伍德盖特和阿杰梅里

社会技术系统宏观伦理原则：分类学和未来方向

杰西卡伍德盖特 yp19484@bristol.ac.uk
尼拉夫·阿杰梅里 nirav.ajmeri@bristol.ac.uk
布里斯托大学计算机科学系,英国布里斯托,BS8 1UB

抽象的

人工智能 (AI) 的迅速采用需要对其伦理影响进行仔细分析。在解决道德和公平问题时,重要的是要检查整个范围的道德相关特征,而不是单独查看单个代理人。

这可以通过将视角转移到嵌入代理的系统来实现,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 STS 中的规范发生冲突时,STS 中就会出现多用户社会困境。要发展满足不同利益相关者需求的公平治理,并以更高的公平目标以令人满意的方式解决这些困境,我们需要在推理中整合各种规范的伦理原则。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。

这项工作开发了一种规范伦理原则的分类法,可以在 STS 的治理中加以实施。我们确定了一系列道德原则,在分类树上有 25 个节点。我们描述了每项原则以前被实施的方式,并建议如何将原则的实施应用于 STS 的宏观伦理。我们进一步解释了每个原则可能出现的潜在困难。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。

一、简介

人工智能系统的快速发展需要了解其伦理影响的重要性（Dastani & Yazdanpanah,2022）。最近代理研究从强调单一代理转向多代理系统（MAS:将多个技术代理部署到一个共同的环境中,R adulescu 等人,2019 年)需要仔细分析 MAS 的伦理影响（Dignum & Dignum,2020 年） ;乔普拉和辛格,2018 年）。要在考虑道德的情况下开发这些系统,公平是关键（Floridi & Cowls,2019）。公平被理解为非歧视,其中歧视是基于敏感属性对人的偏见（Bishr,2018 年;Mehrabi 等人,2021 年）。在追求公平 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（Murukannaniah & Singh,2020）。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,采用以下观点也很重要

宏观伦理学（Chopra & Singh,2018）。宏观伦理学侧重于 STS 的治理,并解决所有与伦理相关的特征,与微观伦理学相反,微观伦理学侧重于 STS 内单个代理人决策的更有限的视角。因此,我们的工作范围在于 STS 的宏观伦理视角,考虑系统的治理。

STS 的治理是指利益相关者试图促进符合他们的价值观（什么对我们生活很重要,Schwartz,2012 年)的结果和规范（预期行为规则,Morris-Martin 等人,2019 年）。这一点很重要,因为伦理应该被理解为一个包含背景的反思性发展过程（K¨okciyan & Yolum, 2020; Morley et al., 2021; Manjarr´es et al., 2021; Zhu et al., 2022）。因此,价值观和规范对于道德推理至关重要（Ajmeri 等人,2020 年;Dignum 等人,2018 年;Singh,2013 年;Yazdanpanah 等人,2021 年）。然而,用户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（Dechesne 等人,2013 年）。因此,在针对多个用户做出决策时出现了挑战（K¨okciyan 等人,2017 年;Liao 等人,2019 年）。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,例如,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。

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（Woodgate & Ajmeri,2022）。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际手段（Murukannaiah & Singh,2020）。通过检查规范伦理理论如何被用于改善公平性考虑,以及它们如何被操作以在人工智能 (AI) 中做出伦理决策,有可能实施规范伦理理论以在 MAS 中做出具有整体公平的目标。在 Binns（2018 年)和 Leben（2020 年)等作品中,规范的伦理原则曾被用于为二进制 ML 算法选择公平性指标的背景下。在代理人（执行行动以实现目标的行为实体,即使用 AI Pedamkar,2021 做出的决策)决策中实施规范道德原则已被用于使代理人能够在特定情况下做出道德判断（Cointe 等人,2021 年）。 , 2016).它们还可以用于改进系统分析中的公平性考虑（Saltz 等人,2019 年;Conitzer 等人,2017 年）。

1.1 伦理原则分类的动机

因此,这项工作的动机源于提高 MAS 中公平性考虑的需要。通过在 STS 治理中实施规范的伦理原则,可以提高公平性（Woodgate & Ajmeri,2022）。道德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合乎道德（Kim 等人,2021 年）。因此,道德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（Conitzer 等人,2017 年）。这些原则有助于指导规范判断、理解不同观点并确定具体行动方案的道德许可（Canca,2020 年;McLaren,2003 年;Saltz 等人,2019 年;Lindner 等人,2019 年）。

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（Burton 等人,2017 年;Robbins 和 Wallace,2007 年）。我们设想道德原则的分类将有助于这种道德思考。

1.2 相关研究的空白

在人工智能伦理的背景下,有两种类型的原则被提及:(1) 那些从规范伦理学中推导出来的原则,例如 Leben (2020) 中发现的道义论和结果论,以及 (2) 那些改编自医学等其他学科的原则和生物伦理学,例如 Floridi 和 Cowls（2019 年）、Jobin 等人建议的那些。 (2019),Fjeld 等人。 (2020),和 Cheng 等人。 (2021) 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。

这两类原则是相关的,但又是不同的领域,很容易混淆。为确保术语清晰,我们将规范伦理学中的原则称为伦理原则,以及 Floridi 和 Cowls（2019 年)和 Jobin 等人强调的原则。 (2019) 作为 AI 基石。

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、可解释性、人类对技术的控制以及应该支撑人工智能的人类价值观的提升AI技术的设计。

伦理原则从哲学理论中推断出的可操作规则,例如
道义论和结果论。

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。 (2019)、Floridi 和 Cowls (2019) 以及 Khan 等人。 （2021）,但是,不符合此处定义的道德原则。 Tolmeijer 等人的工作。 (2021) 有很多相关信息,但是,作者并没有捕捉到我们捕捉到的全部道德原则。此外,Tolmeijer 等人。从机器伦理的角度来看伦理,而不是 AI 伦理及其与 MAS 的关系,正如我们旨在解决的那样。

同样,Yu 等人。 (2018) 确定了道德原则的高层次概述,但未能认识到我们工作中发现的范围,也没有在相同的深度上考虑它们。 Dignum (2019)、Leben (2020)、Robbins 和 Wallace (2007) 对规范伦理进行了总结,但是,他们没有考虑 Linder 等人提到的“不伤害”等原则。 （2019）。为了实现更广泛的适用性,这些作品可能会受益于正式的分类法,包括计算机科学中的其他伦理原则。

1.3 目标与贡献

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）。迄今为止在计算机科学中提出了哪些伦理原则

文学?

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。

Qo （运作化）。伦理原则如何在人工智能和计算机科学研究中得到应用？

这个问题着眼于确定的原则，以检查它们是如何在人工智能和计算机科学中运作的。 Leben (2020) 和 Tolmeijer 等人的作品。 (2021) 就如何实施某些道德原则提供了一些指导，但是，它们并不广泛并且遗漏了一些原则。

Qg （间隙）。人工智能和计算机的伦理和公平研究存在哪些差距
科学，特别是与 STS 中的操作化原则有关？
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。

1.4 组织机构

第 2 节简要解释了该方法，以及对有效性的威胁。这可能有助于未来的研究通过重现此处使用的方法来扩展伦理原则的分类。第 3 部分探讨了我们的发现，评估了每个目标的实现方式并强调了现有差距。第 4 节包括每个 De 本体论伦理原则的详细信息，包括它们以前是如何操作的以及可能出现的潜在困难。第 5 节对目的论原则做同样的事情。

第 6 节总结了我们的主要收获和未来工作的方向。

2. 重现性方法简介

从软件工程研究中对分类法的可重复性和可扩展性的启发，我们遵循 Kitchenham (2007) 的指导方针进行系统的文献回顾，以开发我们的道德原则分类法。图 1 简要概述了我们的方法。

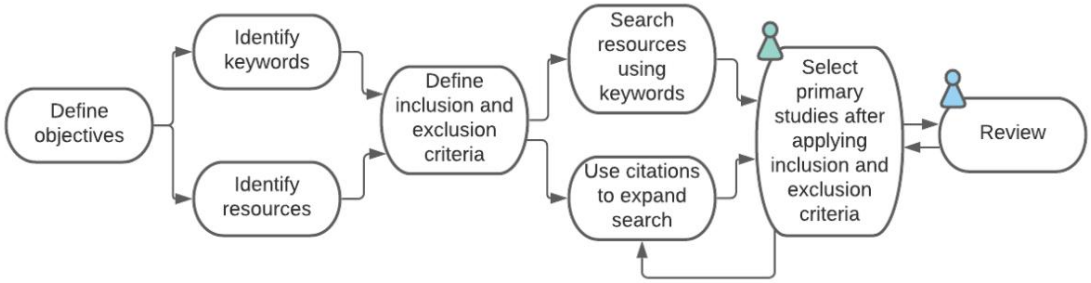

图 1:方法简介

在确定了我们的目标和问题之后，我们制定了通过识别关键字和资源来搜索主要研究的策略。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。使用的搜索字符串是 (AI OR Agent OR ML OR Multiple-User OR Multiagent) AND (Bias OR

“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”）。

在检查每个资源的前 5 页结果后,我们通过对标题应用包含和排除标准来缩小搜索范围,消除明显不相关的研究。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。

2.1 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。我们通过排除这些资源之外的作品来降低风险,并通过遵循主要研究的相关引用来扩大范围。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。

2.2 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 54 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 6 篇论文被纳入审查范围。根据我们的审查,我们通过迭代过程创建了道德原则分类法,在文献中确定新原则时添加节点并相应地修改结构。

3. 计算机科学和人工智能的伦理原则

基于对规范伦理的定义,人工智能和伦理原则的研究大致分为十二个关键原则,以及基于论文结构和贡献的五种研究类型。表 1 和表 2 列出了我们的发现。

在伦理学中,有两个主要的理论分支:道义论和目的论。道义论围绕着规则、权利和义务展开（Murukannaiah & Singh, 2020; Wallach et al., 2008）。另一方面,目的论伦理学从作为要实现的目标的善或可取的东西中推导出责任或道德义务（大英百科全书,2021）。目的论伦理可以进一步分为结果主义、利己主义和美德伦理（SOAS,2021）。图 2 以树状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。

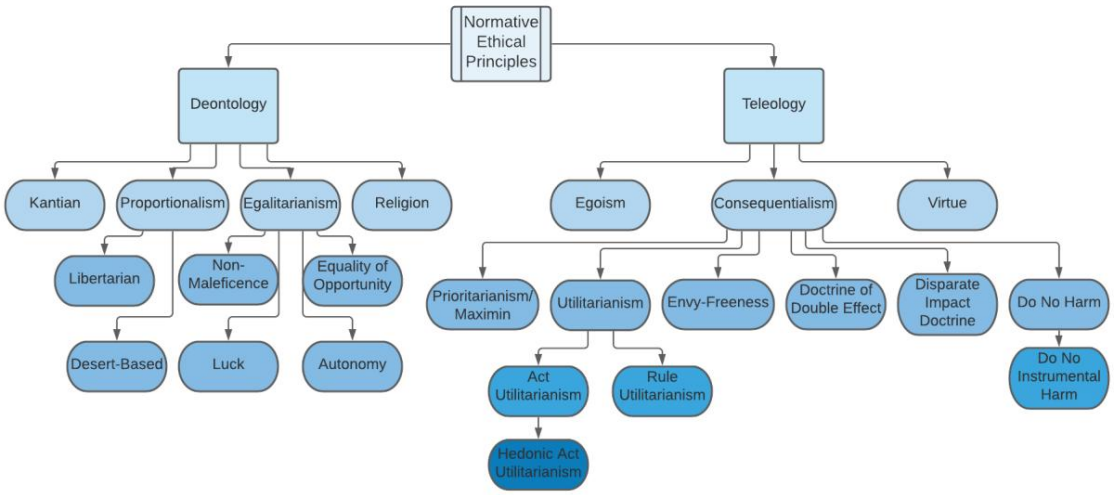

图 2:道德原则分类

我们发现,功利主义等某些原则比不伤害等其他原则得到更多讨论。这也许是因为功利主义是一个历史悠久的理论,因此它是众所周知的,更可能被研究人员利用。

我们还发现,有大量研究将“道义论”和“结果论”作为广义术语引用,但没有具体说明它们指的是什么类型的道义论或结果论,例如 Cointe 等人。 (2016),格林等人。 (2016),以及安德森和安德森 (2014)。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,以便更精确地实施。

3.1 伦理原则的实施

我们迭代了我们审查中确定的论文,以对以前的道德原则运作进行分析。我们发现,架构必须定义为以自上而下、自下而上或混合方法将原则整合到推理能力中。除此之外,福利的定义是必要的,以便

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）
道义论	（Abney,2011 年;Binns,2018 年;Brink,2007 年;Cointe 等人,2016 年;Greene、Rossi、Tasioulas、Venable 和 Williams,2016 年;Leben，2020 年;Murukannaiah 和 Singh,2020 年;Saltz 等人,2019 年;瓦拉赫、艾伦和斯密特,2008）	（Anderson 和 Anderson,2014 年;Berreby、Bourgne 和 Ganascia,2017 年;Dehghani、Tomai 和 Klenk,2008 年;Honarvar 和 Ghasem-Aghaee,2009 年;Limarga、Pag nucco、Song 和 Nayak,2020 年;Lindner 等人， 2019 年;罗宾斯和华莱士,2007 年）
平等主义（Binns,2018 年;Cohen,1989 年;Dworkin,1981 年;	Fleurbaey,2008 年;Friedler、Schei degger 和 Venkatasubramanian,2021 年;Leben,2020 年;Murukannaiah、Ajmeri、Jonker 和 Singh，2020 年;Rawls,1985 年;Sen,1992 年)	（Dwork、Hardt、Pitassi、Reingold 和 Zemel，2012 年）
比例主义（Etzioni & Etzioni,2016 年;Kagan,1998 年;Leben，	2020 年）	（德沃克等人,2012 年）
康德的	（Abney,2011 年;Kant,2011 年;Hagerty & Rubinov,2019 年;Kim 等人,2021 年;Wal lach 等人,2008 年）	（Berreby 等人,2017 年;Limarga 等人,2020 年;Robbins 和 Wallace,2007 年）
美德	（Abney,2011 年;Anderson 和 Anderson，2007 年;Brink,2007 年;Cointe 等人,2016 年;Greene 等人,2016 年;Hagerty 和 Ru binov，2019 年;Murukannaiah 和 Singh,2020 年;Saltz 等人,2019 年;Wallach等人,2008）	(Govindarajulu, Bringsjord, Ghosh, & 萨拉蒂,2019 年;霍纳瓦尔和加西姆阿哈伊, 2009;罗宾斯和华莱士,2007）
结果论（Abney,2011 年;Brink,2007 年;Cointe 等人,2016 年;	Cummiskey,1990 年;Greene 等人,2016 年;Hagerty & Rubinov,2019 年;Leben,2020 年;Saltz 等人,2019 年;Sinnott Armstrong,2021 年;苏卡宁,2017）	（Berreby 等人,2017 年;Limarga 等人,2020 年）
功利主义	（Abney,2011 年;Anderson 和 Ander son，2007 年;Brink,2007 年;Honarvar 和 Ghasem-Aghaee,2009 年;Kim 等人,2021 年;Leben，2020 年;Mill,1863 年;Mu rukannaiah 等人，2020 年;Murukannaiah 和 Singh , 2020 年;Wallach 等人,2008 年）	（Ajmeri 等人,2020 年;Anderson、Ander son 和 Armen,2004 年;Berreby 等人,2017 年;Dehghani 等人,2008 年;Limarga 等人，2020 年;Lindner 等人,2019 年;Rob bins 和 Wallace , 2007)
最大最小值	（生活,2020 年;罗尔斯,1967 年）	（Ajmeri 等人,2020 年）
无嫉妒心（Boehmer & Niedermeier,2021）	-	-
教义的双重效果	-	（Berreby 等人,2017 年;Govindarajulu 和 Bringsjord,2017 年;Lindner 等人,2019 年）
不同影响原则	（宾斯,2018 年）	-
不要伤害	（丹尼斯、费舍尔、斯拉夫科维克和韦伯斯特,2016 年）	（林德纳等人,2019）

表 2:根据提取的原则审查的研究分类:算法和观点或评论研究

类型 / 话题	算法	观点或评论
道义论	（罗德里格斯-索托、塞拉米亚、洛佩兹桑切斯和罗德里格斯-阿吉拉尔,2022）	（Kazim 和 Koshiyama,2020 年;Hagendorff,2020 年;Tolmeijer 等人,2021 年;Yu 等人,2018 年）
平均主义		（李、弗洛里迪和辛格,2021 年）
比例主义 –		-
康德的	-	（Tolmeijer 等人,2021 年;Kumar 和 Choudhury,2022 年）
美德	（罗德里格斯-索托等人,2022 年）	（Kazim 和 Koshiyama,2020 年;Hagendorff,2020 年;Tolmeijer 等人,2021 年;Vanh´ee 和 Borit,2022 年;Yu 等人,2018 年）
结果论（Rodriguez-Soto 等人,2022 年）		（Etzioni & Etzioni,2017 年;Tolmeijer 等人,2021 年;Yu 等人,2018 年）
功利主义	（罗德里格斯-索托等人,2022 年）	（Etzioni 和 Etzioni,2017 年;Kazim 和 Koshiyama,2020 年;Yu 等人,2018 年;Kumar 和 Choudhury,2022 年）
最大最小值	（Diana、Gill、Kearns、Kenthapadi 和 Roth,2021 年;Sun、Chen 和 Doan,2021 年）	（李等人,2021 年）
无嫉妒心（Sun 等人,2021 年）		（李等人,2021 年）
学说 双重效果	-	（邓,2015）
不同影响原则	（帕特尔、可汗和路易斯,2020 年）	-
不要伤害	-	-

了解“善”被理解为什么,或者该原则在其应用中的目标是什么。推理中原则的操作化主要分为三个阵营,在这些阵营中选择行动要么根据:(1)行动如何遵守某些规则,(2)通过评估行动产生的后果,或(3)通过发展美德。还有各种因素取决于特定的原则,以确定它们是否是实施该原则所必需的。

3.1.1 阐明架构

自下而上的机器通过观察学习做出道德决策,而无需学习任何正式规则(Etzioni & Etzioni,2017)。 自上而下的道德规范直接安装到机器中(Kim 等人,2021 年),作为规定什么是道德上正确的行为的规则(Lin 等人,2011 年)。 混合 结合道德推理和经验观察,通过自下而上的上下文观察补充自上而下的规则实施(Berreby 等人,2017)。

为了设计道德敏感系统,从业者必须决定整合伦理理论的架构(Wallach 等人,2008 年)。这些属于两种广泛的方法:自上而下强加伦理理论,以及自下而上构建目标可能明确或未明确指定的系统。

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(Etzioni & Etzioni,2017)。自下而上的方法通常要求机器从经验中学习规范,而自上而下的方法将道德直接植入机器中(Kim 等人,2021 年)。自上而下的方法是基于规则的:道德被理解为通过确定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(Lin 等人,2011)。除了这些之外,还有混合方法,其中结合了道德推理和经验观察。

我们发现许多作品(Limarga 等人,2020 年;Sun 等人,2021 年;Diana 等人,2021 年;Dehghani 等人,2008 年)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。其他作品如 Berreby 等人。 (2017) 实施了混合方法,其中自上而下的规则实施辅以自下而上的上下文信息观察。我们没有发现任何作品以纯粹自下而上的方式使用道德原则。未来的研究可以从经验中推断出伦理原则。

然而,外推的一个困难是如果代理人没有先验知识或定义,如何将道德原则形式化。

3.1.2 福利的定义

福利 评估什么是好的或有价值的,以及什么是令人满意的结果(Binns,2018 年)。

对于所有原则,无论是道义论的还是目的论的,我们发现为了实施它们,有必要对福利进行定义。这是因为在追求公平的过程中必须考虑福利(Fleurbaey,2008)。福利可以是快乐或偏好满足

化 (Cohen, 1989)、收入和资产 (Rawls, 1958; Dworkin, 1981),或做某些事情所需的能力和资源 (Sen, 1992)。它是对什么是好的或有价值的,以及什么是令人满意的结果的评估（Binns,2018 年）。换句话说,这就是应用某个原则的目的。福利的定义可能看起来很抽象,并且因背景和文化而异。因此,需要采用一致且公平的方法来确定福利。 Murukannaiah 等人在 2018 年使用福利来实施平等主义（指出人类在某种基本意义上是平等的,应该努力纠正不平等形式,Binns,2018 年)的一个例子。 （2020）。在这里,作者建议该原则需要在满足利益相关者偏好方面最大化利益相关者之间的差异。因此,在这种情况下,福利被理解为偏好满足,并且在应用平等主义原则时,平等的偏好满足就是目标。

3.1.3 使用规则、后果或美德

我们发现,以前的工作已经通过规则的应用,然后根据它如何符合某些规则来选择一个行动,通过评估后果,然后根据它产生的后果选择一个行动,或者通过发展良性特征来实现原则的操作化。 .

- 应用规则 一些方法建议通过将一组规则应用于可能的行动以确定哪些是令人满意的来实施原则,例如 Abney (2011)、Greene 等人的 De 本体论实施。 (2016),和 Berreby 等人。 (2017)。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大化的规则（Murukannaiah 等人,2020 年）。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（Leben,2020）。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（Tolmeijer 等人,2021 年）。因此,创建将道德原则编码为规则以在 STS 环境中使用的系统方法可能是未来研究的方向。

- 分析后果 后果主义原则可以通过评估不同行动的后果来实施（Limarga 等人,2020 年）。这可以通过根据代理人的后果产生多少总福利来对代理人的选择进行排名来完成（Suikkanen,2017 年）。 Dehghani 等人。 (2008) 通过选择效用最高的选项,用功利主义的原则来说明这一点。 Ajmeri 等人没有选择福利最高的结果。 (2020) 选择通过改善行动后果的最低经验来实施优先主义原则。

使用后果的另一种方式是实施 Envy-Freeness 原则,其中 Sun 等人。 (2021) 促进团体或个人之间嫉妒程度最低的结果。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得到平等或成比例的对待（Patel 等人,2020 年）。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（Greene 等人,

2016 年）。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。

- 培养美德 对于美德伦理,道德源于个人的内在性格（Murukannaiah 和 Singh,2020 年;Wallach 等人,2008 年;Brink,2007 年;Kazim 和 Koshiyama,2020 年;Yu 等人,2018 年;Abney,2011 年）).要根据这一理论解决问题,应运用良性特征（Robbins & Wallace,2007）。因此,该理论可以通过美德的实例化来实施（Tolmeijer 等人,2021）。

  Govindarajulu 等人举例说明了这一点。 (2017) 通过使用计算形式逻辑将情绪形式化、表示特征并建立学习特征的过程来实例化美德。因此,为了将美德伦理实施到机器中,需要培养美德特征。这在以前的文献中已经通过利用计算形式逻辑来建立良性特征来完成,那些具有良性特征的机器将根据定义以道德方式行事。然而,美德伦理很难应用于个别情况（Saltz et al., 2019）,因此跨时间和文化应用美德会带来挑战（Tolmeijer et al., 2021） .因此,未来的研究可以将美德伦理在 MAS 不同背景下的适用性纳入其中。

3.1.4 考虑其他输入

最后,我们确定了各种因原则而异的因素,以确定它们是否是实施所必需的。这些输入可能很难获得,因为它们看起来很主观,尤其是贡献和自主性的输入。因此,进一步研究分析这些输入的作用将是有帮助的,并且也有助于了解是否有任何其他输入未在此处确定。

- 运气 为了应用运气平均主义的原则,必须考虑运气水平,因为它意味着没有人会因运气不佳而最终变得更糟（Lee 等人,2021 年）。
  这意味着人们不应该因为他们无法控制的因素（例如出生环境)而变得更糟或更好。

- 自主性 自主性被理解为代理人行为自由的程度（代理人有理由的行为,Kim 等人,2021 年） ,是平等自治原则的必要输入。该原则表示应达到最低水平的自治并平均分配（Fleurbaey,2008）。因此,为了取得令人满意的结果,有必要了解现有的自治水平,以便应用该原则。

- 贡献 考虑贡献对于实施某些原则（例如自由主义比例主义)是必要的。对于这一原则,可以通过评估每个人在同意时的总贡献来找到令人满意的结果（Leben,2020）。为此,有必要定义人们贡献的资源（例如,努力、金钱、时间等）。

- 效用 对于功利主义原则,有必要获得效用的定义和数量以便应用它们。这是因为这些原则促进了行动

最大化效用或“善”（Limarga 等人,2020 年）。因此,具有最高效用的解决方案将是那些最符合功利主义原则的解决方案。

· 嫉妒 在应用无嫉妒原则时,应消除或尽量减少嫉妒（Sun 等人,2021 年）。因此,为了做到这一点,必须在决策过程中考虑嫉妒程度。

3.2 计算机科学与人工智能在伦理与公平研究方面的差距
智力

我们知道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,特别是与在多代理系统中实施这些原则有关的差距。

扩展分类法。主要差距包括缺乏对双重效应和比例主义等较少利用原则的研究。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。

研究西方学说之外的其他文化的原则也很重要,这些原则应该纳入道德推理和道德人工智能代理的设计中。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。

考虑 STS 中的道德原则。大多数已确定的研究并未明确与 STS 相关。 Tolmeijer 等人。 (2021) 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。 (2020) 在利用价值观和规范进行伦理推理的背景下广泛参考了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。

解决道德困境。最后,调查结果表明,每一项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。这可以通过使用多元主义方法来缓解,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,Tolmeijer 等人,2021 年)可以帮助确定哪个原则最合适环境。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。

4.道义伦理原则

本节检查审查中确定的每个道义原则。对于本节和第 5 节,每个小节的结构首先解释以前的工作如何定义该原则,然后解释以前的工作如何实施该原则。如果之前没有发现实施该原则的工作,我们建议如何在 STS 的背景下实施它。然后我们解释应用该原则的潜在困难。

4.1 道义论

道义论 遵守规则、法律和规范（Murukannaiah & Singh,2020;Ha gendorff,2020）,并尊重源于义务和权利的相关义务和许可（Cointe 等,2016）（Wallach 等,2008； Saltz 等人,2019 年;Yu 等人,2018 年;Kazim 和 Koshiyama,2020 年;罗宾斯和华莱士,2007 年）。

道义论需要遵守规则、法律和规范（Murukannaiah & Singh,2020;Hagendorff,2020）,并尊重源于义务和权利的相关义务和许可（Cointe 等,2016）（Wallach 等,2008； Saltz 等人,2019 年;Yu 等人，2018 年;Kazim 和 Koshiyama,2020 年;Robbins 和 Wallace,2007 年;Rodriguez-Soto 等人,2022 年）。对于道义论而言,行为的可允许性在于行为本身的内在特征;根据道义论的方法,当且仅当行为本身在道德上是善的或不同的,而与其产生的价值无关时,该行为才是允许的（Lindner 等人,2019 年;Limarga 等人，2020 年;Brink,2007 年）。

操作道义论。可以通过采用一组在设计中实施（使用自上而下的架构)或通过学习（使用自下而上的架构)获得的规则来实现道义论的操作,并确保技术与它们一致（Abney， 2011 年;格林等人,2016 年）。这些规则可以补充上下文信息（包括福利的定义)和偏好聚合以获得满意的结果。实施这种方法的框架,例如 Berreby 等人。 (2017),通过收集上下文信息来模拟行动的结果,然后使用道义规范评估该结果的伦理考虑。其他作品如 Limarga 等。 (2020) 使用道义论进行自动化道德推理。同样,Tolmeijer 等人。 (2021) 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,以及它们与规则的契合程度作为机制来描绘道义论的实施。

Deontology 的其他用途包括在不兼容的公平性指标之间进行选择（Binns,2018）,或评估二元分类算法的分布（Leben,2020）。有些人还建议仅在特定情况下使用道义论:Dehghani 等。 (2008) 选择在具有“神圣价值”且结果之间没有数量级差异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。

困难。道义论存在许多困难,其中一些已经在计算机科学文献中进行了讨论。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,它们未能考虑到最可能的后果,因此基本逻辑无法充分捕捉

复杂的伦理见解（Abney,2011 年;Saltz 等人,2019 年）。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。

还有与应实施哪些规则有关的问题。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（Tolmeijer 等人,2021）。确定正确的细节级别对于确保机器的可解释性很重要。另一个问题是当规则之间发生冲突时。这可以通过对规则进行排序或权衡来进行修改,但是仍然必须确定重要性的顺序,并且还假设所有相关规则在使用之前就已确定。

4.2 平均主义

平等主义 人类在某种基本意义上是平等的,因此应该努力减少不平等（Binns,2018）。

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,因此应该努力避免和纠正某些形式的不平等（Binns,2018）。这有时意味着某些有价值的东西应该平均分配。平均主义的福利可能是快乐或偏好满足（Cohen,1989）、收入和资产（Rawls,1985;Dworkin,1981）,或者是做某些事情所必需的能力和资源（Sen,1992）。平等主义的每种货币的重要性在不同情况下可能会有所不同,并且应该考虑到不同的人可能对相同的结果或一组危害和收益有不同的评价（Binns,2018 年）。

实施平均主义。由于平等主义是道义论原则,它可以通过在自上而下、自下而上或混合架构中应用规则来实施。在应用这些规则时,需要对福利进行定义。平等主义在文学中有多种实现方式,例如 Murukannaiah 等人。(2020) 建议在满足利益相关者偏好方面最大化利益相关者之间的差异。在这个例子中,福利就是偏好满足。德沃克等人。 (2012) 将个人公平方面的平等主义视为原则,即任何两个在特定类别方面相似的个人都应该被相似地分类。

另一种方法侧重于权利的分配而不是个人的分类。
Leben (2020) 认为,这意味着应该赋予人口中的每个成员平等的权利（以及平等的份额）,但如果不可能在所有指标上实现对整个人口的平等,他们建议采用最小化距离的分配达到某种公平标准（例如人口规模）。

困难。平等主义可能并不关心不平等的事态本身,而是关心这种事态产生的方式（Binns,2018）。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”,不同的不可通约的公平逻辑可能在其中应用,并且在这些领域之间进行再分配可能不合适,存在着一个突出的争论。

4.2.1 平均主义:非恶意原则

Egalitarian Non-Maleficence 平等主义强加于伤害而非利益;最公平的结果是伤害平均分配（Leben，2020）。

非恶意原则将平等主义强加于危害而非利益（Leben,2020）。它也被称为补偿视图。因此它强调,最公平的结果是任何伤害在每个人（个人或群体)中平均分配。利益可能分配不均,有些人的利益比其他人多,这仍然是公平的。

实施非恶意行为。在所包含的文献中没有发现非恶意原则正在实施的例子。由于它是道义论原则,因此可以通过将原则作为规则应用来实施:平等主义应该在危害而非利益之间施加。这可以通过自下而上、自上而下或混合方法来完成。为了实施它,需要对福利进行定义,以了解什么是有益的,什么是有害的。例如,如果福利是偏好满足,那么收益就是偏好的满足,而危害就是偏好的不满足。

困难。该原则的一个问题是它允许任意大的结果不平等,并假设“富裕”和“贫困”之间存在可疑的区别（Leben，2020）。因此,根据这种批评,很难定义什么是伤害,什么是好处。然而,在 STS 的上下文中,这也许可以通过来自社会层的上下文输入来解决。然而,非伤害原则的另一个问题是,仅仅因为伤害是平均分配的并不一定意味着结果是公平的。如果利益分配极度不平等,直觉上仍然是不公平的;如果利益分配的不平等足够大,它可能会成为一种伤害。然而,为了应对这种情况,一旦它成为一种危害,该原则就强制要求它应该平均分配。因此,只有在利益变得有害的情况下,利益才会分配不均,这也许是公平的。

4.2.2 平均主义:机会均等

机会均等机会应以确保出生环境或随机选择不针对个人的方式平均分配（Friedler 等人,2021 年）。

机会均等的目标是确保不应因个人的出生环境或随机选择而导致负面属性对他们不利（Friedler 等人，2021）。然而,个人仍应对自己的行为负责。因此,个人的福祉应该独立于他们不相关的属性（Dwork 等人，2012 年）。

实施机会平等。机会均等是道义论原则,因此可以通过使用自上而下、自下而上或混合架构的规则应用来实施，并辅以福利的定义。为实施机会均等,Binns（2018 年)建议考虑每个群体是否同样有可能在给定该群体的实际基准率的情况下被预测为理想的结果。李等。 (2021) 认为这意味着所有机会都应该对所有人平等开放

申请人根据相关的优点定义。因此,这个例子中的福利就是获得机会。

困难。但是,从理论上讲,即使只有一小部分人具有获得机会的现实前景,这也可以完全满足（Lee 等人,2021 年）。

只要机会在理论上是可用的,与它是否在实践中可用无关。它也可能无法解决数据中可能已经存在的歧视问题。反对这一观点的另一个论点是,它在理论上允许这样一个社会,在这个社会中,只要一些成员开始时机会均等,他们最终就会陷入贫困（Fleurbaey,2008 年）。

4.2.3 平均主义:运气

运气均等 没有人会因为运气不好而变得更糟,人们应该根据自己的选择获得好处（Lee 等人,2021 年）。

运气均等主义源于机会均等,其目标被理解为消除非选择的不平等（Dworkin,1981）。因此,这意味着没有人应该因运气不好而最终变得更糟,而是应该根据自己的选择为人们提供差异化的经济利益（Lee 等人,2021 年）。

操作运气。以前没有在文学作品中运用平均主义运气的例子。由于它是道义论原则,因此可以通过在自上而下、自下而上或混合架构中应用规则,并辅以福利定义来实施。此外,该原则规定,任何人都不应因运气不佳而最终变得更糟;因此,运气水平也必须输入决策机制。

困难。这样做的一个问题是,通常很难将个人真正控制范围内的内容区分开来。人们需要找出在何种情况下以及在何种程度上,人们应该为他们发现自己所处的不平等地位负责（Binns,2018 年）。理想的解决方案将允许因人们的自由选择和知情冒险而导致的不平等,但忽略那些因运气不佳而产生的不平等。他们认为应该考虑选择和应得等概念的作用:做出的选择可能值得某些奖励和惩罚,但是如果不平等是个人无法控制的环境造成的,则应该予以纠正。因此,运气均等主义只负责创造优势和劣势,而不负责分配它们。然而,Binns 也指出,有时即使是选择造成的不平等也应该得到补偿,例如受抚养的看护人。

4.2.4 平均主义:自治

平等自治 要获得“真正的”平等,必须获得最低水平的自治,具有最低水平的选择多样性和质量,以及最低限度的决策能力（Fleurbaey,2008 年）。

自治平等已被提议为包括全方位的个人自由（Lee 等人,2021）。为了使他们成为“真正的”平等,必须达到最低水平的自主权,应提供最低水平的多样性和质量的选择,并且必须存在最低限度的决策能力（Fleurbaey,2008）。这是阿

可能是因为在其他人没有成本的情况下,希望在未来给人们更多的自由和更多的选择。

实施自治。在文献中没有发现该原则以前被实施过的例子。为了实施平等自治原则,应该通过自上而下、自下而上或混合架构以规则的形式应用（因为它是道义原则）。由于该原则表明应达到最低程度的自治,因此必须实施现有的自治程度,以便根据该原则公平分配。

困难。然而,当权力和信息严重不对称时,理性决策者的自主权无法作为道德目标（Fleurbaey,2008 年）。除此之外,该原则仅适用于个人公平,因为它明确地基于每个人的自由。与个人公平相关的弱点因此被自治原则继承,并让人怀疑它是否真的能实现公平的结果。

4.3 比例主义

比例主义 个人的权利应该根据他们对生产的贡献按比例调整（Leben,2020）。

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（Leben,2020）。这种分配应该根据进入生产过程的因素来管理,例如进入生产的每个人口成员的资源,进入这些资源部署的实际工作量,以及运气的多少进入那些资源。构思比例权利的方式通常分为两种截然不同的方法:自由主义和基于沙漠的方法。

实施比例主义。在文献中没有发现实施该原则的示例。这可以通过在自上而下、自下而上或混合架构中应用规则（因为它是 De 本体论原则)来完成。除了福利的定义外,贡献水平也应输入决策机制,因为原则规定权利应根据对生产的贡献进行调整。

困难。比例主义的一个问题是,在某些情况下,团体或个人可能没有为生产做出那么多贡献,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不应因此而获得较少的权利。但是,考虑到运气的影响可能会减轻这种情况。

4.3.1 比例主义:自由意志主义

自由比例主义 每个人都有权根据他们在同意时的总贡献获得成功率（Leben,2020）。

自由比例主义评估每个人在同意时的总贡献（Leben,2020）。它认为每个群体都有权获得至少与初始贡献一样公平的成功率。该范围内的不平等是公平的/可以接受的（因为它是亲

与原始不平等的比例），任何差异都是不公平的/不可接受的。因此，自由比例主义者只关心确保群体之间的不平等不超过目标特征中预先存在的不平等。自由意志主义理想被理解为主张每个人的自由的价值，只要不伤害任何其他人，这自然会延伸到所有权和资本权（Lee 等人，2021 年）。

实施自由主义。自由比例主义应该通过规则的应用来实施，因为它是道义论原则。这可以通过自上而下、自下而上或混合架构来完成。除了福利的定义外，还需要输入贡献水平，因为这是比例主义的一种形式。

文献中确定的另一种自由意志主义方法通过允许人们为自己定义“善”来应用该原则。对于 Etzioni 和 Etzioni（2016 年），Liber tarians 认为每个人都应该定义善和重要的价值观，国家应该保持中立。他们提出了一种“道德机器人”，它能够应用于各种不同的追求，并以类似于人工智能定向广告的方式解决道德选择问题，提供人与其他智能技术之间的接口。

困难。这种方法的困难在于获得同意的能力。这可能是一个模糊的问题，尤其是在群体公平的背景下。一群人可以在多大程度上同意某事尚不清楚，如果这不可能，那么就很难分析人们贡献了多少以及与此成比例的收益。即使可以明确获得同意，自由意志主义也存在困难，因为它没有针对可能仍然值得减轻的先前存在的不平等现象。

4.3.2 比例主义：基于沙漠

基于沙漠的比例主义 权利与个人努力成正比（Leben，2020）。

另一方面，基于沙漠的比例主义认为权利与个人努力成正比（Leben，2020）。这是因为人口中某种特征的先前流行（自由主义所基于的)可能是不公正环境的结果。一些文献将应得理解为与美德相对应（Kagan，1998）。

实施基于沙漠的比例主义。由于它是道义论原则，因此可以通过在自上而下、自下而上或混合架构中应用规则来实施基于沙漠的比例主义。福利的定义对于理解该原则在其应用中的目标是必要的。由于它是比例主义的一种形式，因此贡献也是必需的输入。根据这一原则，贡献是根据个人努力来定义的。 Dwork（2012)通过在评估沙漠的度量空间中为每个个体分配一些距离来实现基于沙漠的比例，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。

困难。该原则的一个弱点是不清楚什么是“不公正的情况”，因此很难评估哪些特征应该被减轻，哪些协调应得。

4.4 康德

康德的绝对命令意味着行动的理由应该与所有理性代理人都可以从事相同行动的假设一致（罗宾斯和华莱士,2007）,手段 - 目的原则表示将他人视为达到目的的手段是不道德的（阿布尼,2011 年）。

康德 (2011) 伦理学认为,伦理原则源自行动的逻辑结构,首先是将自由行动（行为人有理由的行为)与单纯行为区分开来（Kim 等人,2021）。康德的绝对命令意味着一个理性的代理人必须相信他们的行为理由与假设是一致的,即所有适用于这些理由的理性代理人都可以从事相同的行动（Robbins & Wallace, 2007; Abney, 2011; Kumar & Choudhury, 2022）).所有合法的道德义务都可以基于绝对命令（Wallach 等人,2008 年）。

从绝对命令派生的是手段-目的原则。这表示将他人视为达到目的的手段是不道德的（Abney,2011;Kumar & Choudhury，2022）。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。这是因为所有理性的存在都具有内在的道德价值,而非理性的世界只有工具价值。

可以说,这两项原则共同创造了一个理想世界,在这个世界中,社会可以根据人们的意志准则行事,而不会影响他人的福利（Limarga 等人,2020 年）。
实施康德道义论。这一原则可以通过在自上而下、自下而上或混合架构中应用规则（因为它是道义论的)来实施。

Limarga 等人在以前的文献中已经做到了这一点。 (2020) 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根据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。这可以通过 Berreby 等人（2017 年)的技术来实现,通过定义以下规则来实施手段 - 最终原则:如果一项行动涉及并影响至少一个人,但同时影响是不是行动的目的。如果一项行动导致至少涉及一个人的事件,但该事件不是行动的目的,则该行动是不允许的。允许任何其他操作。

困难。绝对命令的一个问题是它过于宽容。它通过允许任何可以具有普遍性准则的行动来潜在地允许直觉上的坏事（Abney，2011）。一个常见的例子是让谋杀进入你的房子,因为你不能撒谎说他们想要杀死的人不在那里。手段-目的原则也可能过于严格，从字面上解释,它禁止一个人在未经他们同意的情况下影响另一个人的任何行为。

5. 目的论伦理原则

本节检查审查中确定的每项目的论原则,包括有关它们以前如何实施的详细信息以及可能出现的困难。

5.1 美德伦理

美德伦理道德源于个人的内在性格,而不是个人行为的对错（Murukannaiah & Singh, 2020; Wallach et al., 2008; Brink, 2007; Kazim & Koshiyama, 2020; Robbins & Wallace, 2007 ; Yu 等人,2018 年;阿布尼,2011 年）。

美德伦理学认为伦理源于个人的内在品格,而不是个人行为的对错;重要的是一个人的道德品质（Murukannaiah & Singh，2020;Wallach 等,2008;Brink,2007;Kazim & Koshiyama,2020;Robbins & Wallace,2007;Yu 等,2018;Abney,2011）。正确的行为是由品德高尚的人做出的,因此在美德伦理中,一个人不应该问一个人应该做什么,而应该问一个人应该成为什么样的人（Anderson & Anderson, 2007; Rodriguez-Soto et al., 2022）).一个人拥有的品质应该是第一位的,行为是第二位的。美德被描述为以某种方式行事的倾向 (Abney, 2011)。道德美德可以通过习惯和实践来学习,这将美德理论置于文化所提倡的自上而下的明确价值观与个人通过实践发现或学习的自下而上的特质之间。美德伦理学不考虑特定的情况或行为,而是考虑个人生活中的所有行为,并检查这些行为是否共同构成了一个有美德的人的行为（Saltz 等人,2019）。

实施美德伦理。实施美德伦理取决于通过自上而下、自下而上或混合架构发展美德。以前的文献表明,美德的稳定性（如果一个人有美德,就不能像没有美德一样行事)意味着美德伦理可能是一种向机器灌输道德的有用方式（Wallach 等人,2008 年）).罗宾斯和华莱士 (2007) 认为,要实施这一原则,可以通过应用“美德”特征以道德方式解决问题。 Vanh´e 和 Borit (2022) 建议通过对系统设计师的教育来培养这一点。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。 (2021),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,而机制将是美德的实例化。

Govindarajulu 等人举例说明了这一点。 (2017) 通过使用计算形式逻辑将情绪形式化,表征特征,并建立学习特征的过程,以实例化美德。格林等人。 (2016) 认为,基于美德的系统必须了解在特定情况下需要采取一种行动而不是另一种行动的各种特征。因此,为了将美德伦理实施到机器中,需要培养美德特征。这已经在以前的文献中通过利用计算形式逻辑来建立良性特征来完成。

美德伦理也可以与其他方法结合使用;哈根多夫 (Hagendorff, 2020) 认为,道义论方法应该通过观察价值观和性格倾向,将其与美德伦理相结合,从而得到加强。

困难。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或考虑特定动机（Saltz 等人,2019 年）。

进一步的挑战与相互冲突的美德和美德的具体化有关（Tolmeijer 等人,2021 年）。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

是对实际美德的高度抽象。此外,随着时间和文化的变化,美德的概念会发生很大变化,因此,随着美德的变化,现在安装在机器中的那些可能会在未来导致不公平的结果。

5.2 结果论

结果主义 行动的道德取决于不平等对个人和群体的影响（Leben,2020）。

在结果主义方法中,社会正义取决于不平等对个人和群体的影响（Leben,2020）。结果主义是关于确定正确的行动,这些行动可以提升价值（Brink,2007 年;Yu 等人,2018 年;Cointe 等人,2016 年）。

因此,只能通过考虑其后果来判断行为的道德有效性（Saltz 等人,2019 年;Limarga 等人,2020 年;Rodriguez-Soto 等人,2022 年）。这样做的一个优势是,它可以用于评估具有复杂结果的决策,其中某些利益和某些利益受到损害。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,最好的结果是收益超过成本的结果（Sinnott-Armstrong,2021）。

操作结果主义。结果主义原则可以通过分析不同行动的后果来实施,通过自上而下、自下而上或混合架构实现。这在 Limarga 等人中得到了例证。 （2020）,谁实施原则通过考虑不同行动的后果做出适当的判断。

Suikkanen（2017 年)进一步采取了这种做法,他建议根据代理人的选择所产生的总价值来对他们的选择进行排名。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。 (2021) 认为结果主义实施的输入将是行动（及其后果）,决策标准将是相对幸福感,实现它的机制将是效用最大化（然而,这适用于特别是功利主义而不是整体的结果主义）。对于二元分类算法,Leben (2020) 建议可以通过查看如何根据相对社会成本为每个组的结果分配权重来实施结果论。

困难。然而,为每个群体的结果分配权重对于所有受保护群体来说可能是不现实的（Leben,2020）。计算成本可能很高,因为结果论系统需要一台机器来表示它可用的所有操作（Greene 等人,2016 年）。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（Etzioni & Etzioni,2017 年;Saltz 等人,2019 年）。结果主义之外还有道德约束,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（Suikkanen,2017 年）。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（享乐主义）、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)的清单。 （Brink,2007 年;Tolmeijer 等人,2021 年）。

5.2.1 结果论:功利主义

功利主义如果一项行为使所有受影响的人的总预期效用最大化,则该行为是合乎道德的（Kim 等人,2021 年;Lindner 等人,2019 年;Kazim 和 Koshiyama,2020 年;Wallach 等人,2008 年;Murukannaiah 和 Singh,2020 年） ;密尔,1863 年）。

功利主义是一种结果主义理论,它根据行为的后果来评估行为;一种行为是合乎道德的,当且仅当它最大化所有受影响者的总净预期效用（Kim 等人,2021 年;Lindner 等人,2019 年;Kazim 和 Koshiyama,2020 年;Wallach 等人,2008 年;Murukannaiah &辛格,2020 年;库马尔和乔杜里,2022 年;罗德里格斯-索托等人,2022 年）。最终的目的是尽可能远离痛苦并享受尽可能丰富的生活（Mill,1863）。因此,最大幸福原则指出,行为的正确与它们促进的幸福成正比,而错误与它们产生的幸福相反。效用不仅包括追求幸福,还包括预防或减轻不快乐。

实用化功利主义。由于功利主义是结果主义原则,因此可以通过分析行为的后果来实施。此外,该原则还需要对效用及其现有数量进行定义。 Leben (2020) 是以往文献中的一个例子,他建议实施功利主义来证明二元分类算法中公平性指标的设计选择是合理的。这可以通过构建一个函数来完成,该函数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,然后在聚合效用上运行选择程序以最大化总和。同样,Limarga 等人。 (2020) 将功利主义视为一种行为,当且仅当它使善最大化时才是正确的。为了实现这一点,他们为用于最终评估的每个动作分配了一些值（分配给其最坏后果的权重）。 Dehghani 等人。 (2008) 还通过选择具有最高效用的选项在他们的 MoralDM 方法中实施功利主义。 Choudhury 和 Kumar (2022) 建议,根据该理论,可以训练 AI 代理做出判断,从而为最多的人带来最大的幸福。

困难。功利主义的一个常见问题是,它可能导致一小部分用户为了更大的利益而受到不公平对待（Ajmeri 等人,2020 年）。还有人认为,不可能计算每一种替代行动方案的效用,并且该理论不能轻易解释权利和义务的概念或道德区别,例如,杀人与放任自流（Abney,2011 年）).功利主义的另一个公认问题涉及如何量化效用（例如“高等”和“低等”快乐）（Etzioni & Etzioni,2017）。为了缓解这些问题,功利主义或许可以被视为道德行为的额外必要条件,而不是唯一的道德原则（Kim 等人,2021 年）。

5.2.2功利主义:行为功利主义

行为功利主义 道德上正确的行为是在效用方面具有最佳整体效果的行为（Berreby 等人,2017 年;Anderson 等人, 2004 年）。

行为功利主义要求人们应该根据效用原则直接评估行为的道德性,该原则指出道德上正确的行为是具有最佳整体后果的行为（Berreby 等人,2017 年;Anderson 等人， 2004）。机器

利用行为功利主义的行为可以促使考虑可能导致更大净善后果的替代行动,并考虑这些行动中的每一个对所有受影响的人的影响。有人认为,几乎所有结果主义机器伦理的实施都利用行为功利主义（Tolmeijer 等人,2021 年）。

操作行为功利主义。这一原则的结果主义性质意味着它应该通过分析行动的后果来实施。由于它是功利主义的一种形式,它还需要效用的定义和效用的现有数量来确定最合适的解决方案。 Berreby 等人。 (2017) 通过确定领域内行为之间的偏好顺序来实施行为功利主义,然后声明如果存在另一个权重更大的行为,则该行为是不允许的。

困难。行为功利主义的一个常见批评是,一个人可以为了更大的利益而牺牲,而且它可能与正义的概念或人们应得的东西相冲突（安德森等人,2004 年）。这是因为行为的对错完全取决于其未来的后果,而人们应得的是过去行为的结果。

5.2.3功利主义:享乐行为功利主义

享乐行为功利主义道德上正确的行为是从所有替代行为中获得最大净快乐的行为（Anderson 等人,2004 年;Brink,2007 年）。

享乐行为功利主义需要计算最佳行动,从而产生最大的收益
从所有替代行为中获得净乐趣（Anderson 等人,2004 年;Brink,2007 年）。
实施享乐行为功利主义。作为结果主义原则,享乐行为功利主义可以通过分析行为的后果来实施。这个原则是享乐主义的,因此效用将根据快乐来定义,收集快乐水平对于获得解决方案至关重要。在以前的文献中,安德森等人。 (2004) 通过建议作为输入来实施该原则,该原则需要受影响的人数,以及每个人的每个可能行动将发生的愉快/不愉快的强度。对于每个人,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。

困难。这样做的困难在于快乐不一定能推断出公平。
可能会出现对某些人来说真的很愉快,但对其他人来说却非常不公平的情况　例如,为了取悦其他人而羞辱一个人或一个群体。

5.2.4功利主义:规则功利主义

规则功利主义道德上正确的行为是通过理解（一组)道德规则是否会导致最好的整体结果来评估的,假设所有或大多数代理人都遵循它（Berreby 等人,2017）。

规则功利主义涉及通过首先根据效用原则评估道德规则来对行为进行道德评估　假设所有或至少大多数代理人都遵循它,决定（一组)道德规则是否会导致最佳的总体结果（Berreby 等人,2017 年）。例如,一个这样的规则可能是“不偷”。第二步包括

根据第一步的合理性来评估特定的行动。仅当该行为受到维护功利原则的规则的认可时,该行为才被允许,无论该行为本身是否遵守功利原则。因此,与行动功利主义相比,问题不在于哪种行动产生最大效用,而在于哪种道德规则产生效用。

实施规则功利主义。该原则是结果主义的,因此可以通过分析行为的后果来实施,并且需要效用的定义才能做到这一点。 Berreby 等人。 (2017) 通过使用一个谓词来实现这一原则,该谓词将属于特定规则的动作的所有有效权重组合起来,然后通过谓词对这些权重求和。如果存在总体有害的规则实例,即考虑到其所有实例,其不良后果超过其良好后果的规则实例,则该行动被视为不允许的。

困难。规则功利主义的一个常见问题是,有时规则可能会导致非直觉的结果,因此应该被打破。这使得规则功利主义看起来更像是行动功利主义,其中正确的事情是通过每个行动的后果来评估的。

5.3 结果主义:优先主义/最大化

优先主义/最大化最小效用应该通过改善社会中最坏情况的体验来最大化（Ajmeri 等人,2020 年）。

Maximin 原则在于通过寻求改善社会中最坏情况的体验来最大化最小效用;保证每个人的最低效用高于最坏情况（Ajmeri 等人,2020 年）。它将重点转向改善最贫困人群的福祉（Lee 等人,2021 年）。差异原则指出,经济和社会不平等只有在使社会中最弱势的成员受益时才能被证明是合理的（Rawls,1967）。因此,优先主义/最大化侧重于改善最坏情况。

实施优先主义。优先主义是结果主义的,因此可以通过分析结果来实施。它还需要对效用进行定义,以了解如何分配效用。这一原则已在文献中得到实施,例如 Ajmeri 等人。 (2020) 谁提出了一个旨在改善任何用户的最低体验/最坏情况结果的技术代理。 Leben (2020) 认为,通过构建一个对每个潜在分布及其影响建模的函数,该原则可用于证明二元分类算法的公平性指标的设计选择是合理的。然后,应该对聚合效用运行一个选择程序,使最小值最大化。戴安娜等人。 (2021) 使用“minimax”框架提出算法,在该框架中,公平性是通过所有群体的最坏情况结果来衡量的,而不是群体结果之间的差异。因此,他们的目标是最小化所有群体的最大损失,而不是均衡群体损失,以确保情况最差的群体尽可能富裕。另一个例子是 Sun 等人。 (2021),他们通过在所有分配中最小化分配的最大成本,在家务分配问题中使用最大最小原则。

困难。一个问题是,虽然总效用可能会增加,但不一定会减轻歧视的影响（Sun 等人,2021 年）。它仍然允许群体之间存在差异。因此,最有特权的群体可能仍然存在

尽管整体体验得到改善,但他们比最不享有特权的群体享有更多特权。

5.3.1 结果论:无嫉妒

无嫉妒道德行为应该是没有代理人嫉妒另一个代理人的行为（Sun et al., 2021）。

在 Envy-Free 分配中,没有代理人嫉妒另一个代理人（Sun 等人,2021）。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,但只要代理人不互相嫉妒,这就被认为是公平的。

实施无嫉妒。这是一个后果主义原则,因此可以通过分析后果来实施。在这样做的过程中,必须输入嫉妒的程度。
Boehmer 和 Niedermeier (2021) 认为,如果没有代理人更喜欢另一个代理人的（资源)束而不是他们自己的,那么将资源分配给代理人是无嫉妒的。
困难。一个问题是认为重要的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（Lee 等人,2021 年）。此外,可能很难准确衡量嫉妒,因为它是利益相关者可能并不总是公开的主观实体。

另一个问题是,当要分配的项目不可分割时,无法保证存在 Envy-Free 分配,例如需要分配给多个代理的家务活（Sun et al., 2021）。这导致了 Envy-Freeness 的放松,例如最多一个项目的 Envy-Free（一个代理人可能嫉妒另一个,但是通过从嫉妒代理人的捆绑中删除一件杂务,可以消除嫉妒）,以及 Envy-Free最多任何项目（可以通过从嫉妒代理人的束中删除任何正成本的杂务来消除嫉妒）。

5.3.2 结果论:双重效应学说

双效主义 故意造成伤害是错误的,即使它会带来好处（邓,2015）。

双重效应理论 (DDE) 表明,故意造成伤害是错误的,即使它会带来好处（Deng,2015 年）。另一方面,如果不是故意造成伤害,而仅仅是做好事的结果,那么造成伤害可能是可以接受的。对于这一原则,如果行为本身在道德上是好的或中性的,则该行为是允许的,旨在产生一些积极的后果,没有消极后果是实现目标的手段,并且积极后果足以超过消极后果（Lindner 等人,2019 年）).它被解释为在正面和负面影响似乎不可避免的情况下允许采取同时具有正面和负面影响的行动（Govindarajulu & Bringsjord,2017）。

实施双重效应原则。这一原则可以通过分析行动的后果来实施。 Govindarajulu 和 Bringsjord (2017) 自动匹配 DDE,也是三重效应学说的更强版本,使用形式逻辑来操作原理。他们以两种不同的模式使用该框架:从头开始构建符合 DDE 的自治系统,或验证给定的 AI 系统是否符合 DDE。 Berreby 等人的另一种方法。（2017)通过以下方式实施这一原则

有规则禁止一个行为,如果它本质上是坏的,如果它导致一个坏的结果导致一个好的结果,如果它的整体效果是坏的。
困难。 DDE 的一个问题是,只要不是有意的,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。

5.3.3 结果论:不同的影响原则

不同的影响原则任何群体都应该在结果中具有平等或比例的代表性（Patel 等人,2020 年）。

建议将不同影响原则用作群体公平的概念（Patel 等人,2020 年）。它假定任何组在算法提供的解决方案中必须具有近似相等或按比例的表示。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（Binns,2018）。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。

实施不同的影响原则。这个原则可以通过分析行动的后果来实施,因为它是结果主义的。例如,通过使用上下文输入来评估导致解决方案中各组按比例或均等表示的操作。不同的虐待将确保被错误对待的人的存在是平等的或成比例的。

困难。这个问题是对群体公平的常见抱怨,它可能导致个人受到不公平对待,有利于群体。

5.3.4 结果论:不要伤害

不要伤害以任何身份造成伤害都是错误的（Lindner 等人,2019 年）。

该原则强制查看行为造成的伤害,声明没有伤害
应该受到惩罚。因此,任何造成伤害的行为都是不道德的。
实施不伤害。不伤害原则是结果主义的,因此可以通过分析行为的后果来实施。 Lindner (2019) 通过声明技术代理人不得执行会造成任何伤害的动作来实现这一点。丹尼斯等人。 (2016) 提出了一个特定行为违反“不伤害”的例子,即当飞机在地面上时向左移动十米,因此从道德上限制飞机在这种情况下执行此操作。

困难。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。

5.3.5 结果论:不做工具性伤害

不做工具性伤害 伤害是一种副作用,但不能作为实现目标的手段（Lindner 等人,2019 年）。

这个原则允许伤害作为副作用,但不是作为技术代理人的手段
目标（Lindner 等人,2019 年）。

实施无工具性伤害。该原则的实施方式与“不伤害”原则大致相同,只是它允许作为副作用造成伤害。

困难。这可能有助于造成伤害不可避免的情况,尽管它仍然允许某些伤害是可以接受的,这可能导致某些群体或个人受到不公平对待。

5.4 其他原则

除了此处列出的原则外,文献中还提到了其他原则。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。

利己主义是为自己取得最大可能的结果,而不考虑他人（Robbins & Wallace, 2007; Kumar & Choudhury, 2022）。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。

多元主义指出没有一种方法是最好的（Robbins & Wallace,2007）。
使用上下文和各种推理技术来选择原则是合适的。 Tolmeijer 等人。 (2021) 还提倡根据这种方法进行进一步研究,应用多理论模型,机器可以根据情况的类型互换应用不同的理论。他们认为人类道德是复杂的,不能被单一的经典伦理理论所涵盖。该原则未包含在此分类中,因为它本身不是原则,因为它不指导特定的行动过程。

然而,这是一种对道德有帮助的方法,并且可能与帮助开发人员了解如何利用道德原则有关。

特殊主义强调没有规范价值的独特来源,也没有单一的、普遍适用的道德评估程序（Tolmeijer 等人,2021）。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,某个特征是否与道德相关以及它扮演什么角色将对情境的其他特征敏感。作者认为,特殊主义的输入将包括情况（背景、特征、意图和后果）,并且决策标准取决于经验法则和先例,因为所有情况都是独一无二的。决定一项行动的机制取决于它在多大程度上符合规则或先例。他们发现的一些挑战是没有唯一和普遍的逻辑,因此每种情况都需要进行独特的评估。缺乏通用逻辑是未将其纳入分类法的部分原因: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。

关怀和责任的伦理与考虑你与他人的相互联系的感觉有关（罗宾斯和华莱士,2007 年;马丁,2022 年）。为了道德,一个人应该考虑

这些人和你所处的情况。利用你的经验,你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,因为它表示某种行为。它可能与康德的手段-目的原则有关,然而这并没有作为一个单独的原则被包括在分类学中。

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（Hagerty & Rubinov,2019）。他们认为,跨文化存在多种道德框架,这些框架内存在显着差异。对于作者来说,道德和文化是密不可分的,要理解其中一个,您必须先看看另一个。因此,他们认为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。

6. 结论和方向

为了更好地解决人工智能对公平的追求,研究必须以人为本（Dignum & Dignum,2020）。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（Chopra 和 Singh,2018 年）。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（Murukannaiah 等人,2020 年）。为了以令人满意的方式解决这些困境,从而促进更高的公平目标,道德原则可以帮助确定行为的道德允许性（McLaren,2003 年;Lindner 等人,2019 年）。

6.1 要点和方向

根据我们的审查,我们确定了从业者为实施道德原则应了解的关键要点。我们设想,我们开发的分类法和我们确定的关键要点将有助于在 STS 治理中实施道德原则。图 3 总结了要点。

阐明架构。在道德原则的实施中,设计师必须决定是否通过自上而下、自下而上或混合架构方法来实施原则（Wallach 等人,2008 年）。

· 方向。该领域的机会在于进一步研究自下而上的方法,以及辨别每种不同架构适用的环境的正式方法。

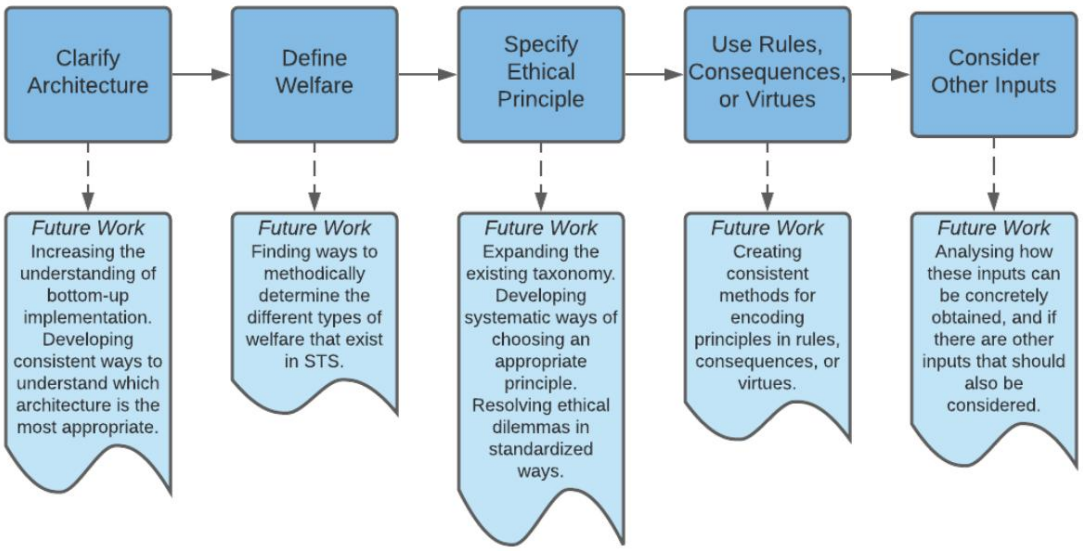

图 3:在 STS 中实施伦理原则的关键要点和未来研究

定义福利。为了理解什么构成令人满意的结果,有必要明确福利的定义以指定什么是好的或有价值的,这在伦理原则的实施中是必要的（Fleurbaey,2008）。

· 方向。为了正确理解福利应该如何定义,需要对与 STS 相关的不同类型的福利以及如何有条不紊地确定福利进行更多的工作。

指定道德原则。研究人员应该具体说明他们在研究中使用了哪些伦理原则（Binns,2018 年）。从 AI 和计算机科学文献中发现了广泛的伦理原则,分类树上有 25 个节点（图 2）。可能存在更多,但在本次审查中未发现。

清楚正在使用的原则将有助于设计人员进一步指定其应用所需的输入,这反过来将提高道德推理能力和决策制定的可解释性（Leben,2020）。除此之外,研究人员应该促进使用较少利用的原则,以避免垄断诸如功利主义等众所周知的原则。

· 方向。因此,这为纳入本次审查中未确定的原则（尤其是西方学说之外的原则)提供了机会,并强调未来的工作明确纳入伦理原则。

· 方向。除此之外,开发一致的方法来选择合适的道德原则将对从业者有益。这是因为规范伦理理论的抽象意味着它们的适用性可能取决于多种因素,包括个人偏好、规范、价值观和背景。方法

因此,有助于确定适用于特定情况的道德原则的学说将很有用。

规则、后果或美德。实施已确定的道德原则有三个关键方向。第一种是使行动符合某些规则（Greene 等人，2016 年）。
第二种是根据行动产生的（潜在)后果选择行动,以及根据所选原则哪种后果最好（Suikkanen,2017 年）。
第三是通过机器的美德实例化;因此,正确的行动是由良性机器产生的（Govindarajulu 等人,2019 年）。

· 方向。展望未来,在规则、后果或美德中对原则进行编码将需要进行大量的进一步研究,以在治理 STS 的背景下开发系统的方法。

原则相关输入。对于输入是否必要,原则与原则之间存在不同的输入选择。这包括对运气、自主性、贡献、美德、效用和嫉妒的考虑。

· 方向。这些输入可能会有些抽象,并且可能难以推断。因此,需要进一步研究如何获得它们,以及是否有更多此处未确定的输入。

参考

Abney, K. (2011)。机器人、伦理理论和元伦理学:困惑者指南,pp. 35-52。麻省理工学院出版社,剑桥。

Ajmeri, N.、Guo, H.、Murukannaiah, PK 和 Singh, MP (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,第 16-24 页,奥克兰。 IFAAMAS。

安德森, M., & 安德森, SL (2007)。机器伦理:创造一个有道德的智能代理。人工智能杂志,28（4）,15。

安德森, M., & 安德森, SL (2014)。 GenEth:一个通用的道德困境分析器。在全国人工智能会议论文集中，卷。 1,第 253–261 页,魁北克。人工智能促进协会。

Anderson, M.、Anderson, SL 和 Armen, C. (2004)。迈向机器伦理。在 AAAI-04 代理组织研讨会:理论与实践,第 1-7 页,圣何塞。美国汽车协会。

Berreby, F.、Bourgne, G. 和 Ganascia, J.-G。 (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,第 96-104 页,圣保罗。

自治代理和多代理系统国际基金会。

R. 宾斯 (2018)。机器学习中的公平性:政治哲学的教训。在 Friedler, S., & Wilson, C.（编辑）,第一届公平、问责制和透明度会议论文集,卷。 81 机器学习研究论文集,第 149-159 页,纽约。 PMLR。

Bishr, ABBB (2018)。人工智能伦理原则和指南。聪明的迪拜。

Boehmer, N., & Niedermeier, R. (2021)。扩大计算社会选择的研究议程:多个偏好配置文件和多个解决方案。在第 20 届自治代理和多代理系统 (AAMAS) 国际会议论文集中,第 1-5 页,伦敦。自治代理和多代理系统国际基金会。

D. 布林克 (2007)。结果主义的一些形式和局限性。牛津手册伦理理论,1 (1),381–423。

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。

Burton, E.、Goldsmith, J.、Koenig, S.、Kuipers, B.、Mattei, N. 和 Walsh, T. (2017)。道德的人工智能课程中的注意事项。人工智能杂志,38 (2), 22–34。

C. 坎卡 (2020)。实施人工智能伦理原则。 ACM 通讯,63 (12), 18–21。

Cheng, L.、Varshney, K. 和 Liu, H. (2021)。对社会负责的人工智能算法:问题,目的和挑战。 JAIR,71,1137–1181。

乔普拉 (Chopra, A.) 和辛格 (Singh, M.)（2018 年）。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,pp. 48–53,新奥尔良。计算机协会。

乔治亚州科恩 (1989)。关于平等主义正义的货币。伦理学,99 (4), 906–944。

Cointe, N.、Bonnet, G. 和 Boissier, O. (2016)。多代理系统中代理行为的伦理判断。在 2016 年自主代理和多代理系统国际会议记录中,第 1106–1114 页,新加坡。 IFAAMAS。

Conitzer, V.、Sinnott-Armstrong, W.、Borg, JS、Deng, Y. 和 Kramer, M.（2017 年）。人工智能的道德决策框架。在第 31 届 AAAI 人工智能会议 (AAAI) 会议记录中,第 4831–4835 页,火奴鲁鲁。美国汽车协会。

Cummiskey, D. (1990)。康德后果论。伦理学,100 (3), 586–615。

Dastani, M., & Yazdanpanah, V. (2022)。人工智能系统的责任。人工智能与社会,1 (1435-5655)。

Dechesne, F.、Di Tosto, G.、Dignum, V. 和 Dignum, F. (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),79–107。

Dehghani, M.、Tomai, E. 和 Klenk, M. (2008)。道德综合推理方法决策。机器伦理,3,1280–1286。

邓 B.（2015 年）。机器伦理:机器人的困境。自然,523,24-26。

Dennis, L.、Fisher, M.、Slavkovik, M. 和 Webster, M.（2016 年）。自治系统中道德选择的形式验证。机器人和自治系统,77,1-14。

Diana, E.、Gill, W.、Kearns, M.、Kenthapadi, K. 和 Roth, A. (2021)。用于（松弛的)极小极大公平性的收敛算法。 CoRR,abs/2011.03108,1-22。

Dignum, V. (2019)。道德决策,第 35-46 页。斯普林格,查姆。

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., G´enova, G., Haim, G., Kließ, MS, Lopez-Sanchez, M., Micalizio, R., Pav´on, J., Slavkovik, M., Smakman, M., van Steenbergen, M., Tedeschi, S., van der Toree, L., Villata, S., & de怀尔特 (2018)。设计伦理:必然还是诅咒?在 2018 年 AAAI/ACM 人工智能、伦理和社会会议论文集中,AIES 18,第 60––66 页,美国纽约州纽约市。计算机协会。

Dignum, V. 和 Dignum, F. (2020)。特工死了。特工万岁!在第 19 届自治代理和多代理系统 (AA MAS) 国际会议记录中,第 1701–1705 页,奥克兰。自治代理和多代理系统国际基金会 (AAMAS)。

Dwork, C.、Hardt, M.、Pitassi, T.、Reingold, O. 和 Zemel, R. (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,第 214–226 页,剑桥。美国计算机协会。

德沃金,R. (1981)。什么是平等?第 1 部分:福利平等。哲学与公共事务,10 (3),185–246。

Etzioni, A., & Etzioni, O. (2016)。人工智能辅助道德。道德与信息技术,
18, 149–156。

Etzioni, A., & Etzioni, O. (2017)。将伦理融入人工智能。伦理学杂志,21, 403–418。

Fjeld, J.、Achten, N.、Hilligoss, H.、Nagy, A. 和 Srikumar, M. (2020)。有原则的人工智能:将基于道德和权利的方法中的共识映射到 AI 原则。伯克曼克莱因中心研究出版物第 2020-1 号,第 1-39 页,剑桥。

伯克曼克莱因中心。

Fleurbaey, M. (2008)。公平、责任和福利。牛津大学出版社,
牛津。

Floridi, L. 和 Cowls, J. (2019)。人工智能在社会中的五项原则的统一框架。
哈佛数据科学评论,1 (1),1。https://hdsr.mitpress.mit.edu/pub/l0jsh9d1。

Friedler, SA、Scheidegger, C. 和 Venkatasubramanian, S. (2021)。公平的（不)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),136–143。

Govindarajulu, NS, & Bringsjord, S. (2017)。关于使双重效果学说自动化。在第 26 届国际人工智能联合会议论文集中,IJCAI-17,第 4722-4730 页,墨尔本。 IJCAI。

Govindarajulu, NS、Bringsjord, S.、Ghosh, R. 和 Sarathy, V.（2019 年）。走向良性机器的工程。在 2019 年 AAAI/ACM 人工智能、伦理和社会会议论文集中,AIES 19,第 29 -35 页,美国檀香山。计算机协会。

Greene, J.、Rossi, F.、Tasioulas, J.、Venable, KB 和 Williams, B.（2016 年）。将道德原则嵌入集体决策支持系统。在第 13 届 AAAI 人工智能会议 (AAAI) 会议记录中,第 4147–4151 页,Snowbird。美国汽车协会

按。

哈根多夫 (2020)。人工智能伦理的伦理:指南的评估。思维与机器,30, 99–120。

Hagerty, A. 和 Rubinov, I.（2019 年）。全球人工智能伦理:对社会影响和影响的回顾
人工智能的伦理影响。 CoRR,abs/1907.07892,1-27。

Honarvar, AR, & Ghasem-Aghaee, N. (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 机器人
与自动化计算智能国际研讨会 - (CIRA),第 290-295 页,大田。

IEEE。

Jobin, A.、Ienca, M. 和 Vayena, E. (2019)。人工智能伦理准则的全球格局。
自然机器智能,1 (9),389–399。

Kagan, S. (1998)。平等与沙漠,第一章。 30,第 298-314 页。牛津大学出版社,
牛津。

康德 I.（2011 年）。伊曼纽尔康德:道德形而上学的基础:德文英文版。剑桥康德德英版。剑桥大学出版社,剑桥。

Kazim, E., & Koshiyama, A. (2020)。人工智能伦理的高级概述。 SSRN,1 (1),1-18。

Khan, AA、Badshah, S.、Liang, P.、Khan, B.、Waseem, M.、Niazi, M. 和 Akbar, MA
(2021)。人工智能伦理:对原则和挑战的系统文献回顾。
CoRR,abs/2109.07906。

Kim, TW、Hooker, J. 和 Donaldson, T. (2021)。认真对待原则:混合体
价值对齐的方法。 JAIR,70、871–890。

Kitchenham, B., & Charters, S. (2007)。在软件工程中执行系统文献综述的指南。技术。代表,基尔大学和达勒姆大学联合
报告。

K¨okciyan, N., Yaglikci, N., & Yolum, P. (2017)。一种解决在线社交网络中隐私纠纷的论证方法。 ACM 跨。互联网技术,
17 (3),
1-22。

K¨okciyan, N., & Yolum, P. (2020)。 Turp:管理信任以规范互联网隐私
东西的。 IEEE 互联网计算,24 (6), 9–16。

Kumar, S. 和 Choudhury, S. (2022)。规范伦理、人权和人工智能
根。人工智能与伦理学,2, 1–10。

莱本 D.（2020 年）。评估机器学习公平性的规范原则。在 AAAI/ACM 人工智能、伦理和社会会议 (AIES) 会议记录中,第 12
页。
86–92,纽约。计算机协会。

Lee, M.、Floridi, L. 和 Singh, J. (2021)。形式化算法公平性之外的权衡:伦理哲学和福利经济学的教训。人工智能与伦理,1
(1),529–544。

Liao, B.、Slavkovik, M. 和 van der Torre, L. (2019)。构建 Jiminy Cricket:利益相关者之间道德协议的架构。在 AAAI/
ACM 人工智能、伦理和社会会议 (AIES) 会议记录中,第 147-153 页,火奴鲁鲁。美国计算机协会。

Limarga, R.、Pagnucco, M.、Song, Y. 和 Nayak, A. (2020)。具有情境演算的机器伦理的非单调推理。在 AI 2020：人工智能的进展中，第 203-215 页，堪培拉。施普林格国际出版社。

Lin, P.、Abney, K. 和 Bekey, G. (2011)。机器人伦理：描绘机械化世界的问题。人工智能，175 (5)，942–949。特别审查问题。

Lindner, F.、Mattmüller, R. 和 Nebel, B.（2019 年）。行动计划的道德许可。AAAI 人工智能会议论文集，33 (01)，7635–7642。

Manjarr´es, A.、Fern´adez-Aller, C.、L´opez-S´anchez, M.、Rodr´ guez-Aguilar, JA 和 Casta ner, MS (2021)。人工智能创造公平、公正、公平的世界。IEEE 技术与社会杂志，40 (1)，19–24。

K. 马丁（2022 年）。作为 AI 道德基础的护理伦理，第 1-6 页。奥尔巴赫公众号系统。

麦克拉伦，BM（2003 年）。扩展定义道德原则和案例：人工智能模型。人工智能，150 (1)，145–181。人工智能和法律。

Mehrabi, N.、Morstatter, F.、Saxena, N.、Lerman, K. 和 Galstyan, A. (2021)。机器学习中的偏见和公平性调查。 ACM 计算调查，54 (6), 1–35。

穆勒，JS（1863 年）。功利主义。朗文斯、格林公司。

Morley, J.、Elhalal, A.、Garcia, F.、Kinsey, L.、M¨okander, J. 和 Floridi, L. (2021)。伦理即服务：人工智能伦理的务实运作。思维与机器，第 31 页，第 239–256 页。

Morris-Martin, A.、De Vos, M. 和 Padget, J. (2019)。多代理系统中的规范涌现：一篇观点论文。自治代理和多代理系统 (JAAMAS)，33 (6)，706–749。

Murukannaiah, PK、Ajmeri, N.、Jonker, CM 和 Singh, MP (2020)。道德多代理系统的新基础。在第 19 届国际自主代理和多代理系统会议 (AAMAS) 会议记录中，第 1706–1710 页，奥克兰。

IFAAMAS。蓝天创意轨道。

PK 穆鲁卡奈亚和国会议员辛格 (2020)。从机器伦理到互联网伦理：拓宽视野。 IEEE 互联网计算，24 (3)，51–57。

Patel, D.、Khan, A. 和 Louis, A. (2020)。背包问题的组公平性。 CoRR，abs/2006.07832，1-36。

Pedamkar, P. (2021)。人工智能中的智能代理。人工智能 https://www.educba.com/中的智能代理/。访问时间：2021-12-21。

R adulescu, R.、Mannion, P.、Roijers, DM 和 Now´e, A.（2019 年）。多目标多代理决策：基于效用的分析和调查。 CoRR，abs/1909.02964，1-48。

罗尔斯，J.（1958 年）。正义即公平。哲学评论，67（2），164-194。

罗尔斯，J.（1967 年）。分配正义。哲学、政治与社会，1， 58–82。

罗尔斯，J.（1985 年）。作为公平的正义：政治而非形而上学。哲学与公共事务，14 (3), 223–251。

Robbins, R., & Wallace, W. (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),1571–1587。特刊集群。

Rodriguez-Soto, M.、Serramia, M.、Lopez-Sanchez, M. 和 Rodriguez-Aguilar, JA (2022)。通过多目标强化学习灌输道德价值观。伦理与信息技术,24 (1), 9。

Saltz, J.、Skirpan, M.、Fiesler, C.、Gorelick, M.、Yeh, T.、Heckman, R.、Dewar, N. 和 Beard, N.（2019 年）。将道德融入机器学习课程。 ACM 跨。电脑。教育,19 (4), 1–26。

施瓦茨, SH (2012)。施瓦茨基本价值观理论概述。在线阅读在心理学和文化中,2 (1),2307–0919。

A. 森 (1992)。重新审视不平等。克拉伦登出版社,牛津。

国会议员辛格 (2013)。规范作为管理社会技术系统的基础。 ACM 智能系统和技术交易 (TIST),5 (1),21:1–21:23。

W. 辛诺特-阿姆斯特朗 (2021)。结果主义。在 Zalta,EN（主编）,斯坦福哲学百科全书（2021 年秋季版）。斯坦福大学形而上学研究实验室,斯坦福大学。

亚非学院（2021 年）。第 1 单元介绍伦理学。 https://www.soas.ac.uk/cedep-demos/000_P563_EED_K3736-Demo/unit1/page_17.htm#。访问时间:2021-09-23。

Suikkanen, J. (2017)。结果论、约束和良好关系:对马克施罗德的答复。伦理学与社会哲学杂志,3 (1), 1–9。

Sun, A.、Chen, B. 和 Doan, XV (2021)。分配不可分割的家务的公平标准和效率之间的联系。 CoRR,abs/2101.07435,1-32。

Tolmeijer, S.、Kneer, M.、Sarasua, C.、Christen, M. 和 Bernstein, A. (2021)。实现机器伦理学:一项调查。 ACM 计算机。生存,53 (6)。

Vanh´ee, L., & Borit, M. (2022)。观点:设计师的道德 - 如何培养人工智能的道德设计师。 JAIR,73,619–631。

Wallach, W.、Allen, C. 和 Smit, I. (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),565–582。

Woodgate, J., & Ajmeri, N. (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) 国际会议记录中,第 1824–1828 页,在线。 IFAAMAS。蓝天的想法追踪。

Yazdanpanah, V.、Gerding, E.、Stein, S.、Dastani, M.、Jonker, CM 和 Norman, T. (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)中,第 57-62 页。

Yu, H.、Shen, Z.、Miao, C.、Leung, C.、Lesser, VR 和 Yang, Q. (2018)。将伦理道德融入人工智能。在第 27 届国际人工智能联合会议论文集中,IJCAI,pp. 5527–5533,斯德哥尔摩。 IJCAI。

Zhu, L.、Xu, X.、Lu, Q.、Governatori, G. 和 Whittle, J. (2022)。 AI 和道德规范 实施负责任的 AI,第 15-33 页。斯普林格国际出版社,Cham。

附录 A. 方法概述

图 4 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。对著作进行定性分析,通过通读和总结要点,然后将其归类为相关的原理和研究类型（见表1和表2）。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,Leben,2020）。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,Sun 等人,2021）。一些作品混合了识别和实现（例如,Kim 等人,2021）。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。

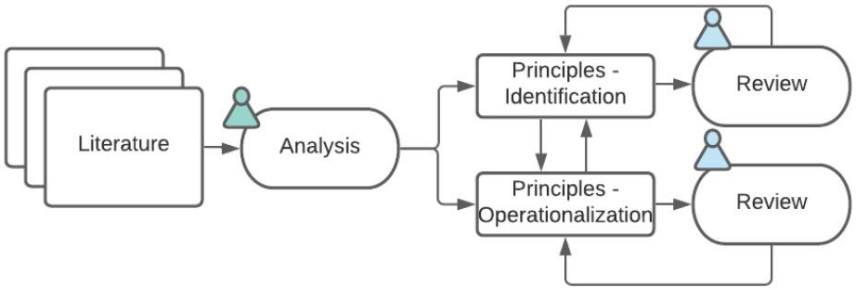

图 4:从 Lit 中提取原理识别和操作化的方法
时代

附录 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。这意味着可能会遗漏其他语言的相关研究,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。初始搜索字符串基于

初步研究,随着审查的继续,确定了更多关键术语。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。

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 (2007) 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而不是客观上影响结果。通过让二级审稿人批判性地分析结果并提出建议以帮助主要审稿人改进研究,可以减轻这种情况。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 3: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	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