伟、陆青华、Guido Governatori 和 Jon Whittle。 2022. AI 和道德规范 实施负责任的 AI。斯普林格国际出版社,Cham，15-33。 https://doi.org/10.1007/978-3-030-72188-6_2

方法概述

图 3 可视化了用于回答研究问题的方法。这是在分析文献中的原则识别(QP)和原则实施(QO)的同时两部分过程中进行的。通过通读和总结关键点对作品进行定性分析,然后将它们归入相关的分类中,这些分类涉及它们所涉及的原理以及它们的研究类型（见表1）。然后汇总这些单独的分析以检查整体的发现。一些作品更具理论性,探索原理的存在以及它们如何与计算机科学相关（例如， [59]）。这些工作对于识别原则(QP) 很有用。其他研究采用既定原则并加以实施,这有助于回答QO （例如， [94]）。一些作品混合了识别和实现（例如， [54]）。该分析是在与第二位作者协商后进行的,第二位作者批判性地检查了所审查的作品和第一位作者提取的发现。

B 对有效性和缓解的威胁

出现了五个对有效性的威胁,这里总结了这些威胁,以及尝试的缓解措施。确定的第一个威胁是,只有写成或翻译成英文的论文才会被纳入我们用于开发分类法的审查中。这意味着其他语言的相关研究

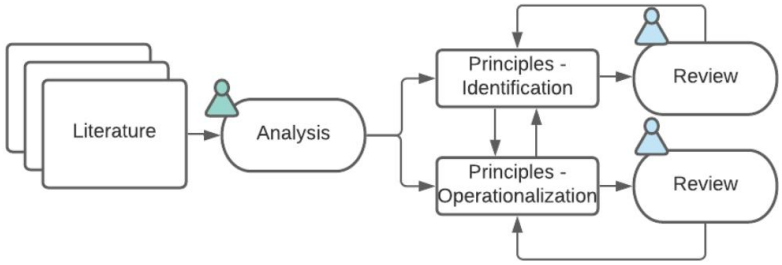

图 3. 从文献中提取原则识别和操作化的方法

可能会被遗漏,这可能会导致文化偏见,从而威胁到研究的外部和内部有效性。内部有效性受到其他语言中引用的缺失伦理原则的威胁,外部有效性受到调查结果跨文化应用减少的威胁。这可以通过寻求具有国际作者身份的论文来缓解,但它被认为是一个悬而未决的问题,可以通过未来将该方法应用于其他语言的研究来解决。

内部有效性的第二个威胁是遗漏关键词的可能性,这可能再次导致相关研究被排除在外。最初的搜索字符串基于初步研究,随着审查的继续,更多的关键术语被识别出来。为解决这一问题,确保审查的目标范围得到仔细界定,从而可以识别出一系列良好的初始相关术语。随着更多术语的识别,确保遵循相关引用并将这些术语包括在内。

存在资源缺失的第三个相关威胁,这对研究的内部有效性具有类似的影响。此处研究的主题涉及广泛的研究领域,人机交互和软件工程等领域未明确包含在搜索中,但可能包含相关研究。这种威胁是通过使用两个大型在线图书馆作为初始资源来解决的,这些资源链接到各种其他资源。还遵循选定研究的引用,扩大了出版物的范围。然而,未来的研究还可能包括在这些其他领域复制该方法。

第四,时间限制威胁到内部有效性,因为只有时间搜索结果的前五页(加上引文)。这可能意味着没有足够的时间进行这些页面之外的相关工作。为了在这个时间限制内尽可能地做最好的研究,我们追求引用,并且广泛遵循Kitchenhams [55]的系统文献综述指南。这有助于有效地识别相关研究。另一方面,通过将我们的方法应用于比此处确定的更多研究的分析,这种限制可能会导致该领域的进一步研究。

研究人员偏见的第五个问题也威胁到内部有效性,因为它可以在特定方向上影响结果而不是客观的。这可以通过让二级审稿人批判性地分析结果并提出建议来帮助主要审稿人改进研究来缓解。这也可以通过将研究选择标准基于研究问题并在审查开始之前对其进行定义来解决。

表 11. 纳入和排除标准

包容	排除
发表作品见于:AIES、FAccT、AAAI、IJCAI, (J)AAMAS, TAAS, TIST, JAIR, AIJ, 自然, 科学恩斯	关于计算机科学以外的元伦理学或应用伦理学的著作
个人和/或团体公平	关于特定 ML 方法的研究
多用户社交困境	非社会困境
规范伦理和多用户 AI 和/或 MAS	多用户 AI 和/或 MAS 的非伦理研究
规范伦理和STS	STS 的非伦理研究
规范伦理原则和人工智能	人工智能基石
与道德原则相关的偏见	不参考伦理原则的偏见研究