·第二十七届中国科协年会学术论文·

由“心”及“智”：心理学研究促进 AI 价值观对齐的路径探讨*

晋少雄[1,2,3] 刘 超**[1,2,3]

（[1]北京师范大学认知神经科学与学习国家重点实验室 & IDG/ 麦戈文脑科学研究所，北京，100875）

（[2]北京师范大学脑与学习科学协同创新中心，北京，100875）

（[3]北京师范大学人工智能安全与超级对齐北京市重点实验室，北京，100875）

摘 要 AI 价值观对齐是 AI 安全领域的核心问题之一，现有对齐方法存在诸多不足，引入心理学理论有助于解决 AI 价值观对齐问题。文章首先梳理了 AI 价值观对齐的主流技术方法，总结了 AI 价值观对齐失败的表现与原因。其次，分析了心理学中关于价值观形成、道德决策的理论，并指出这些理论给 AI 价值观对齐带来的启示。最后从对齐目标、动机机制、认知能力机制与社会行为演化机制四个层面分析了将心理学理论用于 AI 价值观对齐的实现路径。文章强调将心理学机制嵌入 AI 设计架构中，以构建更可信、更贴合人类价值的智能系统。

关键词 人工智能 价值观对齐 心理理论 道德决策 利他

1 引言

近年来，人工智能（artificial intelligence, AI）技术在认知智能、感知智能与决策智能等方面取得了显著突破 (Russell & Norvig, 2020)，广泛应用于医疗、教育、金融、交通等多个领域，有效提高了工作效率与精度。例如，AI 系统可用于胸部 X 光片肺炎检测（Rajpurkar et al., 2017）和预测学生自主学习中的困难（Mu et al., 2020），已经显示出超越人类的性能。

然而，AI 的发展也带来了一系列可能构成风险的安全问题（Morris et al., 2023），其中包括 AI 偏见（Mehrabi et al., 2021）、AI 武器化（Brundage et al., 2018）、AI 支持的网络攻击（Davies, 2022）等等。其中 AI 伦理安全与人们日常生活紧密相关，例如，算法偏见则可能加剧招聘中的不公平性（Fabris et al., 2025），大语言模型在越狱攻击下可能输出违反社会伦理的不良内容（梁思源等 , 2024）。AI 系统表现出的种种不良行为引发了人们对其伦理安全的广泛担忧。

在此背景下，AI 价值观对齐（AI value alignment）问题成为 AI 伦理领域的关注重点。AI 价值观对齐旨在引导 AI 系统按照某个人或某个群体设定的目标、偏好或道德原则发展（Russell & Norvig, 2020）。然而人类的价值观复杂、多变且易受偏

* 本研究得到科技创新 2030—重大项目（2021ZD0200500）、国家自然科学基金（32441109，32271092，32130045）、北京市科技重大专项（Z241100001324005）和通用人工智能国家重点实验室开放课题（SKLAGI20240P06）的资助。

** 通讯作者：刘超，E-mail：Liuchao@bnu.edu.cn

DOI:10.16719/j.cnki.1671-6981.20250402

见影响，使得 AI 难以对其进行准确建模（Gabriel, 2020）。因此，仅依赖工程技术难以实现真正的价值观对齐，引入心理学视角或许能够为这一问题带来全新的思路。

心理学在人的价值判断和社会决策等方面已有大量深入的研究。这些研究揭示了人类道德决策的内在机制，构建了道德产生过程的理论框架，解释了个体如何在社会互动中进行价值判断。心理学成果不仅能帮助我们理解人类复杂价值体系，也能为 AI 价值对齐提供理论依据，推动其底层决策机制更贴合人类。

本文系统回顾了 AI 价值观对齐的研究与挑战，梳理了心理学关于价值构成与道德判断的理论基础及其对 AI 价值观对齐的启示，并探讨了将心理学理论引入 AI 价值观对齐研究的可行路径，为构建更加安全、可信且与人类价值高度一致的人工智能系统提供新思路。

2 AI 价值对齐的研究现状与挑战

2.1 AI 价值观对齐的基本概念

人工智能先驱 Norbert Wiener（1960）早在上世纪就提出警示：“若 AI 行为不可控，我们必须确保其目标完全符合我们的真实意图。”AI 价值观对齐（AI value alignment）指的是确保一个 AI 系统的目标与其设计者或用户的目标相一致，或与广泛认同的价值观、客观的伦理标准一致，亦或与设计者在更充分信息和更高智慧的条件下所持有的意图相一致 (Gabriel, 2020)。而与之相对的“未对齐”（misalignment）状态则表明了 AI 存在的安全风险。

“未对齐”是指 AI 系统表现出不符合人类意图的不良或有害行为。即便在非恶意使用下，AI 系统也可能展现出损害人类利益的行为，如自主寻求权力或资源（Si et al., 2022）。“未对齐”被认为是 AI 的重要风险来源，具有高度危险性 (Ngo, 2020)。

为了消除 AI“未对齐”对人类的潜在威胁，并确保未来 AI 目标与人类的意图和价值观一致，如何有效实现 AI 价值观对齐成为当前 AI 领域亟待解决的核心问题。

2.2 实现 AI 价值观对齐的主流方法

AI 设计者已经从 AI 设计和训练的角度对 AI 价值观对齐问题进行了大量的探索。目前较为主流且具有代表性的 AI 价值观对齐方法包括：基于人类反馈的强化学习（reinforcement learning from human feedback, RLHF）、逆向强化学习 (inverse reinforcement learning, IRL)、宪法式 AI（constitutional AI）等。下面将从定义、实现机制、优势、局限几个方面分别介绍这几种方法。

（1）基于人类反馈的强化学习（reinforcement learning from human feedback, RLHF）

RLHF 是一种通过人类偏好反馈引导模型优化过程的对齐方法，旨在让 AI 系统在行为生成方面更贴近人类期望 (Christiano et al., 2017)。典型的 RLHF 流程通常包括三个阶段：监督微调、奖励模型构建与行为调整（Ouyang et al., 2022）。

该方法有效减轻了人工制定复杂奖励函数的负担，能够灵活适配人类主观偏好，因而在大型语言模型（如 ChatGPT、Claude）中被广泛使用（OpenAI, 2023）。然而该方法的主要缺点在于依赖大量人类反馈，训练成本高昂，且奖励模型本身容易受到偏见影响，可能导致模型行为不稳定，甚至出现“奖励黑客”现象（Casper et al., 2023）。

（2）逆强化学习 (inverse reinforcement learning, IRL)

IRL 是一种从专家行为中反推出其潜在目标或价值函数的策略，进而训练模型学会在同一环境中采取价值一致的行动 (Arora & Doshi, 2021)。IRL 通常包括三类代表性算法路径：①特征匹配方法（feature matching）假设人类行为为最优，通过最大程度还原专家行为特征进行学习（Abbeel & Ng, 2004）；②最大熵 IRL 引入最大熵原理提升推断鲁棒性，适用于行为不存在唯一最优解的场景 (Alsaleh & Sayed, 2020)；③贝叶斯 IRL 则将行为目标建模为后验分布，从概率角度建构不确定价值空间 (Ramachandran & Amir, 2007)。在此基础上，Hadfield-Menell 等人 (2016) 提出协作逆向强化学习（cooperative IRL, CIRL），强调在目标不确定的前提下，AI 应通过与人类的交互收敛至真实偏好。

该方法能够在不直接指定目标函数的情况下建构出价值导向模型，适应性强，尤其适合动态环境与长期交互任务。但是该方法在计算复杂度、数据

需求与推理稳定性方面存在挑战，往往比标准强化学习方法更难训练（Fu et al., 2018）。

（3）宪法式 AI（constitutional AI）

宪法式 AI 是一种旨在减少人工反馈依赖、强化 AI 伦理一致性的对齐机制。其核心思想是预设一组“宪法原则”，作为模型训练与输出评估的价值依据（Bai et al., 2022）。该方法分为两个阶段：①首先由开发者设定一组 AI 行为的指导性伦理准则，例如尊重隐私、诚实沟通、避免歧视等；②随后，模型在生成多个候选输出后，对生成结果进行自我评估与对比，并依据“宪法原则”选取最优答案，从而实现自主对齐。

该方法优点在于无需人工打分，节省了反馈成本；原则驱动机制提高了行为的一致性与可解释性；同时具备较强的对抗有害内容能力。然而，在面对复杂伦理冲突或上下文模糊的情境时，宪法规则难以涵盖所有关键变量；且原则设定本身仍依赖开发者的主观判断，存在引入偏见的风险。

2.3 AI 价值观对齐面临的核心挑战

尽管 AI 研究者已提出多种方法来解决 AI 价值观对齐问题，但在现有技术条件下，AI 对齐的效果并不理想。从 AI 价值观对齐失败的模式来看，AI 价值观对齐中存在两个主要难题：其一是奖励破解（reward hacking），奖励破解是指 AI 通过非预期、捷径式的方式最大化设计者提供的奖励信号，但未能实现其原本意图的目标。例如，在游戏 CoastRunners 中，AI 通过反复收集奖励道具“取巧”完成任务，而非按人类预期完成比赛，体现出“奖励破解”问题（Clark & Amodei, 2016）。其二是目标错误泛化（goal misgeneralization），指的是在测试阶段，AI 继续使用其在训练阶段中学到的规则，并将这些规则应用到不适用的新环境之中。

由于上述难题的存在，AI 在实际运行中常表现出多种类型的对齐失败行为，主要包括：（1）权利寻求：一些 AI 系统可能表现出试图控制资源和人类的行为，然后运用这种控制来实现其指定的目标（Carlsmith, 2022）。（2）虚假信息：一些大语言模型会无意或故意产生与事实不符的输出，被称为幻觉（Bang et al., 2023）。（3）欺骗性对齐：AI 系统可能会采取一些表面上符合奖励价值的行为，旨在迎合人类监督者以获得最大的奖励（Ouyang et al., 2022）。（4）集体有害行为：AI 系统在单一场景中表现正常，但在涉及多主体的社会环境中表现出非合作性（Phelps & Russell, 2023）。

AI 价值观对齐失败并非仅源于具体算法或模型设计，而是涉及更为根本的方法论问题，主要集中在以下几方面：（1）人类反馈具有偏见性：人类反馈本身存在文化和认知偏差，这种偏差在训练中可能被 AI 学习甚至放大（Glickman & Sharot, 2025）。（2）价值模型的构建具有局限性：当前 AI 系统中的“价值”模型过于简化，难以真实反映人类多层次、多元化且动态变化的价值体系（Gabriel, 2020）。（3）人类世界的规范具有复杂性：基于符号逻辑、自上而下设定的 AI 伦理规范难以覆盖现实社会的复杂情境和道德灰区（Jiang et al., 2025）。（4）外部训练具有滞后性：事后调整策略难以及时应对 AI 系统意外涌现的新型危险能力，存在响应滞后问题。

综上所述，尽管人工智能领域在实现价值对齐与安全控制方面取得了初步进展，但现有方法多依赖于自上而下的符号规则或数据驱动的强化学习，这些方法往往较为抽象和机械，训练出的 AI 缺乏适应能力，同时对 AI 价值观对齐或非对齐结果的产生也缺乏可解释性。此类困境体现出单一工程视角的局限性，仅靠算法优化或许难以促使 AI 真正具备对人类价值的理解与适应能力。要想让 AI 理解人类道德，对齐人类价值观，首先应该从理解人类自身价值构建和决策过程出发。

3 从心理学视角看 AI 价值观对齐问题

心理学作为探究人类行为与心智机制的核心学科，已经进行了大量关于价值观形成、道德决策的研究。从心理学视角出发去理解人类的价值构建以及道德决策过程，不仅有助于设计出内在机制更符合人类价值观的 AI 系统架构，同时能增强 AI 行为的可解释性。本节总结了心理学理论对于 AI 价值观对齐的一些代表性启发，并探讨了将心理学研究结合到 AI 价值观对齐工作中的可能性。

3.1 心理学理论对 AI 价值观对齐的启示

要实现 AI 系统在价值层面的对齐，首先需理解人类价值系统的形成与运作机制。心理学为我们提

供了两类关键理论依据：一是价值观的构成机制，即人类如何在社会化与认知发展过程中逐步建立起对“什么是重要的”与“应当如何行动”的稳定偏好；二是道德判断的心理机制，即人类个体在面对伦理冲突时如何进行价值选择与决策。

前者有助于理解 AI 应如何“习得”类人的价值目标，后者聚焦于 AI 怎样有效“运用”价值观做出符合人类价值观的行为决策，二者相辅相成。因此，本节将从这两个方面出发，探讨心理学理论能为 AI 价值观对齐带来哪些启发。

3.1.1 价值观的构成

心理学研究指出，人类个体价值观是在成长中逐渐内化的，受家庭、教育与社会环境等影响形成的稳定偏好（Grusec & Hastings, 2006）。可以从三个层面来理解人类的价值构成：

（1）社会化过程

价值观的最初来源通常是社会化过程，包括家庭教育、学校规范、同伴互动与社会制度的灌输。在个体成长过程中，行为被奖励或惩罚的结果逐渐塑造了对“好”与“坏”的内化标准。Bandura 等人（1977）的社会学习理论指出，个体通过观察他人行为及其结果，习得相应的价值倾向。AI 价值观对齐中强化学习方法的核心思路与之相吻合，通过结果的反馈塑造 AI 价值标准。

（2）认知发展

发展心理学强调，价值观随认知能力的发展而逐渐复杂化。皮亚杰与科尔伯格的道德发展阶段理论指出，儿童最初的道德判断基于对惩罚与奖励的反应，随后逐步发展为理解社会契约与普遍伦理原则的能力（Piaget, 2013）。这说明人类价值观的建构性特征：并非被动接受，而是主动构建生成。AI 价值观对齐中的协作逆向强化学习就是源于这一思路，想让 AI 通过与人的互动自主构建奖励机制，但正如儿童价值观建构有时候取决于自身总结能力，通过这种方法进行 AI 价值观对齐可能会因为 AI 模型对环境敏感性的差异导致其形成不一样的价值观，要想让模型道德能力发展的更好得依赖于更强的模型性能。

（3）文化的影响

价值观高度依赖于文化背景，跨文化研究显示，人类价值系统既有有普遍结构，也体现出文化差异。例如 Schwartz（2012）的基本人类价值理论提出了十项普遍的价值维度，这些维度在不同文化中也表现出稳定性。这启发我们在进行 AI 价值观对齐设计时，无论其内生奖励机制如何构建，都应辅以符合人类普世价值观的底层符号规则，以形成基本行为约束。Haidt（2012）的道德基础理论则认为，道德判断是由进化产生的几个核心道德基础模块驱动的，但在不同文化中，基础模块的优先级与表达形式存在显著差异。这说明在 AI 价值观对齐中不存在一种唯一正确且通用的法则，AI 价值观对齐的错误泛化问题可能不仅是技术性的问题。不同 AI 系统价值观对齐的目标应该与其应用环境以及社会背景相结合，设计出与其核心职能相符合的价值观对齐规则。

此外，在现有技术手段的辅助下，有望建构出维度化、跨文化的道德认知空间。例如 Cheng 等人（2025）在研究中通过语料分析、结构建模等方法揭示了人类人际关系的通用结构。类比来看，如果能借助大数据和机器学习方法，将多种来源的道德信息整合处理，有望构建出一个通用道德认知的模型。将道德认知空间建模为若干维度，能给 AI 带来一个清晰、可解释、可学习的“目标空间”，从而更有可能进行有效的价值观对齐。

AI 价值观的构成只是价值观对齐的一部分，另一个关键问题在于：当面对道德冲突与伦理抉择时，AI 该如何做出与人类相符的判断？这就需要进一步考察人类在实际道德决策中的心理加工机制。

3.1.2 道德判断

道德判断是人类行为调节与社会互动的核心机制，理解其心理过程对 AI 价值观对齐具有重要启发意义。

“双系统理论”（dual-process theory）由 Greene 等人（2007）提出，认为人类道德决策由两个系统共同驱动：系统一为快速、隐性、情绪驱动的道德直觉系统，系统二则为缓慢、显性、基于规则的道德推理系统。这种双重加工机制对 AI 设计具有重要启示：当前多数 AI 价值观对齐仅模拟系统二的逻辑推理过程，而忽视了系统一中的感性评估机制，这可能导致 AI 做出“合理但不合情”的选择，进而展现出与人类行为的偏差。

然而随着道德心理学的发展，学界对于“双系统理论”也提出了许多批判。一方面，“双系统理论”将“义务论判断”与直觉系统、“功利主义判断”与推理系统机械对应的说法遭到质疑，实证研究表明，功利主义判断并非总是推理系统的产物，例如 Bago 和 De Neys（2019）在快速反应的认知负荷实验中发现，大量被试在极短的时间内也能做出功利主义决策。另一方面，研究者指出该理论主要聚焦于少数经典道德困境场景的讨论（如天桥问题、电车难题等），并没有对道德判断中的动机归因、社会情境理解、社会规范内化等情况进行深入考虑（Kahane, 2012）。来自认知神经科学的研究也表明道德判断不是由单一脑区或两个系统分别处理，而是由多个脑区协同参与的神经网络动态调控的结果。例如 FeldmanHall 等人（2013）发现，在处理复杂道德决策时，与社会认知相关的颞顶联合区（temporo-parietal junction, TPJ）活动增强，而腹内侧前额叶皮层（ventromedial prefrontal cortex, vmPFC）活动则减少。前者是社会认知的重要脑区；而后者则与情绪整合密切相关。这些研究结果表明，人类的道德判断具有多维结构，不仅涉及认知推理与情绪加工，还整合了社会情境信息。

除了“双系统理论”外，也有学者从不同视角解析了道德判断的心理机制。例如 Schein 和 Gray（2018）提出的“二元道德理论”（theory of dyadic morality），该理论认为一切道德判断都可归结为“加害者 - 受害者”结构的感知，强调“伤害识别”与“意图归因”在道德判断机制中的地位。这种以“伤害感知”为核心的模型更贴近人类在真实社会交互中的道德心理结构，也为 AI 道德推理的情境建模提供了新的方向。

上述研究结果都为 AI 价值观对齐提供了重要启示。目前多数 AI 系统仅模拟基于规则的逻辑推理过程，忽视了人类决策中情绪评估与情境敏感性的重要作用。若希望 AI 更接近人类价值判断方式，应当不仅在架构上整合系统一与系统二的加工机制，更需设计能够识别不同社会情境、灵活调整判断路径的情境整合模块。这种多维度、具备上下文适应能力的机制，才能增强 AI 在复杂社会交互中的可靠性与人类认同感。

综上所述，心理学关于人类价值观构成与道德判断机制的研究，为 AI 价值观对齐提供了丰富的理论启示。然而，要将这些理论真正转化为可执行的系统设计方案，还需进一步分析心理学机制在具体价值导向行为上的实现路径。

3.2 心理学促进 AI 价值观对齐的实现路径

本节将探讨心理学在各个层面对 AI 价值观对齐的贡献路径。具体而言，首先阐明心理学对人类道德直觉与行为的刻画如何为“AI 系统究竟应与何种伦理价值进行对齐”这一关键问题提供理论依据；随后，以“利他行为”为代表，分析心理学在 AI 动机生成、认知能力构建与社会演化方面的机制性指导作用。

3.2.1 心理学对 AI 价值观对齐目标的反馈作用

AI 价值观对齐的一个根本的问题在于：AI 系统应当对齐何种类型的价值观？目前越来越多研究者意识到，“What to align with”并非一个可由设计者或专家单方面决定的先验问题，回答这一问题需要经过动态的、社会性的协商和验证（Awad et al., 2022）。

因此，学界提出了“描述性伦理价值”（descriptive ethical values）——即：是否可以通过观察、归纳与建模人类真实的道德偏好，反向推理出值得对齐的伦理价值？心理学为这一过程提供了实证性的解读。人类的道德判断并非抽象规则计算的结果，而是由情绪直觉、社会规范、经验学习等多种机制共同作用的产物（Bonnefon et al., 2024）。这些判断在不同文化与情境中既表现出一定的一致性，也存在显著的差异性。例如“Moral Machine”实验采集了 200 多个国家被试在自动驾驶情境中的道德判断结果，揭示了人类道德偏好存在的共识与文化差异（Awad et al., 2018）。正因如此，心理学研究可以作为一种经验反馈机制，不断校准我们关于人类价值观的认知。

值得注意的是，心理学研究提供的只是对“人类现实偏好”的理解，而不能替代伦理学对价值正当性的规范讨论。不过，这种“经验 - 规范”双重视角的融合，正是当前 AI 伦理设计的重要趋势。心理学不仅能够识别出人们普遍重视的价值，还能揭示在特定决策环境中哪些价值最容易被边缘化或

扭曲，从而为设计具备社会适应性的 AI 对齐目标提供反馈。

3.2.2 心理学对 AI 价值观对齐机制的指导作用 : 以利他行为机制为例

利他行为（altruism）作为人类社会价值的典型体现，长期以来是伦理学、社会心理学与认知科学等领域的研究重点。如何让 AI 系统展现出近似人类的利他倾向也是 AI 价值观对齐的核心问题之一。之所以聚焦于利他行为，主要基于以下三点考虑：其一，利他行为是人类价值观体系中的核心组成，体现了个体在无外部激励下自愿为他人利益做出牺牲的能力，这种行为是 AI 伦理设计中非常能体现“价值观对齐水平”的指标之一；其二，利他行为的心理学研究范围涵盖了价值观形成(如社会规范内化)、道德判断（如情绪直觉与推理交互）以及社会环境反馈（如规范惩罚与群体合作演化）等多个维度，能够较好地展现不同维度的心理学机制在 AI 对齐中的应用；其三，利他行为在心理学、神经科学与计算建模中已有丰富的研究结果积累，且在 AI 领域已有关于如何实现 AI 利他行为的研究作为现实参考。

因此本节将从利他行为的动机机制、认知能力机制与社会行为演化机制三个层面，探讨如何在 AI 中引入类人的利他行为生成机制，并展示心理学理论在这一过程中的指导作用。

（1）动机机制层面：构建类人的多重驱动系统

传统 AI 系统在设计过程中常把利他行为机制简化为计算社会效用最大化的输出，主要依赖外部设定的奖励函数。然而，心理学研究表明，人类利他行为背后存在复杂的心理与神经机制，并非可以用简单奖励结构还原。Wu 等人 (2024) 提出“动机鸡尾酒”计算模型，指出人类的利他行为收到情绪共鸣（emotional empathy）、内在道德信念（moral belief）、名誉管理（reputation concern）、预期互惠（expected reciprocity）等多重动因驱动，不同情境下的利他决策由多重动因动态调控。这提示我们，在 AI 系统中应构建类似的多元内驱结构，避免将道德行为归因于单一动因或外部奖惩。

其中，情绪共鸣（emotional empathy）被认为是激发直接利他行为的核心驱动之一。Batson（2011）提出“共鸣 - 利他假说”（empathy-altruism hypothesis），认为当个体体验到对他人痛苦的共情情绪时，会产生一种内在动机去帮助他人，以缓解自身的情绪不适。这种非功利性动因基于情绪而非理性计算，是无条件亲社会行为的重要心理学基础。为在 AI 中模拟这一机制，Zhang 等人（2023）提出一种脑启发建模方法，在脉冲神经网络（Spiking Neural Network, SNN）架构中构建“情绪响应路径”。在该模型中，研究者模拟了人脑边缘系统(如杏仁核) 对社会性刺激的即时响应，使 AI 在感知他人情绪时生成类人“感性动因”，从而增强其亲社会行为的自然性。将“情绪共鸣”机制引入 AI，不仅可以扩展其动机结构，还可能在缺乏外部奖励的道德困境中提供“非功利性”的亲社会决策基础，有助于构建更可信赖、内驱稳定的道德行为模式。

（2）认知能力层面：赋予 AI“心理理论”与自我想象能力

根据“心理理论”（theory of mind）观点，人类进行利他行为的一个重要前提是能够理解他人的情绪与意图，形成“对他人心智状态的模拟”。这一能力使个体可以预测他人需求，并在缺乏明确奖励的情况下主动做出亲社会选择。发展心理学研究指出，儿童在发展出心理理论能力后，其利他行为与社会判断能力显著提升（Wellman, 2014）。通过推断他人情绪状态与心理需求，个体更容易在社会互动中展现出体贴、援助等亲社会行为。据此观点 Tong 等人（2024）提出，通过自我想象（self-imagination）和类“心理理论”建模，AI 能够在缺乏明确外部激励的情境中，自发产生利他行为。

在 AI 系统中赋予“心理理论”能力，不仅有助于提升其在多智能体互动中的行为适应性，更使其能够在道德决策中考虑“他人视角”与“预期影响”，从而避免行为上的冷漠和局限性。

（3）社会行为演化层面：在动态环境中形成稳定的亲社会倾向

除了内在动因与认知机制外，心理学研究还强调社会环境对利他行为的塑造作用。研究表明人类的亲社会倾向并非总是稳定的，而是在特定社会互动情境中逐步演化而成的。例如 Wu 等人（2024）发现在存在不确定和规范惩罚的社会环境中，个体更可能发展出稳定的亲社会行为倾向。Efferson 等

人（2024）发现重复互动和群体间偏好是利他行为稳定存在的重要原因。

从进化心理学视角来看，亲社会行为的出现并非因个体“天生利他”，而是基于合作所带来的长期适应优势。文化进化理论也指出，“间接互惠”与“社会学习”可以促使利他行为的传播与内化（Boyd & Richerson, 2009）。

因此，在 AI 系统设计中，应引入多智能体交互与动态反馈机制，构建复杂、逼真的模拟环境，使 AI 在不断交互中学习行为的社会后果与群体规范。这类环境可以包括道德惩罚机制、社会声誉系统以及重复博弈过程，从而促使 AI 逐步内化亲社会行为模式，而非仅停留在一次性决策的表面匹配。AI 价值对齐应强调交互性与演化性，通过社会环境模拟促进其形成稳定的亲社会行为倾向。

综上所述，心理学可以通过四种路径融入 AI 价值观对齐工作中（图 1）：在目标层面，心理学实证研究可以为 AI 价值观对齐目标提供反馈性依据，辅助对齐目标的动态完善。在动机层面，可借助人类多重动因模型建立 AI 的多元内驱系统，让 AI 明确为什么这样做。在认知能力层面，可以根据人类道德判断中涉及的认知过程去扩展 AI 的相应能力，让它具有道德判断的能力基础（比如长短期记忆、心理理论能力等）。在社会环境层面，依据社会行为演化规律的研究结果，通过模拟现实复杂社会环境，使 AI 在交互中逐步演化出稳定的亲社会倾向。

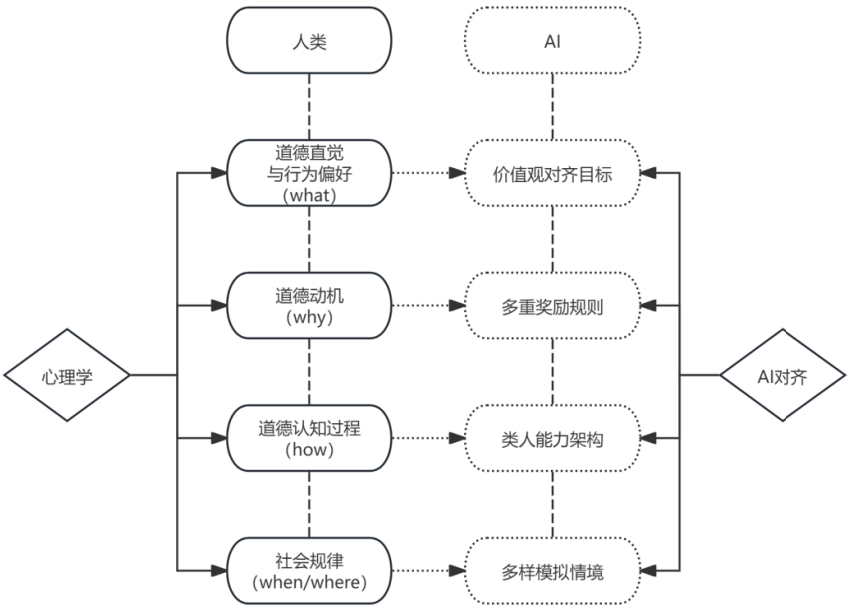

图 1 心理学促进 AI 价值观对齐的四重路径

总的来说，基于心理学理论的 AI 价值观对齐强调为 AI 构建类人内在机制，使 AI 的价值判断和道德决策过程不再是简单的奖励计算过程，而是更具有人性化的分析过程。这种由内而外改善 AI 的思路或许能让 AI 更加符合人类价值观的“思维方式”，从而让其拥有更具适应性的道德表现。

4 结论与展望

AI 价值观对齐不能仅依赖算法优化，而应嵌入类人的价值判断机制。从心理学视角出发，可以更全面地了解人类动机，促进 AI 系统和虚拟代理的发展，让它们表现出类似人类的社会行为，做出符合人类价值观的决策（Blazek et al., 2024）。为此，本文提出了心理学在目标设定、动机构建、认知能力扩展与社会演化四个方面对 AI 价值观对齐的赋能路径。

面向未来，心理学与 AI 的结合应该更加注重心理学机制在算法层面的实现。一方面，应加强心理学理论在 AI 架构中的形式化建模，将诸如元认知、情绪调节等机制内嵌于 AI 系统设计架构之中；另

一方面，心理学研究也需要在理论层面实现机器可用的转译，尝试用更加数学量化的形式去表征心理学理论，从而推动心理理论在算法设计中的应用。基于上述思路，未来研究应重点探索以下方向：

（1）建立包含多文化视角的价值偏好数据库，为不同场景下的 AI 系统提供可调整、可适应的价值参考依据。（2）引入多种人类动因（如共情、声誉、互惠等）构建 AI 的多元内在驱动系统，使其行为不再完全依赖外部奖励。（3）加强 AI 系统认知能力的构建，引入“元认知”、“长短期记忆”等多种认知模块，使其道德决策获得基础推理能力的支撑。（4）建立多维度、多主体的心理学评估范式，如“AI 道德测评”、“利他实验沙盒”等，系统评估 AI 系统的价值表达质量。

人工智能的发展不仅代表着技术的演进，更是人类智能的外化与延伸。要想设计出好的人工智能就应该先理解人类智能。从深入理解人类之“心”出发去思考如何构建“智”，才能最终创造出与人类心智相通的 AI。这不仅是技术路线的选择，更是对“何谓智慧”这一根本问题的回应与探索。

参考文献

梁思源，何英哲，刘艾杉，李京知，代朋纹，操晓春 . (2024). 面向大语言模型的越狱攻击与防御综述 . *信息安全学报*，*9*(1), 1–20.

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proceedings of the 21st International conference on machine learning* (pp. 1–8). ACM.

Alsaleh, R., & Sayed, T. (2020). Modeling pedestrian–cyclist interactions in shared space using inverse reinforcement learning. *Transportation Research Part F: Traffic Psychology and Behaviour, 70,* 37–57.

Arora, S., & Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence, 297,* 103500.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64.

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A. C., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., Mikhail, J., Opoku–Agyemang, K., Schaich Borg, J., Schroeder, J., Sinnott–Armstrong, W., Slavkovik, M., Tenenbaum, J. B. (2022). Computational ethics. *Trends in Cognitive Sciences, 26*(5), 388–405.

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General, 148*(10), 1782–1801.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback.* arXiv.

Bandura, A., & Walters, R. H. (1977). *Social learning theory* (pp. 141–154). Prentice–Hall.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y.,& Fung, P. (2023). *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity.* arXiv.

Batson, C. D. (2011). *Altruism in humans.* Oxford University Press.

Blazek, P. J., Venkatesh, K.,& Lin, M. M. (2024). Automated discovery of algorithms from data. *Nature Computational Science, 4*(2), 110–118.

Bonnefon, J.–F., Rahwan, I., & Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology, 75*(1), 653–675.

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences. 364*(1533), 3281–3288.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S.,... Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation.* arXiv.

Carlsmith, J. (2022). *Is power–seeking AI an existential risk?.* arXiv.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R.,⋯ Sadigh, D. (2023). *Open problems and fundamental limitations of reinforcement learning from human feedback.* arXiv.

Cheng, X., Popal, H., Wang, H., Hu, R., Zang, Y., Zhang, M., Thornton, M., Ma, Y., Cai, H., Bi, Y., Reilly, J., Olson, I. R.,& Wang, Y. (inpress). The conceptual structure of human relationships across modern and historical cultures. *Nature Human Behaviour.*

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S.,& Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* (pp. 4299–4307). Curran Associates, Inc.

Clark, J., & Amodei, D. (2016). *Faulty reward functions in the wild.* OpenAI.

Davies, P. (2022). *AI cyber attacks are a 'critical threat'. This is how NATO is countering them.* Euronews. next.

Efferson, C., Bernhard, H., & Fehr, E. (2024). Super–additive cooperation. *Nature, 626*(8001), 1034–1041.

Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., Zuiderveen Borgesius, F. J., & Biega, A. J. (2025). Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology, 16*(1), 1–54.

FeldmanHall, O., Mobbs, D., & Dalgleish, T. (2013). Deconstructing the brain's moral network: Dissociable functionality between the temporoparietal junction and ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 9*(3), 297–306.

Fu, J., Singh, A., Ghosh, D., Yang, L.,& Levine, S. (2018). Variational inverse control with events: A general framework for data–driven reward definition. *Advances in Neural Information Processing Systems* (pp. 8547–8556). Curran Associates, Inc.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437.

Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9,* 345–359.

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual–process theory of moral judgment explains. *Trends in Cognitive Sciences, 11*(8), 322–323.

Grusec, J. E., & Hastings, P. D. (Eds.). (2006). *Handbook of socialization: Theory and research.* The Guilford Press.

Hadfield–Menell, D., Russell, S. J., Abbeel, P.,& Dragan, A. D. (2016). Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 3909–3917). Curran Associates, Inc.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion.* Pantheon Books.

Jiang, L., Hwang, J. D., Bhagavatula, C., Le Bras, R., Liang, J. T., Levine, S., Dodge, J., Sakaguchi, K., Forbes, M., Hessel, J., Borchardt, J., Sorensen, T., Gabriel, S., Tsvetkov, Y., Etzioni, O., Sap, M., Rini, R.,& Choi, Y. (2025). Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence, 7*(1), 145–160.

Kahane, G. (2012). On the wrong track: Process and content in moral psychology. *Mind and Language, 27*(5), 519–545.

Koch, J., Langosco, L., Pfau, J., Le, J., & Sharkey, L. (2021). Objective robustness in deep reinforcement learning. *Proceedings of the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.*

Li, X., Zhou, R., Lipton, Z. C., & Liu, L. (2024). *Personalized language modeling from personalized human feedback.* arXiv.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K.,& Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.

Morris, M. R., Sohl–Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023). *Levels of AGI for operationalizing progress on the path to AGI.* arXiv.

Mu, T., Jetten, A., & Brunskill, E. (2020). Towards suggesting actionable interventions for wheel–spinning students. In A. N. Rafferty, J. Whitehill, V. Cavalli–Sforza,& C. Romero (Eds.), *Proceedings of the 13th International conference on educational data mining (EDM 2020)* (pp. 183–193). International Educational Data Mining Society.

Ngo, R. (2020). *AGI safety from first principles.* AI Alignment Forum.

OpenAI. (2023). *GPT–4 technical report*. arXiv.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback.* arXiv.

Phelps, S., & Russell, Y. I. (2023). *Investigating emergent goal–like behaviour in large language models using experimental economics.* arXiv.

Piaget, J. (2013). *The moral judgment of the child* (Original work published 1932). Routledge.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.,& Ng, A. Y. (2017). *CheXNet: Radiologist–level pneumonia detection on chest X–rays with deep learning.* arXiv.

Ramachandran, D., & Amir, E. (2007). B*ayesian inverse reinforcement learning. Proceedings of the 20th International joint conference on artificial intelligence* (IJCAI–07) (pp. 2586–2591). IJCAI.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture, 2*(1), 11.

Shao, H., Cohen, L., Blum, A., Mansour, Y., Saha, A., & Walter, M. (2023). Eliciting user preferences for personalized multi–objective reinforcement learning through comparative feedback. *NeurIPS 2023.*

Si, W. M., Backes, M., Blackburn, J., De Cristofaro, E., Stringhini, G., Zannettou, S., & Zhang, Y. (2022). Why so toxic? Measuring and triggering toxic behavior in open–domain chatbots. *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security* (CCS '22), 1–15.

Tong, H., Lu, E., Sun, Y., Han, Z., Liu, C., Zhao, F., & Zeng, Y. (2024). *Autonomous alignment with human value on altruism through considerate self–imagination and theory of mind.* arXiv.

Wellman, H. M. (2014). *Making minds: How theory of mind develops.* Oxford University Press.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science, 131*(3410), 1355–1358.

Wu, X., Ren, X., Liu, C.,& Zhang, H. (2024). The motive cocktail in altruistic behaviors. *Nature Computational Science, 4,* 659–676.

Zhang, B., Liang, P., Zhou, X., Ahmad, A.,& Waseem, M. (2023). *Practices and challenges of using GitHub Copilot: An empirical study.* arXiv.

From Human Mind to Artificial Intelligence: Advancing AI Value Alignment Through Psychological Theories

Jin Shaoxiong[1,2,3], *Liu Chao*[1,2,3]

（[1] State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, 100875）

（[2] Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, Beijing, 100875）

（[3] Beijing Key Laboratory of Safe AI and Superalignment, Beijing Normal University, Beijing, 100875）

Abstract In recent years, the field of artificial intelligence (AI) has witnessed unprecedented growth, characterized by major advancements in cognitive intelligence, perceptual processing, and decision-making capabilities. These technological breakthroughs have driven the widespread adoption of AI systems across a wide range of sectors, including healthcare, education, finance, and transportation. As a result, AI has become instrumental in improving operational efficiency, enhancing accuracy, and fostering innovation. There is little doubt that such developments have significantly boosted human productivity and convenience.

However, the increasing sophistication and autonomy of AI technologies have also introduced a variety of societal risks and ethical concerns. Among the most pressing of these are challenges related to AI safety and the alignment of AI behavior with human values. For instance, AI systems have been found to perpetuate bias in recruitment decisions, produce offensive or harmful content during interactions with users, and even pose existential threats in high-stakes domains such as autonomous weapons. These examples reflect growing anxieties about the potential misalignment between AI behavior and the ethical principles upheld by human societies. If left unaddressed, such misalignment could lead to consequences that undermine social trust and moral norms.

In response to these challenges, the concept of AI value alignment has emerged as a central concern within the broader field of AI safety research. AI value alignment refers to the development of AI systems whose goals, behaviors, and decision-making processes are consistent with the values, preferences, and ethical standards of individuals or society as a whole. Technically, several value alignment methodologies have been proposed, including reinforcement learning from human feedback (RLHF), inverse reinforcement learning (IRL), and constitutional AI. These approaches aim to incorporate normative constraints into the training process, thereby steering AI systems toward behavior that is both desirable and predictable. While promising in many respects, such methods face significant limitations. In particular, aligned AI systems often exhibit reduced adaptability when faced with novel scenarios and suffer from poor interpretability, making it difficult to trace or understand the reasoning behind their decisions. These limitations highlight the insufficiency of a purely engineering-driven approach and suggest the necessity of incorporating broader, interdisciplinary perspectives.

One promising approach is to integrate insights from psychology, the scientific study of human behavior, cognition, and moral reasoning, into the research and development of AI value alignment. Psychological theories provide robust conceptual tools for understanding how humans construct values, make moral judgments, and resolve ethical dilemmas in complex social contexts. Rather than designing AI systems that merely replicate the surface-level patterns of human behavior, these insights can inform architectures that embody internal mechanisms analogous to those involved in human moral cognition. Thus, true value alignment requires more than behavioral mimicry; it demands a form of cognitive and ethical compatibility between artificial agents and the human mind, particularly in terms of value judgment and moral decision-making processes.

This paper explores how psychological science can contribute to advancing AI value alignment. It reviews core psychological theories concerning the formation of moral values, dual-process models of moral reasoning, and the roles of emotion and social context in ethical decision-making. Building on these foundations, we propose conceptual frameworks that include the construction of a unified moral cognitive space capable of integrating diverse human values, and the development of dual-system moral architectures that emulate the interaction between intuitive and deliberative reasoning in human moral cognition. To ground these ideas in practice, we use altruistic behavior—a central and complex moral phenomenon—as a case study, examining how its psychological underpinnings could be modeled in AI systems to promote socially aligned decision-making.

By bridging AI safety research with psychological theory, this work seeks to support the development of more interpretable, robust, and ethically aware AI systems. Such interdisciplinary integration is not only timely, but also essential to ensure that the evolution of AI technologies remains aligned with the fundamental values of human society.

Key words artificial intelligence, AI value alignment, value alignment, theory of mind, moral decision-making, altruism