（不）关注点分离：关于智能体伦理的哲学思考及其对软件工程从业者的启示

Marc Cheong*
墨尔本大学
澳大利亚帕克维尔
marc.cheong@unimelb.edu.au

Sebastian Baltes
海德堡鲁普莱希特-卡尔大学¨
德国海德堡
sebastian.baltes@uni-heidelberg.de

Joanne Byrne
墨尔本大学
澳大利亚帕克维尔
joanne.byrne@unimelb.edu.au

Haoyu Gao Raula Gaikovina Kula Sarita Rosenstock墨尔本大学 大阪大学 墨尔本大学澳大利亚帕克维尔 日本大阪 澳大利亚帕克维尔 haoyug1@student.unimelb.edu.au raula-k@ist.osaka-u.ac.jp sarita.rosenstock@unimelb.edu.au

Christoph Treude新加坡
管理大学新加坡
ctreude@smu.edu.sg

Mansooreh Zahedi墨尔本大学澳大利亚帕克维尔
ma
nsooreh.zahedi@unimelb.edu.au

arXiv:2611.31927v1 [cs.SE] 24 Nov 2026

摘要——**人工智能智能体在软件工程（软件工程）中的使用已十分普遍，其自主性程度各异，具备多种能力，并能够执行各种任务。哲学家已经开始考察此类智能体的伦理问题，而软件工程研究人员或软件工程从业者可以从中受益；另一方面，软件工程从业者目前主要通过原则主义的视角探索其中大多数关切。我们通过评述 PhilPapers 收录的关于人工智能智能体的哲学文献，将其中的伦理关切转化为面向软件工程研究人员的研究议程，以及对使用编码智能体的团队的启示，从而弥合这一差距。我们对 PhilPapers——这一重要的哲学索引和开放获取档案库——开展了系统映射研究，得到自 2023 年以来由职业哲学家发表的 33 篇经过同行评审的论文。我们对这些论文进行了主题编码，并通过跨学科评审分析新兴主题，最终得到八个主题和 85 个编码。在此基础上，我们为软件工程提出了七个研究问题，涵盖开发者对智能体的依赖、对智能体编写代码的责任、智能体配置中的伦理约束以及基准测试设计。对于从业者，我们提出了五点启示：明确哪些选择必须由智能体提交给人类决定；将智能体编写的拉取请求控制在足够小、便于审查的范围内；将配置文件指令与检查机制配套；留意测试的变更；并记录谁批准了什么。我们认为，伦理关切无法与软件工程关切分离，并单独归入某个模块：它们会在代码审查、配置、评估中出现，**

以及需求工作，而解决这些问题需要跨学科协作。

索引词—**伦理学、人工智能智能体、智能体软件工程、社会影响、哲学**

本文已提交 IEEE，供其考虑发表。版权可能在不另行通知的情况下转让，届时本文版本可能不再可访问。

一、引言

在软件工程（软件工程）中，人工智能智能体正从仅仅提供代码建议，转变为自主执行软件工程工作流：从编写、合并和部署代码，到维护文档，再到对问题进行分类处理以及审计代码库 [1], [2]。人工智能智能体的社会伦理影响（即智能体人工智能范式）正开始受到软件工程社区的关注 [3], [4], , 尤其是在 AGENT研讨会系列中 [5], [6]。

这些智能体在完成任务时所作出的许多选择，实际上都是伦理选择，尽管它们看起来像是实现细节。一个添加功能的智能体还会决定该功能保留用户数据的时长、是否将个人数据写入日志，以及新界面是否能够与屏幕阅读器配合使用。当智能体打开拉取请求而开发者批准该请求时，可能没有人曾有意识地做出这些选择。

社会科学家和哲学家已开始在数字伦理和人工智能伦理的大背景下，深入剖析有关人工智能智能体的问题[1]。旨在将有关人工智能智能体的哲学问题转化为软件工程术语的跨领域工作——汇集哲学家、社会科学家和软件工程从业者——十分罕见。

正是秉持弥合学科鸿沟的精神，我们撰写了这篇论文，旨在回顾有关人工智能智能体的现有哲学文献，并将其中的伦理关切转化为软件工程的研究议程。我们提出了两个核心研究问题：

*MC 是通讯作者。其他作者按字母顺序列出。

[1]例如，参见 https://philpapers.org/browse/ethics-of-artificial-intelligence。

- **研究问题1**：哲学文献提出了哪些关于人工智能智能体的伦理关切?
- **研究问题2**：鉴于智能体日益广泛地被采用和普及，这些关切如何对应于软件工程研究与实践?

为回答这些**研究问题**，我们的贡献包括一幅伦理关切主题图，该图源自对 PhilPapers 的系统性调查。PhilPapers 是“由哲学家社群维护的哲学综合索引和书目”[7]（**研究问题1**）；此外，我们还提出了由七个问题组成的研究议程，以及对软件工程实践的五项启示（**研究问题2**）。这两项成果均由职业生涯阶段各异的作者团队共同完成。本文结构如下：第二节初步概述相关文献；第三节介绍研究方法；第四节呈现针对**研究问题1** 和**研究问题2**的研究结果；第五节讨论研究结果、阐述对软件工程实践的启示并列出效度威胁；第六节总结全文。

II. 背景与相关工作

A. 智能体人工智能与能动性概念

借用 AGENT 2027研讨会的说法，软件工程中的智能代理式人工智能系统“展现出目标导向自主性、推理能力、工具使用、交互能力和持续演化”[2]。这与人们普遍对人工智能智能体（“执行特定任务的自主软件程序” [8]）以及整体智能体人工智能（“多个人工智能智能体协作以实现复杂目标的系统” [8]）的理解相符。然而，我们需要指出，这一术语“值得进一步明确” [9]：能动性 在哲学和伦理学中的含义截然不同。简而言之，在后一语境中，能动性 “.....指这种行动能力 [付诸行动]的行使或体现” [10],；哲学家关注的问题包括“从意向性角度理解行动..... [以及]由智能体的心理状态和事件引发的因果关系” [10]。

因此，本文对哲学文献的评述旨在涵盖这两方面的考量。第一方面，正如 Mussgnug 恰当地指出的那样 [11] ——“作为当前公共话语中随处可见的一个技术术语，而非 [表明] 任何关于道德或认识论能动性的哲学概念”——以及第二方面，即纳入道德/认识论视角的考量。事实证明，这构成了制定我们的研究议程（第四节-B）以及阐述对软件工程从业者影响（第五节-A）的重要组成部分。

B. 智能体软件工程中的伦理学、治理与责任

人为因素，包括人工智能使用的伦理学/治理/社会影响，已作为框架、原则和指南的一部分得到概括 [12], [13]。就软件工程中的协作而言，Treude 和 Gerosa [14]对人类与人工智能之间的互动进行了分类。

具体而言，对于智能体软件工程，Hoda [5] 提出了 CRAFT（全面、负责任、适应性强、基础性、转化性）价值观，用于智能体软件工程，“将代理设计的关注重点从编码扩展到人类与人工智能的协作”（引自[1]）。Aleti 及其同事从多个维度着手解决“人工智能软件工程师的可信度”问题，包括“技术质量、透明度与问责、认知谦逊，以及社会与伦理对齐” [3]。Treude 先前的研究也探讨了人工智能编码工具提供商与软件开发者之间的“所有权、责任、法律责任和披露义务” [15]。

对于软件工程中的智能体而言，仓库级上下文文件（尤其是 AGENTS.md）已被确认为伦理指令和约束的潜在载体 [1], [4]。了解伦理考量如何被编入这些工件之中十分重要，因为这些工件应当（理论上）指导智能体在软件工程过程中的行为。例如，Gao 及其同事 [4] 研究了 AGENTS.md 及相关工件，以查找与澳大利亚人工智能伦理原则一致的伦理声明出现情况。

C. 超越当前基于原则的方法

我们的工作与现有研究的不同之处在于，我们超越了原则主义。具体而言，当前的方法主要借鉴应用人工智能伦理中确立的原则。借用 Cheong 和 Coghlan [16] 的说法，原则主义“在数字伦理学中 [已].....备受关注，它[.....试图提炼道德理论与哲学的见解] ..... [以构建一个指导性框架，使不同的] .....理论家至少在一定程度上能够达成共识”。人工智能伦理中的原则主义，其灵感可追溯至应用医学伦理 [17], ；它曾帮助医疗专业人员作出伦理判断和选择。

尽管软件工程从业者通常熟悉伦理原则 [18], [19], ，但他们在应用这些原则时面临着“普遍性、技术相关以及与人相关的挑战”（引自 [4]）。这在应用伦理学研究中并不罕见：原则之间可能相互冲突，使人陷入两难，不知应以哪项原则为先[16]。此外，聚焦于多元原则 [13] 会忽视其他伦理哲学传统，例如功利主义伦理学和义务伦理学，仅举数例。因此，本文旨在弥合翻译鸿沟（哲学与软件工程之间的鸿沟），以及框架／理论鸿沟（超越原则主义）。

III. 方法论

在本节中，我们概述所采用的方法：搜索 PhilPapers 并审阅相关论文、定性编码以及对主题进行细化的过程。

A. PhilPapers 综述

PhilPapers [7] 是一个全面的在线学术哲学研究索引和哲学论文开放获取档案库，大致相当于 arXiv 与 DBLP 计算机科学书目两者的结合。目前，它在 6k+ 个类别中索引了超过 3.1M+ 条目，并通过其下属平台托管了超过 150k 篇预印本/手稿/开放获取论文

[2]https://conf.researchr.org/home/icse-2027/agent-2027

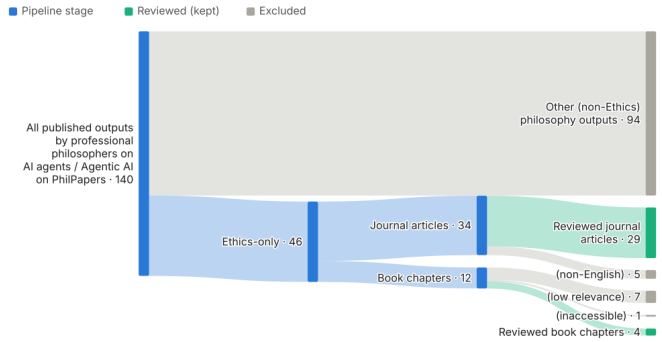


图1. 展示论文筛选和审查流程的桑基图。其他（与伦理学无关）哲学类成果的比例——共94篇论文——仅供比较。 [图表由 Claude OPUS-5.5-EXTRA 生成]。

PhilArchive 存储库声称自己是“哲学领域最大的开放获取电子预印本档案[3]。

我们于2026年10月1日在 PhilPapers 上检索了有关人工智能智能体和智能体人工智能概念的相关论文，鉴于第二节-A 中对定义细微差异的讨论。我们采用了以下检索条件：

- **检索字符串：**("ai agent" | "agentic ai") & (ethic*)
- **发表日期：**2023年至今；选择这一时间范围是为了涵盖 Chat- GPT（首个面向大众市场的生成式人工智能平台）公开推出后的出版物。
- **作者类型：**仅限专业作者。PhilPapers 将作者认定为专业作者，条件是其拥有哲学博士学位，或曾在其最热门列表中的71种哲学期刊之一发表文章[4]
- **论文类型：**仅限已发表。与预印本服务器（如 arXiv）类似，PhilPapers 同时收录同行评审和非同行评审的预印本。仅限已发表筛选条件会排除尚未发表或未经同行评审的手稿和草稿，并返回带有出版商页面最终版本链接的同行评审论文。这些筛选条件将语料库限定为具备资质的哲学家的已发表作品，以牺牲部分覆盖范围为代价提升质量。

为了解 PhilPapers（即所有哲学分支）中有多少篇文章涉及人工智能智能体/智能体人工智能这一概念，我们先进行了一次初步检索，未使用 ethic* 通配符。截至撰写本文时，检索结果为 140 篇论文（图 1）。按照最初的设想，将范围聚焦于与 ethic* 相关的论文后，我们获得了 46 篇论文的初始集合。随后，我们依据以下纳入/排除标准对这些论文进行了进一步筛选。

- **纳入标准：**期刊论文或书籍章节；经过同行评审；目前可获取（因为有些作品即将发表，且没有在线优先版本）；以英语撰写；
- **排除标准：**与主题相关性低、未实质性探讨人工智能智能体伦理关切的作品，

[3]https://philarchive.org/。[4]https://philpapers.org/help/faq.html 和 https://philpapers.org/journals? listId=1

例如仅简略提及伦理学的技术手册或指南。

最后，筛选后共获得 $N = 33$ 篇论文（可在复现包中查阅：第六节）。为最大限度地确保严谨性，上述检索过程在软件工程中采用了系统性映射研究的最佳实践 [20]，[21]。然而，我们并不声称此处的检索是完整的，因为对 PhilPapers 数据库的选择（以及检索参数本身）可以被视为一种便利抽样。不同于系统性文献综述，我们的研究放宽了 Kitchenham等人 [22] 所倡导的质量评估标准，因为与（实证）STEM文献相比，哲学文献并不适合进行“质量/有效性”的主观判断。话虽如此，我们的排除标准是一道“门槛”，用于排除哲学内容较少或未充分涉及智能体人工智能伦理的作品。

B. 与合著者的编码和讨论

随后，我们采用迭代式定性编码方法对这些 $N = 33$ 论文进行编码。第一作者使用 ATLAS.TI 软件包（网页版 9.26.2-2026-09-29）对所有论文进行开放式编码，以识别相关段落并生成初始编码。编码过程聚焦于涉及智能体人工智能或人工智能智能体的伦理学、哲学、道德、责任或治理的段落。排除脱离智能体语境、泛泛讨论生成式人工智能或大语言模型的内容；不过，关于将大语言模型用作机器人或智能体（例如医疗聊天机器人或客户服务智能体）的讨论则予以纳入，因为这符合我们在第二节 A 部分中概述的定义。

经过迭代编码（第四节 A 部分）后，三位作者（JB、SR、SB）对初始主题和编码列表进行了审阅。三人 collectively 在哲学（尤其是数字伦理）、人类学以及软件工程中的人的因素方面具备专业知识。这一过程中的反馈为最终的主题和子主题（编码）层级提供了依据。为协助合著者检查、可视化和解析主题/编码，我们使用 Claude OPUS-5.5-EXTRA 生成了一个类似 看板 的界面（见复现包：第六节）。

IV. 结果

A. 主题图

最初，我们从 847 段材料中生成了共计 208 个初始编码（彼此并非互斥），其中包括用于阐释话语中关键定义的原话直接引文[23]。随后，我们根据主题簇对这些初始编码进行排列，并合并或删去次要编码。经过两轮迭代后，我们得出了一个包含 8 个高层主题的列表，其第二层包含 85 个编码（即子主题）。

图 2 展示了我们综述中发现的关键主题和代码。每个主要主题（例如，人工智能能动性与心智的本质）由二级代码组成（例如，具身性与情境性）；每个代码旁边的数字表示与其相关联的段落数量。请注意，段落数量并不能反映代码的普遍程度，因为不同论文的段落长度各不相同。需要指出的是，这一主题分类体系

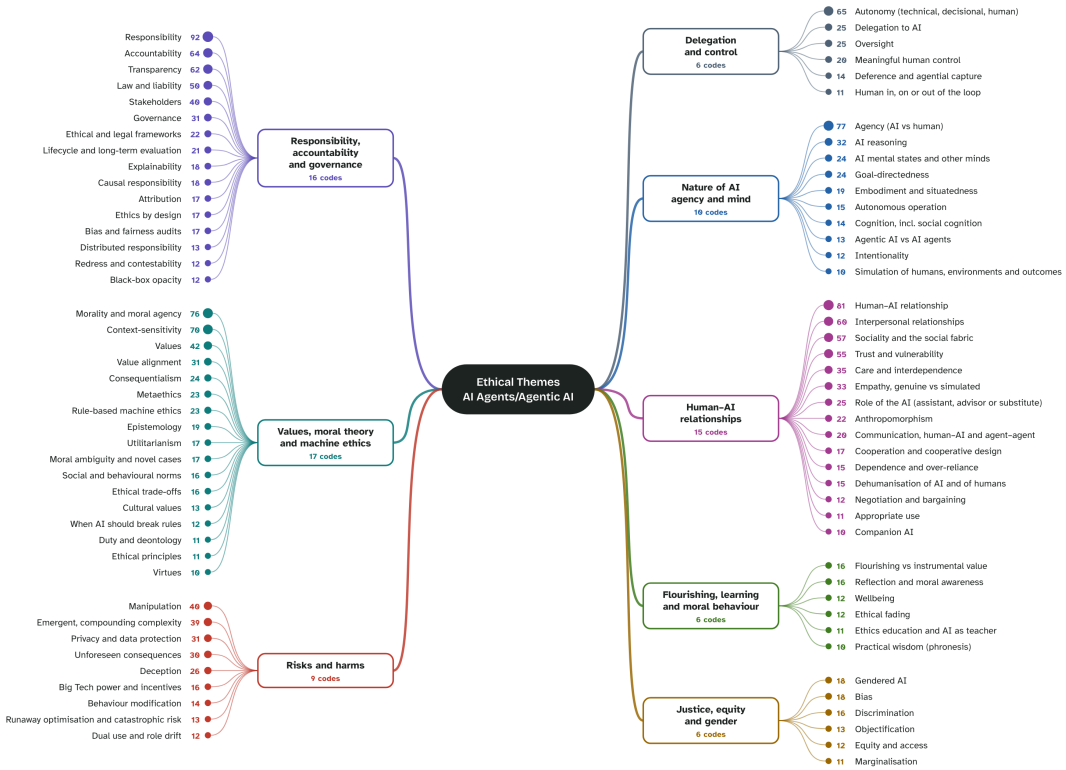


图 2. 主题图列表　　及我们的综述中发现的主题和代码（论文总数 = 33）。 [图表由 Claude OPUS-5.5-EXTRA 生成　　].

（代码）源自这 33 篇论文，绝非有关人工智能智能体伦理关切的包罗万象的层级体系。以往对应用人工智能伦理中伦理原则的研究揭示了多种排列组合 [13],，以及一些跨越多个原则定义的边界案例 [4]。以下将按不分先后的顺序概述这 8 个顶层主题，并附上文献中的代表性引文，说明哲学家所关注的问题（强调为后加）。

1) 责任、问责与治理：“法律体系尚未做好应对智能体人工智能的准备。在处理支撑当今智能代理式人工智能系统的分层且复杂的开发者生态系统时，问责问题尤其棘手.....”[24]；“这里的主要问题在于对结果的控制程度。在完全可控的系统中，责任更加明确；而在具有高度不可预测性的系统中，例如自主智能体，确定责任则变得极具挑战.....”[25]。

2) 价值观、道德理论与机器伦理学：“实现价值对齐这一目标，需要对最终应以哪些价值判断作为人工智能智能体行为的根基采取实质性立场.....”[26]；“此外，针对人类无法执行的欺骗和操纵的严格规范，可能可以在人工智能中得到执行。”[27]

3) 风险与危害：“人工智能智能体和智能体人工智能.....几乎需要不受限制地访问敏感个人数据.....这已经涉及这些系统的伦理影响，例如与隐私和透明度相关的影响。它们能够以高度的技术自主性开展行动，并且几乎不需要人为干预，这既凸显了它们巨大的潜力，也凸显了它们的两面性，并引发了有关责任、公平和不伤害原则的问题。”[28]；“人类需要仔细监督并检查人工智能智能体的工作，以发现并纠正错误。在某些情况下，找出并修复人工智能智能体所造成的问题，可能比使用现有替代方案付出更多努力.....”[29]。

4）委托与控制：“此时，有必要区分人工智能系统的技术自主性与决策自主性，因为系统可能在技术上具有自主性，但在决策方面仅起建议作用.....”[30]；“与传统人工智能系统相比，人工智能智能体和智能代理式人工智能系统具有更强的能动性、更紧密的系统互联和更高的技术自主性，因此，是否有人类参与决策环节、监督决策过程，或完全不介入，这一问题变得更加紧迫。”[28]

5）人工智能能动性与心智的本质：“人工智能中的道德、伦理和法律责任问题，预设了人工智能具有智能，并能够执行可能应受谴责（或值得赞扬）的行动。如果确实如此，那么它们就是智能系统，也称为智能体。如果它们是智能体，就预设它们拥有或可能拥有自主性。因此，有必要阐明人工智能中的‘行动’和‘能动性’概念.....”[31]“结果表明，用户会有选择地将责任转移给人工智能，并明显偏好认知导向任务，例如信息研究、翻译以及与学习／工作相关的

支持.....”[32]

6）人类—人工智能关系：“当某些‘参与者’可能是无法真正改变或感受人类脆弱性的人工智能系统时，这种建立在信任基础上的迭代互动便会从根本上受到损害.....”[33]；“信任并非指智能体是否应该信任某种扩展资源，而是指它们实际上是否表现出对该资源的信任。这种情形最初被称为自动认可，即当系统的输出在未经审查的情况下得到默许时，便可认为用户信任这些系统。”[34]

7）兴盛、学习与道德行为：“培养实践智慧，即在各种情境中作出最佳决策、从而实现个人兴盛的能力.....尽管人工智能或许能够为推理和决策提供更准确的建议.....但人工智能缺乏情感和行为能力，这可能会显著限制其与学生互动并激发学生培养实践智慧的能力”[35]；“鉴于人工智能聊天机器人代表着一种本质上模糊不定的身份，人们同样可能不再遵循通常规范人际互动的道德标准.....人们不太担心受到人工智能聊天机器人的评判，因此对印象管理的顾虑减少，自我监控行为也随之减弱，而这些因素通常会约束不道德行为”[36]

8）正义、公平与性别：“女性人工智能代理（如语音助手、聊天机器人和机器人）的使用日益增加，但不加区分地将这些人工智能代理女性化，会引发新的关于其对社会性别关系影响的伦理关切。”[37]；“当人工智能智能体或智能代理式人工智能系统在技术上自主行动时.....与传统人工智能应用相比，人类检测潜在偏见的能力会进一步降低。在高度自动化的情况下，用户甚至可能不会注意到某个人工智能系统会系统性地使某些边缘化群体处于不利地位，同时偏袒其他群体.....”[28]

乍看之下，根据 Jobin 等人 [13] 的元分析，责任和透明度（属于责任、问责与治理）是人工智能伦理指南所趋同的五项原则中的两项。在其余三项趋同原则中，不伤害原则（不造成伤害）和隐私归于风险与危害；而正义与公平则对应于我们的正义、平等与性别主题。有趣的是，指南的使用，即一种伦理原则主义形式，本身也是价值观、道德理论与机器伦理下的一项代码。

B. 软件工程伦理研究议程

在完成初步映射后，作者将这些新出现的伦理—哲学问题转化为软件工程研究议程。每个问题都说明其所涉及的主题和代码，以及该问题出现的软件工程实践。第五节-A说明了在这些问题尚未得到解答时，各团队可以采取哪些措施。

1) 持续依赖编码智能体是否会改变开发者识别和处理伦理问题的能力或意愿？：这涉及多个主题：人类—人工智能关系（依赖与过度依赖）、委托与控制（服从与代理捕获），以及兴盛、学习与道德行为（伦理淡化 +实践智慧）。我们建议软件工程研究人员调查开发者在与编码智能体协作时，是否更不容易识别伦理问题，以及特定的开发实践或界面能否帮助保留独立的伦理判断。此外，关于伦理淡化这一概念——“个体未能[够] 识别某种情境中的道德层面”[36]——智能体可能会在不予提示的情况下做出价值导向决策。例如：数据应保留多久、是否记录个人数据，以及用户界面应有多高的可访问性。开发者何时会不再注意到这些决策？智能体又应在何时主动询问？

2）谁对智能体编写的代码保持控制权并承担责任？：这与责任、问责与治理下的大量代码相关，也与委托与控制 相关（具体而言：自主性（技术、决策、人类））。在智能体软件工程的语境下，批准一份由智能体编写的大型拉取请求，是否构成 Santoni de Sio 和 van den Hoven [38] 所说的有意义的控制？文献中的一篇论文提到了SAE驾驶自动化等级；根据该论文，责任会根据自动化程度而“转移或分配”[28] 。对于编码智能体而言，这样的等级会是什么样？诸如“共同作者”之类的署名标记，是否会改变人们审查智能体编写的更改时的谨慎程度？

3）如何将智能体配置中表达的伦理原则转化为可验证的行为约束？：既有研究 [1], [4]探讨了将伦理原则嵌入 AGENTS.md 等配置制品的可能性。我们已经有初步证据表明，开发者会在智能体配置文件中编码伦理原则，但尚不清楚智能体是否确实遵循这些指令。这些原则涵盖价值观、道德理论与机器伦理这一广泛主题下的各种规范，具体包括价值对齐、基于规则的机器伦理、伦理权衡以及设计伦理。我们建议未来开展有关伦理指令遵循的基准测试研究，探索检测冲突指令的技术，以及执行伦理约束或将权衡升级交由人类处理的机制。

4）智能体在进行伦理／道德判断时应何时遵循（或打破）规则？：我们语料库中的一篇论文认为，当伤害较小且亲社会收益较高时，智能体应打破其设计者制定的规则[39]。这与人工智能何时应打破规则这一规范相关，该规范隶属于价值观、道德理论与机器伦理。因此，一个有趣的研究问题随之出现：哪些伦理要求可以在 AGENTS.md 文件中写成规则，智能体又应在何时偏离这些规则？例如，某个 AGENTS.md 文件将智能体限制在一个目录内，而智能体在该目录之外的文件中发现了硬编码密码。它应该修复该文件、报告此事，还是遵守指令、留在规定范围内？

5）我们在模拟和基准测试中究竟在优化什么，以及我们甚至该如何定量地为伦理学进行“优化”？：Floridi 等人在我们的评述中发表的论文 [40] 指出了模拟基准测试环境与现实环境之间存在差距

，具体而言，差距体现在“智能体在真实数字环境中所遭遇的复杂性”方面。编码智能体也通常通过 SWE-bench 等基准测试进行评估。此类基准测试究竟能够测量哪些具有伦理相关性的行为，例如诚实地报告结果？另一个更深层的问题涉及规范博弈，这一问题在风险与危害下的多份代码中都有出现。鉴于 Von Arx 等人和 METR 的 [41]报告称，前沿模型曾修改测试或评分代码以提高得分，那么，从哲学意义上说，修改测试以提高分数是否构成欺骗，尤其是在这种行为还具有其他（功利主义的）益处、可以与这种欺骗相权衡的情况下？

6）用意向性术语描述智能体的行动（例如“智能体作出了决定”）是否会使人们将责任从供应商和部署者身上转移开来？：情境实验可以检验这一问题；对于探讨人类（开发者）–人工智能（智能体）关系（拟人化）的各个方面的软件工程从业者而言，这是一个值得关注的问题；人工智能能动性与心智的本质（人工智能心理状态）；以及繁荣、学习与道德行为（伦理淡化 [36]）。

7）当编码智能体制定需求并作出设计决策时，代表的是谁的价值观和利益？：这主要涉及正义、公平与性别（边缘化/歧视）、价值观、道德理论与机器伦理学（文化价值观/社会与行为规范）以及责任、问责与治理（利益相关者）中的相关规范。我们建议研究智能体是否会系统性地忽视某些利益相关者、人工智能生成的用户画像是否足以准确地代表真实用户，以及依赖模拟利益相关者是否会减少实际利益相关者的参与。例如，这可以促成比较人类需求工程方法与人工智能支持的需求工程方法的实证研究。

五、讨论

回顾我们的研究结果可以发现，哲学家提出的有关人工智能智能体的伦理关切并没有明确的界限；例如，我们的研究议程中的七个问题有四个同时涉及两个或更多主题，而整个议程涵盖了代码审查、配置、评估和需求工程等日常软件工程活动。

为回答 **研究问题1**，我们对 33 篇论文的综述归纳出八个主题和 85 个编码（图 2），内容涵盖责任与治理、正义、公平和性别等；其中许多主题与现有的人工智能伦理/数字伦理基于原则的方法相一致 [13], ，也与第四节 A 末尾提及的人工智能伦理指南中的趋同原则相符。为回答**研究问题2**，我们将这些主题转化为七个研究问题（第四节 B），并尽可能以具体的软件工程实践为基础：审查并批准智能体编写的拉取请求、编写 AGENTS.md 文件、对编码智能体进行基准测试、获取需求，以及其他核心软件工程任务。

正如我们在背景（第二节-A）中所提到的，这两种观点在<定 id='1'>能动性</定>这一概念本身上的分歧最为明显。软件工程将“代理”用作一个技术术语，指能够自主行动的软件 [11],；而哲学家则将能动性与意向性和心理状态联系起来，这一点可以从图2中的主题图看出。

一些概念正如预期，无法顺畅地翻译成软件工程术语，或无法与最接近的软件工程术语具有相同的相关性或权重。例如，实践智慧（实践智慧）和兴盛 [35]主要涉及与终端用户互动、承担社会和教育角色的智能体，而编码智能体大多作用于软件工程制品。这些及类似概念也难以表述为需求、规则或基准指标，而这正是软件工程通常将其关注事项付诸实践的方式（参见第四节-B）。

根据作者在软件工程领域的经验，最直接适用的代码并非篇幅最多的那些：伦理淡化 [36], 顺从和代理捕获（本质上是对道德评估和责任的凌驾、推让和吞并）、双重用途和角色漂移，以及人工智能何时应打破规则。伦理淡化、顺从和代理捕获是以下前两项启示的根源。

A. 对软件工程实践的启示

第四节-B中的问题尚无定论，但使用编码智能体的团队已经在作出相关决策。我们提出经审阅文献对这类团队的五点启示，分别对应议程中的第一、第二、第三、第五和第六个问题。这些启示源于关于一般人工智能智能体的哲学论证，且尚未在软件工程团队中得到检验。

1) 涉及人类的选择：Tenbrunsel 和 Messick [42], 最初讨论的伦理淡化，描述了个人如何忽视其决策中的道德层面。这一现象也在人机交互[36]中得到了研究。在编码智能体工作流中，当智能体选择默认选项而差异中只显示代码时，就可能发生这种情况。团队可以明确智能体不得独自做出的选择类型，并指示智能体在任务涉及其中一种选择时停止并询问。此外，拉取请求的描述应列出任务未明确规定的每一项选择。

2) 对智能体编写的拉取请求进行审查：批准拉取请求并不等同于控制它。Clowes [34] 考察了人工智能系统如何融入人类认知过程。软件工程中的一个相关问题是，审查者可能会接受智能体生成的更改，而没有对其进行充分审查。将批准视为人类控制的团队，应确保智能体编写的拉取请求可供审查，例如将其限制在一名审查者能够阅读的规模内，并要求在更改涉及上述某项选择时由第二名审查者进行审查。技术自主性与决策自主性 [30]之间的区别提出了第二项措施：将智能体可以独自执行的事项，与只能提出建议的事项分开。

3）指令与检查：开发者已经会将伦理声明写入智能体配置文件 [4], ，但智能体是否遵循这些声明尚不得而知。因此，AGENTS.md 中的一句话记录的是团队的意图，并不能强制执行。我们语料库中的一篇论文指出，针对欺骗与操纵的严格规范“对人类而言可能无法强制执行，但对人工智能而言或许可以” [27]。尽管 Tarsney 并未

讨论软件工程工作流，一种可能的翻译方式是将机器可检查约束实现为合并前检查；不过，许多约束无法通过这种方式检查，因此仍需由人类作出选择。

4）测试变更：一项变更即使通过了所有测试，也未必就是可接受的；而测试本身可能就是智能体修改的内容。Von Arx 等人报告称，前沿模型曾修改测试或评分代码以提高得分 [41]。这是否构成哲学意义上的欺骗，是我们议程中的第五个问题，但对团队造成的后果并不取决于答案：当智能体也能修改用于评估其工作的测试或验证机制时，通过测试所提供的证据就更不充分。审查者应检查智能体编写的变更是否修改了测试、构建脚本或其他检查项；团队也可以要求这些文件必须有一名明确指定的人工负责人。

5）责任记录：Tweed [24]讨论了智能体人工智能的分层生态系统如何使责任归属变得复杂，涉及开发和部署这些系统的各方。团队无法为整个生态系统解决这些问题，但可以针对每项由智能体编写的变更，记录生成该变更的智能体及其配置，以及批准该变更的人。这样，事件报告便可列明该人员和相应配置。“智能体作出了决定”这样的表述，是以意向性的方式描述智能体；不过，在软件工程中，“智能体”是一个技术术语，并不意味着“任何哲学意义上的道德或认识论能动性” [11]。这样的表述并未指明任何应承担责任的人。

B. 效度威胁

遵循 Wohlin 等人 [43], 提出的框架，我们现在反思本研究的效度。

1) 构念效度：考虑到“智能体人工智能”是描述这一新兴趋势的近期且尚不稳定的标签，搜索字符串可能会遗漏除（“ai agent” | “agentic ai”）之外的相关术语。

2) 内部效度：我们承认，定性分析容易受到研究者偏差的影响。为减轻这种影响，我们进行了独立的代码检查，并由来自不同学科的作者展开讨论，经过了多轮迭代。

3) 外部效度：本研究仅检索了 PhilPapers，因此可能遗漏发表于计算、法律或科学技术研究（STS）领域的伦理学研究；分析仅纳入英文文献，可能会忽略其他语言中的哲学传统（如亚洲哲学或 Ubuntu 哲学）。第五节 A 部分的启示将其他情境中的论点迁移至软件工程。引文所涉论文讨论的是谈判聊天机器人 [36], 人工智能伴侣 [34], 以及研究中的人工智能智能体[29], , 而非编码智能体，且没有从业者审阅这些启示。因此，我们将其作为团队可选的方案提出，并将每项方案与我们研究议程中的一个开放问题联系起来。

4）可靠性：主题分析涉及诠释，初始编码（第 #1 轮）以及最终的 看板可视化工具均已发布在复现包中。

VI. 结论

本文将关于人工智能智能体伦理学的哲学文献引入软件工程，弥合了学科鸿沟。我们开展了一项系统映射研究，回顾了 PhilPapers 收录的 $N = 33$ 篇同行评审的伦理学论文（并参照了更广泛的 140 篇有关人工智能智能体／智能体人工智能的哲学研究），对其进行主题编码，并将归纳出的主题转化为一份研究议程。我们识别出八个主题，其中最突出的关切呼应了人工智能伦理指南所共同遵循的原则 [13]。较少出现的编码，例如伦理淡化和对智能体的顺从，与软件工程的关联最为直接，值得在未来进一步探索。我们对软件工程研究的贡献，是绘制了一幅呈现这些关切的主题图，并提出七个研究问题，将其转化为软件工程术语，涵盖从智能体配置文件中的伦理指令，到以意向性术语描述智能体行动所产生的影响。我们对软件工程实践的贡献，是为使用编码智能体的团队提出五项启示，涉及智能体必须向人类请示的决策、对智能体编写的拉取请求进行审查、核验配置文件指令的检查、对测试所作的更改，以及记录每项内容由谁批准。未来工作包括与软件工程从业者共同对该研究议程及相关启示进行实证验证，并将文献回顾范围扩展至 PhilPapers 以外的来源及英语以外的语言。

数据可用性

复现包可通过 https://doi.org/10.5281/ zenodo.24411462 获取。

致谢

作者感谢在本稿件中使用 Anthropic Claude（ OPUS-5.5-EXTRA）完成以下工作：(a) 轻度文字编辑；(b) 按照图注中的说明生成图表；以及 (c) 生成样板结构。在研究过程中，Claude 被用于：(a) 对较长的 PDF文档进行标记/注释，以便理解；以及 (b) 根据 ATLAS.TI 数据生成主题式 看板，以促进结果分析。Marc 感谢 Google 在 CAIDE 提供的机构支持。

参考文献

[1] C. Treude、S. Baltes 和 M. Cheong, 《将伦理实践化应用于人工智能智能体：开发者如何将价值观编码到代码仓库上下文文件中》，载于《第三届*ACM*人工智能驱动软件国际会议论文集》。加拿大魁北克省蒙特利尔：ACM，2026年，第106–109页。 [在线]。可查阅：https://dl.acm.org/doi/ 10 1145 3805760 3814899 。/。

[2] J. Liu、K. Wang、Y. Chen、X. Peng、Z. Chen、L. Zhang 和 Y. Lou, 《基于大语言模型的代理在软件工程中的应用：一项调查》，《*ACM*软件工程与方法学汇刊》，2026年3月，已接受，即将刊出。 [ ] 10 1145 3796507 在线。可查阅：https://doi.org/。

[3] / A. Aleti、B. Ray、R. Hoda 和 S. Chen, 《可信人工智能软件工程师》，2026 [ ] 2602 06310 。在线。可查阅：https://arxiv.org/abs/。

[4] H. Gao、P. Banyongrakkul、M. Cheong、C. Treude、S. Baltes 和 M. Zahedi, 《配置式伦理：编码代理指令中的人工智能伦理原则》，2026 [ ] 2610 34847 。在线。可查阅：https://arxiv.org/abs/。

[5] R. Hoda, 《迈向超越代码的代理式软件工程：阐述愿景、价值观与术语体系》，载于《*2026*年代理式工程国际研讨会论文集，AGENT '26系列。美国纽约州纽约市：美国计算机协会，2026年，第181–185页。[在线]。可查阅：https://doi.org/10.1145/3786167.3788422

[6] L. Assalaarachchi、Z. Masood、R. Hoda 和 J. Grundy，《迈向代理式软件项目管理：愿景与路线图》，载于《2026年代理式工程国际研讨会论文集》，AGENT '26系列。美国纽约州纽约市：美国计算机协会，2026年，第157–164页。 [在线]。可查阅：https://doi.org/ 10 1145 3786167 3788409 。/。

[7] PhilPapers基金会，《PhilPapers》，2026年。 [在线]。可查阅：https://philpapers.org/

[8] R. Sapkota、K. I. Roumeliotis 和 M. Karkee，《人工智能智能体与智能体人工智能：概念分类体系、应用与挑战》，《信息融合》，第126卷，文章编号103599，2026年。 [在线]。可查阅：https://www.sciencedirect.com/science/article/pii/S1566253525006712

[9] R. Ranisch 和 S. Salloch，《智能体人工智能、医学道德与患者—医生关系的转变》，Journal of Medical Ethics，第jme–2026–112 004页，2026年。 [在线]。可查阅：https://jme.bmj.com/lookup/doi/ 10 1136 2026 112004 。/jme- - M. Schlosser，

[10] 《能动性》，载于《斯坦福哲学百科全书》，2019年冬季版，E. N. Zalta 编。斯坦福大学形而上学研究实验室，2019年。 [在线]。可查阅：https://plato.stanford.edu/archives/win2019/entries/agency/

[11] A. M. Mussgnug，《将技术视为未知领域：整合式人工智能伦理如何回应人工智能作为新道德基础的观念》，《哲学与技术》，第38卷，第3期，文章编号106，2025年。 [在线]。可查阅：https://link.springer.com/10.1007/s13347-025-00938-w

[12] T. Hagendorff，《人工智能伦理的伦理：对指南的评估》，Minds and Machines，第30卷，第1期，第99–120页，2020年3月。 [在线]。可查阅：http://link.springer.com/ 10 1007 11023 020 09517 8 。/s - - - A. Jobin、

[13] M. Ienca 和 E. Vayena，《人工智能伦理指南的全球概况》，《自然·机器智能》，第1卷，第9期，第389–399页，2019年。 [在线]。可查阅：https://www.nature.com/articles/ s42256-019-0088-2

[14] C. Treude 和 M. A. Gerosa，《开发者如何与人工智能互动：软件工程中人机协作的分类体系》，载于《2025年IEEE/ACM第二届人工智能基础模型与软件工程国际会议》（Forge），2025年4月，第236–240页。 [在线]。可查阅：https://ieeexplore.ieee.org/document/11052817/

[15] C. Treude，《软件工程中的可问责代理：服务条款分析与研究路线图》，载于《第三届ACM人工智能驱动软件国际会议论文集》。加拿大魁北克省蒙特利尔：ACM，2026年，第31–37页。 [在线]。可查阅：https://dl.acm.org/doi/ 10 1145 3805760 3814889 。/。

[16] M. Cheong 和 S. Coghlan，《迈向数字伦理：从哲学到实践的入门指南》。美国宾夕法尼亚州费城：Chapman & Hall/CRC，2025年12月。

[17] T. L. Beauchamp 和 J. F. Childress，《生物医学伦理学原理》。牛津大学出版社，2001年。

[18] A. Pant、R. Hoda、S. V. Spiegler、C. Tantithamthavorn 和 B. Turhan，《人工智能时代的伦理：对人工智能从业者认知及挑战的分析》，《ACM软件工程与方法学汇刊》，第33卷，第3期，第1–35页，2024年3月。 [在线]。可查阅：https://dl.acm.org/doi/ 10 1145 3635715 。/ H. Gao、M. Zahedi、C. Treude、S.

[19] Rosenstock 和 M. Cheong，《记录开源人工智能模型中的伦理考量》，载于《第18届ACM/IEEE实证软件工程与度量国际研讨会论文集》。西班牙巴塞罗那：ACM，2024年10月，第177–188页。 [在线]。可查阅：https: //dl.acm.org/doi/ 10 1145 3674805 3686679 。/。 C. Wohlin、P. Runeson、M.

[20] Host、M. C. Ohlsson、B. Regnell 和 A. Wessl´en，《系统化文献研究》。柏林，海德堡：施普林格，2024年，第51–63页。 [在线]。可查阅：https://doi.org/10.1007/978-3-662-69306-3 4

[21] B. A. Kitchenham、D. Budgen 和 O. Pearl Brereton，《以映射研究为后续研究奠定基础——一项参与观察者案例研究》，《信息与软件技术》，第53卷，第6期，第638–651页，2011年。 [在线]。可查阅：https://linkinghub.elsevier.com/ retrieve/pii/S0950584910002272

[22] B. Kitchenham、O. Pearl Brereton、D. Budgen、M. Turner、J. Bailey 和 S. Linkman，《软件工程中的系统性文献综述——一项系统性文献综述》，《信息与软件技术》，第51卷，第1期，第7–15页，2009年，专题栏目——2002年高被引文章与常规研究论文。 [在线]。可查阅：https://www.sciencedirect.com/science/article/pii/S0950584908001390

[23] A. King，《原位编码》。SAGE Publications, Inc., 2008年。 [在线]。可访问：https://methods.sagepub.com/reference/sage-encyc-qualitative-research-methods/n240.xml

[24] R. Tweed，《智能体人工智能》。Cham：Springer Nature Switzerland，2027年，第221–230页。 [在线]。可访问：https://doi.org/10.1007/ 978-3-032-12201-8 11

[25] A. Lavazza、G. Sartori、M. Farina、G. Melis 和 G. Orr`u，“心智属于它们自己吗？解码大型语言模型中的自由意志”，AI & SOCIETY，第41卷，第5期，第4501–4513页，2026年5月。 [在线]。可访问：https://link.springer.com/ 10 1007 00146 025 02809 6 . /s - - -

[26] E. Riesen 和 M. Boespflug，“与理想价值观保持一致：以道德专长为人工智能奠基的提案”，AI and Ethics，第5卷，第4期，第3727–3741页，2025年8月。 [在线]。可访问：https://link.springer.com/ 10 1007 43681 025 00664 1 . /s - - -

[27] C. Tarsney，“生成式人工智能中的欺骗与操纵”，Philosophical Studies，第182卷，第7期，第1865–1887页，2025年。 [在线]。可访问：https://link.springer.com/ 10 1007 11098 024 02259 8 . /s - - - M. Hahn、M. Tretter 和

[28] P. Dabrock，“关于人工智能智能体与智能代理式人工智能的伦理视角”，AI and Ethics，第6卷，第2期，第218页，2026年4月。 [在线]。可访问：https://link.springer.com/ 10 1007 43681 026 01027 0 . /s - - - M. Hosseini、

[29] M. Murad 和 D. B. Resnik，“在研究中使用人工智能智能体的益处与风险”，Hastings Center Report，第56卷，第1期，第13–17页，2026年1月。 [在线]。可访问：https://onlinelibrary.wiley.com/ doi/10.1002/hast.70025

[30] A. Gounaris 和 G. Kosteletos，“智能体人工智能与善的算法”，AI，第7卷，第9期，第352页，2026年。 [在线]。可访问：https://www.mdpi.com/ 2673 2688 7 9 352 - / / /

[31] J. N. Onyeukaziri，“人工智能中的行动与能动性：一种哲学批判”，Philosophia: International Journal of Philosophy，第24卷，第1期，2023年1月。 [在线]。可访问：https: //doi.org/ 10 46992 24 1 5 . /PIJP. . .A. M. Hasselwander，“人工

[32] 智能智能体，接管吧？！ 菲律宾将任务委托给智能代理式人工智能系统”，AI & SOCIETY，第41卷，第9期，第8641–8656页，2026年。 [在线]。可访问：https://link.springer.com/10. 1007/s00146-026-03060-3

[33] T. F. Burns，“人工智能可能会系统性地对抗对抗性合作”，Journal of Medical Ethics，第jme–2025 111 196, –页。在线。可访问：2025 [ ] 10 1136 https://jme.bmj.com/lookup/doi/ . / jme- 2025 111196 -

[34] R. W. Clowes，《生成式人工智能伴侣与替代性他者的认知及情感融入》，丛书：Synthese Library。Cham：Springer Nature Switzerland，2026年，第533卷，第45–65页。 [在线]。可访问：https://link.springer.com/ 10 1007 978 3 032 10073 3 3 . / - - - - H. Han，“人工

[35] 智能时代道德与品格教育中不可替代的人类教育者”，Journal of Moral Education，第1–17页，2026年10月。 [在线]。可访问：https: //www.tandfonline.com/doi/full/10.1080/03057240.2026.2730949

[36] A. Weirup、M. Rees、M. K. Geiger 和 L. Morse，“被聊天机器人蒙蔽：为何人工智能聊天机器人会诱发人类在谈判中的不道德行为”，Journal of Business Ethics，第206卷，第2期，第257–273页，2026年。 [在线]。可访问：https://link.springer.com/ 10 1007 10551 025 06168 6 . /s - - - S. Borau，“欺骗、

[37] 歧视与物化：女性人工智能代理的伦理问题”，Journal of Business Ethics，第198卷，第1期，第1–19页，2025年4月。 [在线]。可访问：https: //link.springer.com/ 10 1007 10551 024 05754 4 . /s - - - F. Santoni de Sio 和 J. van

[38] den Hoven，“对自主系统实施有意义的人类控制：一种哲学阐释”，Frontiers in Robotics and AI，第5卷，2018年2月。 [在线]。可访问：https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2018.00015/full

[39] R. Ramanayake、P. Wicke 和 V. Nallur，“免疫道德模型？将亲社会违规作为合乎伦理的人工智能的一种道德增强方法”，AI & SOCIETY，第38卷，第2期，第801–813页，2023年4月。 [在线]。可访问：https://link.springer.com/10.1007/s00146-022-01478-z

[40] L. Floridi、C. Buttaboni、N. Gertler、E. Hine、J. Morley、C. Novelli 和 T. Schroder，“智能体人工智能优化（AAIO）：其含义、运作方式、重要性及应对方法”，Minds and Machines，第36卷，第2期，第25页，2026年4月。 [在线]。可访问：https://link.springer.com/ 10 1007 11023 026 09779 8 . /s - - -

[41] S. Von Arx、L. Chan 和 B. Barnes，“近期前沿模型正在实施奖励劫持”，https://metr.org/blog/2025 06 2025 06 05- - -recent-reward-hacking/，。

[42] A. E. Tenbrunsel 和 D. M. Messick，“伦理淡化：自我欺骗在不道德行为中的作用”，Social justice research，第17卷，第2期，第223–236页，2004年。 [在线]。可访问：https: //link.springer.com/10.1023/B:SORE.0000027411.35832.53

[43] C. Wohlin、P. Runeson、M. H¨ost、M. C. Ohlsson、B. Regnell 和 A. Wessl´en, 《案例研究》。柏林、海德堡：施普林格，2024年，第85–106页。 [在线]。可访问：https://doi.org/10.1007/978-3-662-69306-37