arXiv:2302.08759v1 [cs.AI] 17 Feb 2023

自主智能体的价值工程

Nieves Montes[1],Nardine Osman[1],Carles Sierra[1],和 Marija Slavkovik[2]

[1]人工智能研究所（IIIA-CSIC），西班牙巴塞罗那[2]挪威卑尔根大学{nmontes,nardine,sierra}@iiia.csic.es
marija.slavkovik@uib.no

2023年2月20日

摘要

机器伦理（ME）关注人工道德智能体（AMA）的设计，即能够根据道德价值观进行推理和行动的自主智能体。以往的方法将价值观视为与某些行动或世界状态相关联的标签，而非智能体推理的组成部分。人们也常常忽视这样一点：价值观引导的智能体会与其他价值观引导的智能体共同处于一个受规范支配的环境中，因此忽略了人工道德智能体的社会维度。在这篇前瞻性论文中，我们提出一种植根于道德与社会心理学的新型人工道德智能体范式，将价值观作为依赖情境的目标内化于智能体之中。这些目标通过评估现行规范最鼓励的结果，将个体层面的价值观与 $norms$ 集体层面的价值观紧密联系起来。我们认为，这种规范推理——即赋予智能体理解规范道德含义的能力——能够使自主智能体具备价值意识。此外，这种能力还为智能体协调其所处社会中施行的规范与其内化的人类价值观铺平了道路：在基于价值观对规范进行推理的基础上，辅以协商机制，帮助智能体集体商定最符合其人类价值观的一套规范。总体而言，我们的智能体模型不再将价值观视为惰性标签，而是将其与规范推理以及将价值感知智能体融入现代人机混合社会所需的社会功能联系起来。

关键词 – 价值观、价值意识、规范多智能体系统、机器伦理

1 引言

我们的社会是一个社会技术系统，其中包括人类智能体、借助计算设备增强自身能力的人类，以及能够在越来越大程度上自主行动的人工智能体。多智能体系统这一学科致力于识别并实施智能体之间能够促进协调的互动行动。

智能体协调不仅需要理性和自主性，还需要理解道德价值观——“机器的自由度越高，就越需要道德标准” [22]。Floridi 和 Sanders 的 [12]道德智能体定义明确包含人工智能体：“当且仅当一项行动能够导致道德善或道德恶时，该行动才被称为具有道德可评价性。当且仅当一个智能体能够实施具有道德可评价性的行动时，该智能体才被称为道德智能体。”我们如何构建人工道德智能体?

研究如何构建能够进行道德推理的人工道德智能体的领域，就是机器伦理（ME）。机器伦理“关注机器对人类用户及其他机器的行为” [2]。机器伦理的目标是使自动化和增强算法能够坚持“正确”的价值观。然而，机器伦理并不直接关注如何将道德推理与智能体的交互结合起来。价值观是什么，以及价值观如何形成，也仍然是一个悬而未决的问题。为了实现一个集体价值观与个体价值观相一致的社会技术社会，我们需要能够构建“理解”价值观的人工道德智能体。

我们在此提出一种新型人工道德智能体范式，使人类选定的价值观能够嵌入人工智能体的推理过程中。在对“价值”进行形式化定义时，我们遵循道德心理学而非道德哲学。我们在第 2 节阐述提出这一推理与方法的动机，并在第 3 节介绍该方法。在该节中，我们说明如何赋予智能体以价值观的含义，以及我们对规范所持的观点。我们还提出一种机制，使价值观与支配共享环境中行为的规范保持一致。最后，我们在第 4 节讨论相较于以往方法而言，本方法的新颖之处，并概述未来的研究方向。

2 关于价值观和规范的背景

2.1 什么是价值观

在更广泛的人工智能伦理领域中，关于应将哪些价值观强加于人工智能体（包括其使用与开发）的问题，主导着整个领域[14, 15]。然而，“价值”一词常常未经定义，也没有真正与伦理学联系起来。

在道德哲学中，伦理价值通常是相对于伦理原则来定义的。伦理原则是规范理论的一部分，该理论为道德规则和/或道德决策提供正当性依据或辩护，并反映了关于

对与错或“应当”如何的客观立场 [11]。IEEE 7000-2021《IEEE 系统设计过程中应对伦理关切的标准模型流程》[1]将伦理价值定义为“在人类文化语境中支持对何为对、何为错作出判断的价值”。从广义上理解，价值观如此定义后，便难以为计算代理所操作化。因此，研究人员采取了许多不同的方法，而这些方法难以比较或评估。此外，这种定义也不利于代理交互。具体而言，现有人工道德智能体的开发范式和原型使得智能体在推理自身的行动和状态时，未将他人的道德价值观纳入考量。

在强化学习中，这种方法正日益被用于实现具有道德行为的智能体 [8, 20, 5], ，我们可以看到，价值观的形式化（和量化）方式多种多样。在其他机器伦理方法中，例如 [17, 3, 9], ，价值观被用作行动、选项或推理过程中的其他要素的标签。

要构建具有价值观“意识”的智能体，我们需要将价值观形式化为智能体推理过程中的一个要素。所选的形式化方式应能对智能体价值观“意识”的程度以及代理交互进行界定。

2.2 什么是规范（以及多智能体系统）

在共享环境中交互的智能体会受到规范的约束。规范是一种社会机制，引导人类行为朝着某些期望的结果发展。规范负责构建智能体之间的交互，并塑造它们的整体组织方式。历史案例表明，当人类社群的成员遵循一套自行制定、定义明确的规范时，这些社群能够取得成功并持续发展 [21]。人们遵循某一特定规范的原因多种多样，可能是因为这样做对他们有利，也可能是因为他们认为这是一种道德或职业义务。

用于表征规范的形式语言丰富多样：基于约束 [13], 基于时间 [1], 或基于事件 [16], ，仅举几例。在许多情况下，尤其是在人类社会中，它们以自然语言表达的形式出现。然而，大多数表征语言，无论是否正式，都是结果性的，也就是说，规范通过惩罚或奖励发挥作用。某些人类行动会使人付出代价，另一些则会带来益处。

规范是一种系统级构造，适用于整个智能体社会。这一特点使规范有别于价值观，后者存在于每个个体代理之中。

3 构建价值感知智能体

本节提出我们关于人工价值感知智能体的方案。当智能体具备对价值之操作性

[1] https://standards.ieee.org/ieee/7000/6781/

含义的明确表征，并能据此解读多智能体系统的状态时，我们便称其为价值感知智能体。本文的核心论点如下：如果智能体具备价值感知能力，就能够从规范的价值对齐视角推理规范。换言之，它们可以分析一组规范所促成的结果（或其引发的多智能体系统行为），以及这些结果与期望价值观的对齐程度。这为从道德视角创建和选择规范，并从系统内部影响多智能体系统的价值对齐，开启了道路。

本节分为三个部分。第一部分介绍我们对价值观的看法，以及如何通过具体目标来表征价值观，从而为开发价值感知智能体打开大门。第二部分介绍我们对规范的看法。最后，第三部分将价值观与规范联系起来，为从规范所对齐的价值观视角对规范和多智能体系统行为进行推理奠定基础。

3.1 对于价值感知智能体而言，价值观是什么

作为构建人工道德智能体的起点，我们采用施瓦茨基本人类价值观理论（STBHV）[25, 26]。这是道德与社会心理学领域一套成熟的理论，为价值观提供了定义，阐述了价值观在社会生活中发挥的功能，并提示了价值结构的组织方式。此外，该理论所概述的许多价值观特征，都与社会科学与人文科学中的其他框架相兼容 [24]。

受到 STBHV 的启发，我们首先明确价值观是什么，以及它们如何在人类身上发挥作用。Schwartz 认为，价值观是“概念或信念”，能够“超越具体情境”[25, 第 4]页。尽管价值观具有抽象性，但它们与个体所处的情境密切相关：“价值观的主要内容方面，是其所表达的目标类型或动机关注点。（.....）价值观以明确目标的形式，体现了人类生存的三项普遍要求，所有个体和社会都必须对此作出回应（.....）”[25, 第 4]页。由此，STBHV 将价值观与另外两种具体事物联系起来：价值观在特定情境中所激发的明确目标，以及这些目标旨在实现的最终功能。

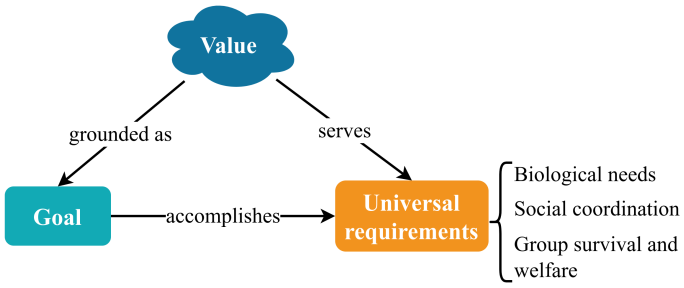

图1：施瓦茨基本人类价值观理论的三个主要组成部分。

这些目标旨在实现的最终功能。价值观、目标及其所满足的人类生存需求之间的关系见图1。

我们将 STBHV 中的这些动机目标作为实体，使价值观能够在计算环境中付诸实践。换言之，目标是智能体所追求的价值观的体现，并能捕捉某个价值观在特定情境或领域中的含义。因此，将目标编码到软件智能体中，可作为其所受驱动的价值观的替代表征。价值观具有超越性，而与之对应的目标则是依赖情境的。例如，在职业环境中，“性别平等”这一价值观应体现为在薪资和职业晋升方面同工同酬、同等认可；而在家庭环境中，这一价值观则更适宜体现为家务的平等分担。

传统上，人工智能中的“目标”指的是一种硬目标，其定义明确，评估结果非真即假 [30]。我们认为，这种二分式目标不够细致，无法反映价值观的复杂性。因此，我们提议使用模糊逻辑陈述来编码受价值观驱动的目标，以便在连续尺度上表示满意度的不同程度 [31, 23]。此外，这种方法还可以通过为相应目标分配满意度阈值，处理价值观的相对重要性（STBHV 中指出的另一项特征），这与 [7]中定义的分级欲望类似；如此一来，智能体就不必尽可能充分地实现目标，而只需达到令人满意的程度。正如受价值观驱动的目标依赖情境，价值观的相对重要性（即它们之间的偏好排序）也依赖情境，因此，标注相应目标的阈值亦是如此。

例如，如果在财富分配的语境中，平等 这一价值被视为重要，那么这一价值可以具体化为目标经济平等，并根据基于智能体对财富分配的感知所计算出的基尼指数来评估其真实程度。如果智能体非常重视这一价值，其满意度阈值就会较高（即基尼指数较低）；而如果平等在其偏好列表中的优先级较低，其满意度阈值就会较低（即使基尼指数较高，也不会促使智能体采取行动）。

既然我们已经将价值编码为目标，就可以讨论这些目标的用途。这里，我们同样借鉴STBHV，其核心观点是，价值通过其目标代理评估工具 [25]发挥作用。这些价值导向的评估可以应用于各种构成要素，例如行动、计划（即行动序列）、状态、结果，或上述要素的组合。因此，施瓦茨理论并未明确采取义务论立场或功利主义立场。具体而言，价值驱动的目标也可以评估规范，即关于行为的模式或指令，既可以通过其规定或禁止的行动，也可以通过其实施所带来的结果来进行评估。规范与价值之间的关系是我们提案的核心，详见第3.3节。

STBHV 未涉及的一个问题是价值观的起源。从价值观所服务的人类生存普遍需求（见图1）可以推断，价值观是进化的结果，而

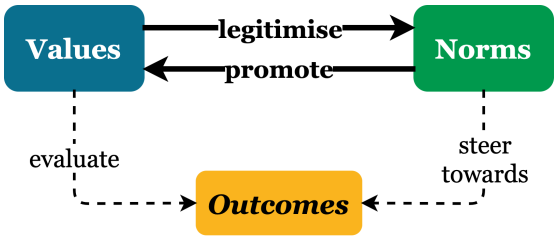

图2：价值观与规范通过结果建立的关系。

它们减轻了人们必须不断从纯粹生存角度思考所带来的认知负荷。在人工软件智能体的语境中，这类问题并不适用。尽管如此，价值观、受其驱动的目标及其相对重要性仍须有其来源。不过，现阶段我们的提案对价值 elicitation 过程持中立态度。我们关注的是价值感知智能体的内部运作，而暂不涉及价值驱动目标是以何种方式以及向谁询问得来的具体细节。

3.2 价值感知智能体的规范是什么

本文从以下抽象层面考察规范：规范通过施加奖励和惩罚，使由智能体行动引发的某些环境转移比其他转移更有可能发生。这种对“制度性”规范的看法相当宽泛，也包括“简单”规范，例如在满足给定前置条件时，对行动施加道义算子（如禁止、许可或义务），并由此产生相应后果的规范。

3.3 连接价值观与规范

如上所述，规范支配行为：它们激励行为朝特定方向发展。因此，我们认为规范作为主要的价值促进机制发挥着核心作用。如果经过精心设计，规范能够在很大程度上促进实现那些奠定价值观意义的目标，而这些目标植根于智能体所处的环境之中。

当实施一项新规范（或一组规范）带来的结果，就某种价值而言被视为高度正面时，我们称该规范与该价值相一致。因此，规范与价值观之间的关系本质上是结果性的 。规范本身并无道德属性；只有当它在社会中产生的影响符合成员以目标形式表达的价值观时，它才具有道德属性。

图2以示意图形式呈现了这两个实体之间的关系。表面上，价值观与规范构成一个反馈循环：价值观使强制执行的规范具有正当性，而规范在得到执行时会促进价值观。在更根本的层面上，两者通过结果相连：规范引导系统朝向这些结果，而这些结果会根据价值观受到正面评价。

确保人工智能与人类价值观保持一致，被认为是伦理人工智能的主要关注点之一 [10]。为解决这一问题，并考虑到规范在驱动行为方面的作用，一些关于自动化规范合成的研究应运而生，重点是根据规范所支持的道德价值观来选择规范，参见例如 [28, 27]。另一类研究是价值敏感设计（VSD）：由人类设计者手工设计在道德上适当的技术规范，而这一过程是在多智能体系统之外进行的。

我们认为，应由自主智能体尝试使规范与人类设计者赋予它们的价值观保持一致。据我们所知，只有 [18] 提出了一种架构，旨在通过智能体的参与实现指令性规范的内生涌现。然而，我们尚未发现该研究有任何后续工作。

我们主张，具有明确表征和语法的指令性规范应由系统中的智能体负责处理，并借助它们对价值观的理解（即价值观所奠基于其上的目标）作为评估依据加以评估。负责对智能体进行编程的设计者，无须直接编写技术规范。不过，设计者有责任纳入必要的机制，使智能体能够据此制定规范、推断规范可能导致的结果，并对这些结果进行伦理评估。

系统中的智能体由人类赋予价值观，并由人类将这些价值观的含义奠定为持久目标，因此人类能够完全掌控价值观的含义。可以为每个智能体提供其专属的、可能各不相同的同一价值观奠基目标版本，因此不同智能体可能拥有相互冲突的价值奠基目标。另一方面，规范是为集体设计的，旨在调节整个智能体社会的行为。如上所述，我们的立场是，智能体有责任使规范与人类设计者赋予它们的价值观保持一致。

在此可以应用协商技术，利用计算社会选择、论证和协商机制 [6]。当然，这可能会导致集体规范与某些个体价值观相冲突，但总体而言，这些规范能够代表系统中的价值观。

规范的价值对齐程度可以由智能体通过分析方法或模拟进行评估。分析方法涉及形式化推理，用于分析规范所诱导的行为结果。另一种方法是运行模拟，使智能体能够观察规范产生的结果。在这两种情况下，目标都是评估价值奠基目标在该结果中得到满足的程度，而这代表了规范的价值对齐程度。

总而言之，我们的提案探讨了智能体如何通过集体选择规范来约束彼此的交互，从而确保实现价值对齐行为。这是通过智能体具备推理规范的价值对齐程度的能力来实现的，进而从道德角度将规范或规范集合与价值观联系起来。这可以理解为通过赋予智能体价值感知能力来增强其能力。具备从道德视角分析规范的能力，便会

使智能体在制定、选择、组合规范，甚至决定遵守或违反规范时，能够作出价值感知的决策。这正是本文的核心主张。

4 结论

随着人工智能体自主性的增强，将伦理价值融入人工智能体已成为必然。当前构建人工道德智能体的方法，主要基于实现某种特定的道德理论，或通过学习（通常是强化学习）道德行为。伦理价值被用作语法层面的构造，而非智能体推理过程的一部分。具体而言，人类程序员将价值观作为标签赋予特定选项或计划。另一种做法是将价值观用于设定人为约束或严格规范，限制人工智能体可采取的行动，从而约束其行为，如 [9, 4] 所示。此外，人工道德智能体的“设计”通常也忽略了能动性的社会层面 [19, 29], ，也就是说，智能体存在于由法律和规范所支配的环境中，并与其他同样是道德智能体的主体共享这些环境。

我们提出改变智能体范式，将价值观设定为语义构造，使其不仅可用于智能体推理过程，还可用于调整共享环境中的规范。除了推理规范之外，我们还希望智能体具备价值感知能力，能够推理自身的行动，以及何时应遵守或打破规范。

本文未讨论的是如何嵌入具有价值观丰富的心智理论的智能体。价值观丰富的心智理论将使智能体能够直接推理其他智能体的行为，进而促成更具价值感知的交互。在我们提出的范式中，智能体只对规范进行推理，而这些规范会受到其他智能体价值观的影响。不过，关于智能体行动推理的讨论，我们留待另一篇论文展开。

参考资料

[1] Ågotnes, T., van der Hoek, W., Rodríguez-Aguilar, J. A., Sierra, C., and Wooldridge, M. J. 多模态计算树逻辑：完备性、复杂度及其应用。*Stud Logica 92*，1（2009年），1–26。[2] Anderson, M., and Anderson, S. L. 机器伦理的现状：来自AAAI研讨会的报告。*Minds Mach. 17*，1（2007年3月），1–10。[3] Anderson, M., and Anderson, S. L. Geneth：通用伦理困境分析器。载于第二十八届*AAAI*人工智能会议论文集，*2014*年*7*月*27*日至*31*日，魁 ´北克市，魁 ´北克省，加拿大（2014年），C. E. Brodley和P. Stone编，AAAI出版社，第253–261页。

[4] Arkin, R. C., Ulam, P., and Wagner, A. R. 自主系统中的道德决策：执行、道德情感、尊严、信任与欺骗。*IEEE*会议录 *100*，3（2012年），571–589。[5] Armstrong, S. 面向人工智能体的动机性价值选择。载于第二十九届*AAAI*人工智能会议研讨会（2015年）。[6] Baum, S. D. 人工智能中的社会选择伦理。*AI Soc. 35*，1（2020年），165–176。[7] Casali, A., Godo, L., and Sierra, C. 用于表征偏好并进行推理的分级BDI代理模型。《人工智能》*175*，7（2011年），1468–1478。《偏好的表征、处理与学习：理论与实践挑战》。

[8] Chaput, R., Duval, J., Boissier, O., Guillermin, M., and Hassas,S. 一种结合推理与学习以实现伦理行为的多智能体方法。发表于*AIES '21*：*AAAI/ACM*人工智能、伦理与社会会议，线上活动，美国，*2021*年*5*月*19*日至*21*日 (2021)，M. Fourcade、B. Kuipers、S. Lazar 和 D. K. Mulligan 编，ACM，第13–23页。

[9] Dennis, L. A., Fisher, M., Slavkovik, M., and Webster, M. 自主系统中伦理选择的形式化验证。机器人与自主系统 *77* (2016)，1–14。

[10] Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila,R., Dennis, L. A., Génova, G., Haim, G., Kliess, M. S., López-Sánchez, M., Micalizio, R., Pavón, J., Slavkovik, M., Smakman,M. H. J., van Steenbergen, M., Tedeschi, S., van der Torre, L.,Villata, S., and de Wildt, T. 设计伦理：必要性还是诅咒？载于《2018年AAAI/ACM人工智能、伦理与社会会议论文集》，AIES 2018，美国路易斯安那州新奥尔良，2018年2月2日至3日（2018），J. Furman、G. E. Marchant、H. Price 和 F. Rossi 编，ACM，第60–66页。[11] Fieser, J. 伦理学。载于《互联网哲学百科全书》，M. Boylan 编。ISSN 2161-0002，2021年。[12] Floridi, L., and Sanders, J. W. 人工智能体的道德性。《心智与机器》14，3（2004），349–379。[13] García-Camino, A., Rodríguez-Aguilar, J. A., Sierra, C., andVasconcelos, W. W. 电子制度中规范的基于约束规则的编程。《自主智能体与多智能体系统》18，1（2009），186–217。[14] Hagendorff, T. 人工智能伦理的伦理学：对指南的评估。《心智与机器》30，1（2020），99–120。

[15] Jobin, A., Ienca, M., and Vayena, E. 人工智能伦理指南的全球格局。《自然·机器智能》（2019）。[16] Kowalski, R. A., and Sergot, M. J. 基于逻辑的事件演算。载于《时间的语言——读本》，I. Mani、J. P ustejovsky 和 R. J. Gaizauskas 编。牛津大学出版社，2005年，第217–240页。[17] Liao, B., Anderson, M., and Anderson, S. L. 价值驱动型智能体中的表征、论证与解释：基于论证的方法。CoRR abs/1812.05362（2018）。[18] Morris-Martin, A., Vos, M. D., and Padget, J. 多智能体系统中的规范涌现：一篇观点论文。《自主智能体与多智能体系统》33，6（2019年9月），706–749。[19] Nallur, V. 机器实现的伦理概览。《科学与工程伦理》26，5（2020），2381–2399。[20] Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R.,Madan, P., Varshney, K. R., Campbell, M., Singh, M., and Rossi,F. 使用强化学习和策略编排教授人工智能智能体伦理价值。载于《第二十八届国际人工智能联合会议论文集》，IJCAI-19（2019年7月），国际人工智能联合会议组织，第6377–6381页。[21] Ostrom, E. 《治理公地：集体行动制度的进化》。Canto Classics。剑桥大学出版社，1990年。[22] Picard, R. W. 《情感计算》。麻省理工学院出版社，1997年。[23] Ramík, J. 目标规划问题中的模糊目标与模糊备选方案。《模糊集与系统》111，1（2000），81–86。[24] Rohan, M. J. 名字不重要？价值观构念。《人格与社会心理学评论》4，3（2000年8月），255–277。[25] Schwartz, S. H. 价值观内容与结构中的普遍性：20个国家中的理论进展与实证检验。载于《实验社会心理学进展》。Elsevier，1992年，第1–65页。[26] Schwartz, S. H. Schwartz 基本价值观理论概览。《心理学与文化在线读物》2，1（2012年12月）。[27] Serramià, M., López-Sánchez, M., and Rodríguez-Aguilar, J. A.一种构建价值对齐的规范系统的定性方法。载于《第十九届自主智能体与多智能体系统国际会议论文集》（南卡罗来纳州里奇兰，2020年），AAMAS '20，国际自主智能体与多智能体系统基金会，第1233–1241页。

[28] Serramià, M., López-Sánchez, M., Rodríguez-Aguilar, J. A., Morales, J., Wooldridge, M., and Ansotegui, C. 利用道德价值观选择恰当的规范。载于《*2018*年*AAAI/ACM*人工智能、伦理与社会大会论文集》（2018年12月），ACM。[29] Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., andBernstein, A. 机器伦理的实现：一项综述。*CoRRabs/2001.07573*（2020年）。[30] Wooldridge, M., and Jennings, N. R. 智能体理论、架构与语言：一项综述。载于《计算机科学讲义》。施普林格·柏林·海德堡，1995年，第1–39页。[31] Zadeh, L. A. 模糊逻辑与近似推理。*Synthese 30*，第3–4期（1975年），407–428页。