arXiv:2305.02739v1 [cs.AI] 4 May 2023

多智能体系统中的人类价值观

Nardine Osman 人工智能研究所（IIIA-CSIC）巴塞罗那，加泰罗尼亚，西班牙
nardine@iiia.csic.es

Mark d'Inverno 伦敦大学金史密斯学院 伦敦，英国
dinverno@gold.ac.uk

摘要
当今伦理人工智能面临的主要挑战之一，是开发能够证明其推理和行为与人类价值观保持一致的计算系统。然而，人类价值观向来具有模糊性、矛盾性和不断变化的特点。为了弥合这一差距，并使我们进一步接近能够以形式化方式推理如何将价值观融入人工智能的目标，本文提出了一种以社会科学为基础的价值观形式化表示。我们利用这一形式化表示，阐明了在多智能体系统（MAS）中实现价值一致行为所面临的关键挑战，并提出了一份旨在应对这些挑战的研究路线图。

关键词
人工智能中的人类价值观、计算价值对齐、概念基础、多智能体系统

1 引言

人们普遍认为，对人类价值观进行计算建模，对于设计由人类和人工智能体组成的混合群体所涉及的伦理多智能体系统（MAS）至关重要 [2,7, 16, 21]。我们提出一种直观、基础且具体的价值表征，该表征以社会科学为基础，是构建必要基础计算机制的前提，这些机制用于推理多智能体系统中的价值。我们认为，已发表的文献中尚不存在这样的模型。借助我们提出的形式化表征，我们展示了如何明确构建具有价值一致行为的多智能体系统所面临的计算挑战。通过努力借鉴其他学科，尤其是社会科学领域的研究成果，我们有意为跨学科研究团队携手合作铺平道路，并为其提供共同的概念基础。为此，我们展示了如何运用形式化模型界定构建伦理多智能体系统的四项关键研究挑战，具体如下：

(1) 挑战1。我们如何识别正在处理的价值？——价值识别与分类问题。(2) 挑战2。我们如何从个体价值转向集体价值？——价值聚合与共识问题。(3) 挑战3。智能体如何决定当前重视什么，以及接下来要做什么？——价值感知决策问题。(4) 挑战4。我们如何构建可持续且与价值一致的多智能体系统？——伦理多智能体系统问题。

2 价值表示的形式化模型

我们对价值的立场与社会科学的观点一致，即价值是指导行为的抽象概念，但其确切含义和解释会因语境和/或其他因素而存在很大差异。时间 [15, 17]。（社会科学中使用的建模方法各不相同，但我们致力于开发一种尽可能广泛借鉴这些方法的模型。）然而，要对价值进行有意义的评估，就需要对价值有具体的计算性表征理解。也就是说，尽管我们可能会泛泛地谈论公平这一我们希望拥有的价值，但在特定社群中，公平需要有更具体的定义。例如，在我们最近实施的一个用于支持互助社群的系统中，公平被理解为："任何成员都不会寻求远多于自己曾自愿为他人提供的帮助。"前者是一个抽象概念，而后者则是通过某项属性与公平这一价值相联系的具体共享理解（含义）；该属性是否满足（或满足程度）可以自动验证。从抽象价值转向具体的基于规则的实现这一思路，使我们提出应使用分类体系来定义价值。于是，任何一般性的价值概念（如公平）在分类体系中向下展开时都会变得更加具体，并在叶节点处成为具体、可计算且可验证的内容。（这种方法与价值敏感设计中关于价值变迁分类体系的研究一致 [20]。）

社会科学中另一个被认为是核心的重要概念是价值重要性，其中，个体或社群的价值的相对重要性决定了其行为。我们通过为价值分类体系中的每个节点附加一个重要性度量来纳入这一概念，但不规定该度量应采取何种形式。

我们通过分类体系进行价值表示的形式化方案见定义2.1。

定义*2.1*（价值分类体系）。价值分类体系 $\mathrm{V} = (N, E, I)$ 被定义为一个有向无环图，其中：

（1）节点集合 $N = N_l \cup N_\phi$ 表示价值概念，由两类节点组成：i）通过标签指定的节点，其中 $N_l \subset \mathrm{L}$ 表示标签节点的集合，L 是所有价值标签 的集合，这些价值标签表示诸如"公平"或"互惠"等抽象价值概念；ii）通过具体属性指定的节点，其中 $N_\phi \subset \Phi$ 表示属性节点的集合，Φ 表示所有价值属性 的集合，这些价值属性的满足情况可以在不同的世界状态下自动验证，例如，在互助社群中，一个人寻求帮助的次数不超过其提供帮助次数的 125%。（2）边集合 E：$N \times N$ 是有向边（n_p，n_c）的集合，$\in E$ 表示价值概念 n_p 与 n_c 之间的关系（分别为父节点和子节点），说明价值概念 n_p 比 n_c 更为一般。

（3）重要性函数 I： $N \rightarrow CD$ 为 N 中的每个价值概念分配一个取自陪域（值域） CD 的重要性值。

我们认为，价值分类体系中的属性节点使得以计算方式推理价值及解决价值对齐问题成为可能（对某一价值的属性满足程度越高，与该价值的对齐程度就越高——更多内容见第 3.4 节）。节点的重要性使得价值感知决策成为可能，其中包括选择采取哪些行动、遵守哪些规范（更多内容见第 3.3 节和第 3.4 节）。最后，分类体系的结构使得在不同情境中对价值作出不同解释成为可能（更多内容见第 3.1 节）。它也使得我们能够推理和审议价值的含义（更多内容见第 3.1 节和第 3.2 节）。

我们正在进行的工作 [**匿名引用**] 阐述了我们提出的价值表示方案，论证了这一方案的必要性与学术意义，并详述了我们的形式化方案与社会科学领域重要研究成果之间的对齐。该工作还列举了设计者可选用的实现实例，以及相应的机制和算法，以便：1）确保价值分类体系中价值重要性的一致性；2）支持计算价值对齐模型的实现。另一方面，这篇前瞻性论文运用我们提出的价值表示，指出了在多智能体系统中实现计算价值对齐所面临的关键研究挑战，并据此提出未来跨学科研究的路线图。

3 实现多智能体系统中价值对齐行为的路线图

3.1 价值识别与分类问题

价值识别与分类的挑战，在于确定任何多智能体系统中的价值是什么，并识别这些价值之间的相互关系，以及它们对于任何当前或设想中的多智能体系统的重要性。这包含两个部分。首先，如果我们希望加入一个组织并了解如何在其中取得成功，就需要理解该系统据以运行的价值，以及这些价值与我们自身价值之间的关系。同样，如果我们的挑战是设计一个新的多智能体系统，那么我们就希望与所有利益相关者合作，识别该系统运行过程中应遵循的价值。

相关工作。 人工智能领域目前针对这一主题的研究旨在从人际互动（通常是书面记录）中引出并学习相关价值。研究人员正在利用自然语言处理技术，以（半）自动化的方式从文本中估计潜在的人类价值观。例如， [11] 根据电子商务评论中使用的词语对价值进行了分析，而 [9] 则通过结合推文中的文本特征以及来自维基百科的情境知识，估计推文中所体现的相关价值。然而，只有在预先选定了高层次的价值列表之后，才会采用这些技术，例如广为人知的施瓦茨价值体系 [17]。使用任何预先定义的固定列表不仅存在列表是否适用于特定情境这一假设方面的局限性，还会阻碍价值随时间发生变化。这也是我们与价值敏感设计社群 [20]所持有的共同观点。在确实做到这一点的各种方法中，并非从预先定义的价值列表出发，而是旨在识别相关价值的研究可见于 [22],，该研究提出了一种众包算法，用于生成一般价值层级。另一个类似的研究见于 Axies，该研究结合人工和自动化技术，利用自然语言处理识别特定情境价值观 [10]。

确定未来研究路线图中的前进方向。这些现有方法对价值的理解通常仍停留在抽象层面。它们通过文本标题（例如“公平”）加以表述，却没有进一步探究所列各项价值的具体含义，也没有提供用于审议和推理这些价值列表的机制。我们将对这些被忽视的机制进行高层次描述，为进一步的研究与开发奠定基础。随后，这些描述将引出有关价值含义（通过属性节点体现）以及不同价值标签之间关系的问题。具体而言，我们确定了未来面临的一些关键研究挑战：

(1) 扩展现有关于价值识别的研究（例如 [10]），以便建立人类/人工智能流程最初识别出的价值之间的关系，从而构建我们所提出的价值分类体系。(2) 开发用于构建价值属性节点的机制（这些节点通常依赖于上下文），并将这些属性节点链接到抽象的标签节点。这对于任何旨在构建能够明确推理价值的人工智能系统的计算方法都至关重要。

(3) 开发自动传播机制：给定分类体系中某些节点的重要性，计算所有其余节点的重要性，并以确保整个分类体系中的重要性保持一致性的方式进行计算。开发此类机制在实践中将非常有用，因为获取每个单独节点的重要性通常并不简单（见下文讨论）。

我们认为，解决这些挑战对于在实践中真正推动将价值观引入人工智能系统至关重要。在这些研究挑战取得进展之前，很难想象我们如何才能信任人工智能系统，使其真正按照我们的价值观运行——这是人工智能最关键的伦理关切。明确界定的价值分类体系为反映相关利益相关者的人类价值观提供了一种显式机制，利益相关者可以查看并核验这些分类体系。人们要明确指定自己的价值分类体系并非易事。尽管许多从事价值敏感设计的伦理学家一直在明确引出利益相关者重视的价值及其相互关系，但要求技术用户完成这一过程，在实践中会过于繁琐且耗时。我们可以期待，普通用户／利益相关者能够大致了解人工智能系统从其价值体系中学到了什么，以及系统选择以何种明确方式对这些价值体系进行建模（即构建出的价值分类体系）。此外，我们可以期待他们能够认可或否定所学习到的价值体系的各个方面，从而引导人工智能学习和呈现价值观。但我们不能期待普通人深入了解每个节点的价值重要性、节点之间的确切关系等细节。因此，只有通过人工智能与人类利益相关者的协作，才能实现这种平衡

：人工智能向人类告知其正在学习的内容，而人类的输入则有助于引导学习过程。

3.2 价值聚合与协议问题

价值识别与分类侧重于识别单个实体（例如，人、社区、组织、公司等）的重要价值；价值聚合与共识则侧重于构建集体价值分类体系所需的机制。问题在于，我们如何从一组个体价值分类体系转向集体价值分类体系?

相关工作。 [7] 认为，我们生活在一个多元主义世界中，不同实体持有不同的价值体系。为了确保多智能体系统中的行为符合人类价值观，我们需要对任何多智能体系统的价值体系作出决策。要确定该价值体系，就必须处理个体乃至个体子群体之间可能相互冲突的价值体系。 [7] 将这一问题定义为：识别一种价值体系，尽管人们的道德信念存在广泛差异，它仍能获得“反思性认可”。 [13, 14]着重指出了水政策制定领域中处理个体利益冲突的挑战，并报告了围绕不同利益相关者的价值体系展开审议如何有助于解决此类冲突。该领域的一些研究 [8] 运用计算社会选择来聚合个体价值体系，并得出一个共识价值体系。这种方法考虑了多种伦理路径，从功利主义（最大效用）到平等主义（最大公平性）。

通过为未来研究制定路线图来确定前进方向。尽管关于价值聚合与共识的研究开始逐渐涌现，但仍有许多挑战亟待解决，其中包括以下问题。现在，我们已经为价值体系提供了一个正式、具体的模型，这些问题也因此变得更加明晰：

（1）开发用于计算社会选择的机制。这些 e
可以考虑我们关于价值体系的形式化提案中通过属性节点所定义的价值含义。换言之，我们需要一种复杂的聚合机制，其目的并非聚合各个价值概念的价值重要性，而是将完整的价值分类体系聚合为一个聚合价值分类体系。

(2) 制定价值共识机制。除了计算集体价值体系的聚合机制之外，还需要协议技术（如论证和谈判机制），以支持各个组成成员通过审议价值的含义（属性节点）和重要性，就拟议集体所采用的价值体系达成共识。

我们注意到，在这两项挑战中，个体的价值体系可能会发生变化，也可能不会发生变化，因为重点在于为集体就某一价值体系达成一致。因此，个体价值体系与集体价值体系之间可能会产生冲突。如果不一致性足够严重，这可能会促使个体不再参与该集体，并寻找与其自身价值体系更加契合的替代选择。在其他情况下，个体智能体可能被迫在集体内部进行互动，因此会认可社群的价值体系（无论其是否决定采取遵循集体价值体系的行动）。

3.3 价值感知决策问题

如前所述，识别个体和集体的价值体系，为基于价值进行推理提供了基础。智能体掌握自身的价值体系以及其当前所处集体的价值体系后，便能够推理应如何行动。计算方面的挑战在于开发增强型决策机制，将不同的价值体系，尤其是个体和集体的价值体系，纳入考量。

相关工作。在价值驱动决策领域，说服一直是促使智能体以特定方式行动的一种方法。在 [1], 中，提出了一种论证框架，其立场是：说服依赖于论据的强度，而论据的强度取决于所促进的社会价值观。在 [5], 中，描述了一种智能体模型，其中智能体行动同时受其需求和价值驱动，价值用于确定这些需求的优先级。

在其他研究中， [3], 信任的概念被作为一种影响决策的机制加以探讨，其中利用智能体过去行动的可靠性来判断该智能体是否值得信任。研究者指出，当无法利用过去的经验来评估他人的可靠性时，信任方与受托方之间的价值共享可以发挥作用，并提出了一种根据共同价值观的建立程度来评估信任的方法。在 [4], 中，则利用价值推理帮助智能体在需要采纳的计划中做出选择。

通过未来研究路线图明确前进方向。如前所述，我们的立场源自社会科学，其中价值是指导行为的抽象概念。现在，我们可以更清晰地识别这一领域面临的一些挑战：

（1）研究关于行动的推理，其中要明确认识到与行为选择相关的价值及其（当然）相对重要性。可以研究不同的方法，例如在认知智能体模型中采用实践推理，或扩展现有的BDI模型以纳入价值分类体系。还可以研究一种增强型心智理论：智能体能够观察彼此的行动，运用心智理论据此建立彼此的模型，并对这些行动及其潜在意图进行推理。此过程旨在支持观察者自身的决策过程。这些机制的重点是纳入价值分类体系，从而推理他人行为背后的价值，并据此做出考虑价值的决策。迄今为止，文献大多只是将价值用作标签，并未真正理解价值的含义；而使用价值分类体系，则可以丰富此类推理机制。

（2）开发价值驱动的审议机制，通过说服、论证或协商影响行为。这将拓展现有研究，例如 [1],的工作，并进一步研究价值共识。例如，人们或许不再基于行动的价值对齐来劝说他人应当如何行动，而是尝试就价值体系本身及其更新方式进行说服或论辩。说服他人改变其价值分类体系，可以促使其以某种方式行动。研究将聚焦于对价值分类体系中各节点的审议，并参照其重要性。

(3) 开发更完善的可解释性机制，帮助人类理解和探究自身行动可能产生的影响所涉及的伦理 意义，同时更好地理解驱动智能体以特定方式采取行动的伦理动机。这需要建立相应的机制，以便在理解所选行动可能产生的影响的同时，对这些影响进行推理；此外，还需要理解价值的具体含义，而我们的价值分类体系及其属性叶节点正是对此的体现。

3.4 多智能体系统中的伦理问题

第三项挑战侧重于个人决策过程中的价值对齐，而第四项挑战则关注如何开发多智能体系统，使其与人类价值观的对齐能够在较长时间内得到持续验证，尤其是考虑到用户和利益相关者所持有的不同价值体系可能会随着时间推移而发生变化。

相关工作。与人类价值观保持对齐的技术设计，是社会科学中的一个成熟领域，被称为价值敏感设计（VSD） [6]。VSD 依赖线下参与式设计以及线下评估，而我们的研究计划则通过提供一种在线验证机制来补充这一方法，该机制以计算方式评估对齐程度。

由于规范传统上一直被用于在多智能体系统中调节行为，因此，用于评估多智能体系统对齐程度的机制，已被简化为评估多智能体系统规范的价值对齐程度。如果一组规范促成的结果与给定价值体系更加一致，则称该组规范及其对应的多智能体系统与该价值体系相对齐。该领域的研究主要集中于选择一组最优规范，以最大限度地优化多智能体系统的价值对齐 [12, 18]。在 [18], 中，规范综合实现了自动化，并基于一些关于哪些规范能够促进哪些价值的初步知识。在 [12]中，提出了一种促进价值的规范综合方法，其本质上优化了 [19]中提出的价值对齐机制。在 [19], 中，价值偏好被理解为对世界状态的偏好，而一组给定规范的价值对齐程度，则取决于这些规范在多大程度上推动我们接近偏好的状态。

通过未来研究路线图确定前进方向。尽管伦理多智能体系统领域提出了多种机制，但该领域仍有许多挑战有待解决，例如：

(1) 根据我们的价值体系分类体系所提供的价值计算含义，改进价值对齐机制。现有的价值对齐机制存在两大主要缺陷。首先，它们需要人类投入大量手工工作，来界定价值（即我们分类体系中的属性节点）的含义及其重要性。我们认为，可以通过前文第 3.1 节所述的价值识别与分类机制解决这一问题。其次，许多现有机制在推理价值时，并不了解这些价值的含义。同样，我们认为，为价值提供计算含义（通过属性节点）能够提升价值对齐推理能力，从而给出更好的解释，接下来我们将对此进行介绍。

（2）开发解释机制，帮助人类理解：相对于给定的价值体系，为何一组规范比另一组规范更受偏好。鉴于我们引入了价值的计算含义，解释将更加详细。这类解释可以有力地支持新多智能体系统的设计，以及一般意义上的政策制定和协议设计，例如医疗协议、应急协议或用于规范灌溉实践的政策的设计。

（3）通过价值驱动的规范达成机制，为多智能体系统中的自我治理提供新机制。自我治理型多智能体系统面临的主要挑战之一，是就管理这些社会的规范达成共识。其目标是开发相关机制，支持智能体群体找出用于调节彼此互动的最佳规范集合。首先，规范的形式化分析、多智能体仿真以及基于人工智能的优化技术，都是可用于探索规范系统空间、寻找最优协调规范集合的技术。随后，可以开发价值驱动的审议机制，以支持围绕所选规范达成集体协议的过程。如上一个挑战中所述，解释还可以进一步增强审议机制。

4 结论

我们提出了一种价值体系的形式化基础表示，使我们能够对多智能体系统（MAS）中的价值进行计算推理。我们提出了一个受到社会科学领域多项研究影响的、我们认为最为直观的高层模型。不过，始终会存在其他建模方法；但通过阐述这一尝试，我们希望能够看到相应的反向提案。

随后，我们利用这一新表述，阐述了在多智能体系统中实现价值一致行为所面临的研究挑战，以及应对这些挑战所需的路线图。

人工智能的伦理设计这一愿景，需要持续开展协调一致的跨学科研究。我们有意将研究工作与社会科学领域关于价值的广泛研究相结合，不仅是为了夯实我们的形式化方案，也是为了提供一个概念框架，为跨学科研究的开展奠定基础。

致谢

这项工作得到了欧盟资助的 VALAWAI（# 101070930）和 WeNet（# 823783）项目，以及西班牙资助的 VAE（# TED2021- 131295B-C31）和 Rhymas （ # PID2020-113594RB-100 ）项目的支持。

参考文献

[1] Trevor Bench-Capon、Katie Atkinson。2009。抽象论证与价值观。载于人工智能中的论证，Guillermo Simari 和 Iyad Rahwan（编）。Springer US，马萨诸塞州波士顿，45–64。https://doi.org/10.1007/978-0-387-98197-0_3 [2] Raja Chatila、Virginia Dignum、Michael Fisher、Fosca Giannotti、Katharina Morik、Stuart Russell 和 Karen Yeung。2021。可信人工智能。载于关于人工智能与人类的思考。计算机科学讲义，第 12600 卷。Springer，–。https://doi.org/。/ - - - - Kinzang Chhogyal、Abhaya C. Nayak、Aditya Ghose 和 Hoa Khanh Dam。2019。多智能体系统的基于价值的信任评估模型。载于*IJCAI*。ij-cai.org，194–200。[4] Stephen Cranefield、Michael Winikoff、Virginia Dignum 和 Frank Dignum。2017。不给你披萨：BDI智能体中的基于价值的计划选择。载于第二十六届国际人工智能联合会议论文集，*IJCAI-17*，178–184。ijcai.org，–。https://doi.org/。/ijcai。/ Gennaro di Tosto 和 Frank Dignum。社会行为仿真：实现具有价值观和驱动力的智能体。载于基于多智能体的仿真 *XIII*——第*13*届国际研讨会，*MABS 2012*，西班牙瓦伦西亚，*2012*年*6*月*4*日至*8*日，修订版精选论文（计算机科学讲义，第*7838*卷），Francesca Giardini 和 Frédéric Amblard（编）。Springer，1–12。https://doi.org/ 10 1007 978 3 642 38859 0_1 。/ - - - - Batya Friedman、David G. Hendry 和 Alan Borning。2017。价值敏感设计方法综述。人机交互基础与趋势 11，2（2017年11月），63–125。https://doi.org/。/ Iason Gabriel。人工智能、价值观与对齐。心智与机器 30, 3（2020）：411–437。https://doi.org/。/s - - - Roger Lera-Leri、Filippo Bistaffa、Marc Serramia、Maite López-Sánchez 和 Juan A. Rodríguez-Aguilar。2022。迈向多元价值对齐：通过*lp*回归聚合价值体系。载于第*21*届自主智能体与多智能体系统国际会议，*AAMAS 2022*，新西兰奥克兰，*2020*年*5*月*9*日至*13*日，Piotr Faliszewski、Viviana Mascardi、Catherine Pelachaud 和 Matthew E. Taylor（编）。国际自主智能体与多智能体系统基金会（IFAAMAS），780–788。https://doi.org/ 10 5555 3535850 3535938 。/。Ying Lin、Joe Hoover、Morteza Dehghani、Marlon Mooijman 和 Heng Ji。2018。获取背景知识以改进道德价值预测。*2018*年*IEEE/ACM*社会网络分析与挖掘进展国际会议（*ASONAM*）（2018），552–559。[10] Enrico Liscio、Michiel van der Meer、Luciano C. Siebert、Catholijn M. Jonker、Niek Mouter 和 Pradeep K. Murukannaiah。2021。Axies：识别和评估特定情境价值观。载于第*20*届自主智能体与多智能体系统国际会议论文集（线上活动，英国）（*AAMAS '21*）。国际自主智能体与多智能体系统基金会，里奇兰，南卡罗来纳州，799–808。[11] 刘辉、黄英辉、王子超、刘凯、胡先根和王伟军。2019。人格还是价值观：基于在线评论增强型推荐系统的心理特征细分比较研究。应用科学（2019）。[12] Nieves Montes 和 Carles Sierra。2021。价值引导的参数化规范系统综合。载于*AAMAS '21*：第*20*届自主智能体与多智能体系统国际会议，线上活动，英国，*2021*年*5*月*3*日至*7*日，Frank Dignum、Alessio Lomuscio、Ulle Endriss 和 Ann Nowé（编）。美国计算机协会，907–915。[13] K. Pigmans、H. Aldewereld、V. Dignum 和 N. Doorn。2019。价值审议在促进水治理议题中的利益相关者参与方面的作用。水资源管理33（2019年10月），4067–4085。https://doi.org/ 10 1007 11269 019 02316 6 。/s - - - Klara Pigmans、Neelke Doorn、Huib Aldewereld 和 Virginia Dignum。2017。水治理中的决策：从利益冲突到共同价值观。施普林格国际出版公司，查姆，165–178。https://doi.org/ 10 1007 978 3 319 64834 7_10 。/ - - - - Meg J. Rohan。玫瑰不论叫什么名字？价值观这一构念。人格与社会心理学评论4，3（2000），255–277。https://doi.org/ 10 1207 15327957 0403_4 。/S PSPR S. Russell。人类兼容：人工智能与控制问题。企鹅出版集团。https://books.google.es/books?id=M1eFDwAAQBAJ [17] Shalom H. Schwartz。2012。施瓦茨基本价值观理论概述。心理学与文化在线读物2，1（2012）。https://doi.org/ 10 9707 2307 0919 1116 。/ - 。Marc Serramia、Maite López-Sánchez 和 Juan A. Rodríguez-Aguilar。2020。综合价值对齐规范系统的定性方法。载于第*19*届自主智能体与多智能体系统国际会议论文集，*AAMAS '20*，新西兰奥克兰，*2020*年*5*月*9*日至*13*日，Amal El Fallah Seghrouchni、Gita Sukthankar、Bo An 和 Neil Yorke-Smith（编）。国际自主智能体与多智能体系统基金会，1233–1241。[19] Carles Sierra、Nardine Osman、Pablo Noriega、Jordi Sabater-Mir 和 Antoni Perelló。2021。价值对齐：一种形式化方法。*CoRR* abs/2110.09240（2021）。arXiv: 2110 09240 。https://arxiv.org/abs/2110 09240 。https://arxiv.org/abs/。Ibo van de Poel。面向价值变化的设计。伦理与信息技术 23（2018）：27–31。https://doi.org/ 10 1007 10676 018 9461 9（），—。https://doi.org/。/s - - - T. L. van der Weide、F. Dignum、J. J. Ch. Meyer、H. Prakken 和 G. A. W. Vreeswijk。2010。运用价值观进行实践推理。载于计算机科学讲义。施普林格柏林海德堡，79–93。https://doi.org/ 10 1007 978 3 642 12805 9_5 。/ - - - - Steven Wilson、Yiting Shen 和 Rada Mihalcea。2018。构建和验证层级词典：以个人价值观为例。载于社会信息学（计算机科学讲义（计算机科学讲义））， Steffen Staab、Olessia Koltsova 和 Dmitry I. Ignatov（编）。施普林格国际出版股份公司，瑞士，455–470。https://doi.org/10.1007/978-3-030-01129-1_28 10届社会信息学国际会议，SocInfo 2018；会议日期：2018年9月25日至28日。