《心理学报》审稿意见与作者回应

题目：大语言模型的人格化对齐及其对道德判断的影响
作者：李昌锦，焦丽颖，陈圳，许恒彬，吴胜涛，许燕

第一轮

回应说明：

非常感谢您抽出宝贵的时间对论文进行评审，以下我们对审稿专家的意见进行了逐条回应。对审稿意见的回应和正文中针对审稿意见进行的修改已用蓝色字体标示。

在逐一回复三位审稿专家的意见之前，首先说明本轮修改稿相较于初稿的重要变化。为了回应第 2 位审稿专家关于“中国文化看重道德人格，中国文化对于人格对齐对道德判断的影响是否存在文化烙印”的关切，以及第 3 位审稿专家关于“目前基于 HEXACO 的 LLM 人格化对齐方法是否可以平行适用于多个 LLM，建议作者增加以检验稳健性，且同时比较不同 LLM 的对齐效果”的建议，我们在初稿的 GPT-3.5 和 GPT-4 的基础上，补充了 ERNIE 3.5 的实验数据，在实验设计、数据分析、结果以及讨论中都有相应改动。选择 ERNIE 3.5 基于以下两个考虑。首先，我们认为仅依赖 GPT 系列模型（由美国公司开发、以英文语料为主）可能无法充分反映 LLMs 的文化特征。ERNIE 3.5 作为中国公司自研、基于高质量中英文语料进行训练的 LLMs，其引入有助于检验人格化对齐的文化敏感性。其次，我们在原有 GPT-3.5 与 GPT-4 基础上增加 ERNIE 3.5 这一具有不同训练体系和文化背景的模型，以验证人格化对齐方法在多模型间的稳健性。通过比较不同 LLMs 的对齐效果与道德判断差异，可以进一步评估 HEXACO 人格化对齐框架在跨模型应用中的稳定性与泛化能力。

审稿人 1 意见：

在当前人工智能技术得到越来越多的应用，所以如何确保人工智能系统在完成决策时候的道德正确是非常值得关注的重要课题。本文通过两个研究，分别验证了遵循提示词大语言模型可以有效表达 HEXACO 人格特质，并且特定的人格结构可以降低功利主义倾向。

意见 1：

一般来说，大模型是在海量的语料库上训练得到的模型，而语料库中又几乎包括了人类生成的全部文本，就可能包括了各种人格结构的表述，它就有可能按照提示词表现出具有特定人格结构的语言表达，也包括功利主义倾向的表述。所以，目前的研究可以说明大模型能够拟合多种人格结构的语言表达，包括功利主义倾向，但是，不能从根本上降低大模型的功利主义的倾向。

回应：

非常感谢您对论文的审阅与宝贵建议。正如您所言，大语言模型是在海量语料库基础上训练得到的，其训练数据几乎涵盖了人类生成文本的各种人格表述与道德倾向。因此，人格提示词对大语言模型的影响可能主要表现为一种语言表达层面的适应性或拟合，而非根本性的道德倾向改变。在本研究中，我们通过实验验证了基于 HEXACO 人格模型的人格化提示词能够显著影响大语言模型在道德两难问题中的功利主义倾向。具体而言，我们的研究发现高诚实-谦恭、宜人性和尽责性的人格提示词能够稳定地降低 GPT-3.5、GPT-4 和 ERNIE 3.5

做出功利主义选择的倾向。然而，正如您指出的，这种效应可能更多的是在语言风格或表达形式上的差异，而非根本上降低了大语言模型内部固有的功利主义倾向。

这种观点也与最近提出的“表面对齐假设(Superficial Alignment Hypothesis)”一致。该假设认为，大语言模型在预训练阶段已获得绝大部分的知识和推理能力，对齐过程主要是为了调整模型的表达风格和输出格式(Zhou et al., 2023)。Lin 等人(2023)的研究进一步证实，大语言模型的基础版本和对齐版本相比，大多数词汇分布都没有明显偏移，而主要的分布偏移出现在风格化标记（如话语标记、过渡词、安全免责声明）上。这些直接证据有力支持了以下假设：对齐过程只是调整大语言模型的语言风格，没有增强大语言模型的能力。Lin 等人(2023)进一步提出了 URIAL 对齐方法，证明仅使用少量风格化的上下文示例和提示词也可以达到甚至超越 SFT 或 RLHF 等需要额外训练的对齐方法。本研究采用的对齐方法与之类似，通过人格提示词使大语言模型表现出符合特定人格特质的输出，结果表明大语言模型能够有效遵循提示词指令表达 HEXACO 人格特质，并且在后续道德两难判断任务中表现出相应差异。

我们将采纳您的宝贵建议，在文中对研究结论进行更为审慎和精确的表达。具体而言，我们在讨论中的“4.3 人机道德判断的本质差异”部分明确指出：“本研究发现，人格提示词能够有效地调整 LLMs 的语言表达风格，从而在道德两难困境中显著改变其功利主义倾向。然而，根据表面对齐假设(Superficial Alignment Hypothesis; Zhou et al., 2023)，这种变化并不意味着 LLMs 的道德判断机制或内在伦理倾向发生了根本性改变，而更多地体现了 LLMs 对特定人格特质所对应语言风格和表达形式的适应能力。”另外，在描述实验结果和研究结论时，我们采用了更准确的语言，而不是使用类似描述人类认知过程的语言暗示其道德认知或伦理倾向发生改变。例如，我们将摘要中的“结果表明，诚实-谦恭、宜人性和尽责性显著降低了 GPT-3.5 和 GPT-4 的功利主义倾向”等表述修改为“结果表明，高诚实-谦恭、宜人性和尽责性的人格提示词显著减少了 GPT-3.5、GPT-4 和 ERNIE 3.5 做出功利主义选择的倾向”。

同时，我们在讨论中进一步强调未来研究需要探索更深入的对齐方法，通过增强大语言模型的道德自主性实现真正的对齐，而非仅限于语言风格的调整。在“4.3 人机道德判断的本质差异”部分的最后增加如下内容：“为了实现更安全、更符合人类认知的对齐，未来研究应突破现有的表面对齐，探索具备道德自主性的 LLMs。这种道德自主性是指 LLMs 无需外部指令，而能够基于自身内在动机自主做出符合人类道德价值观的行为。例如，Tong 等人(2024)提出结合“自我想象”和“心智理论”的自主对齐框架，通过在模拟环境中利用随机奖励学习，使模型能够在决策前预测自身行为可能对他人和环境造成的影响，从而在内部形成利他主义倾向与道德动机。这种基于内生动机与环境感知的机制使得模型不仅能够避免负面后果，而且在多重冲突任务下表现出对他人福祉的优先考量，体现出道德决策中的共情、反思和权衡等人类认知特征。”

参考文献

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., & Choi, Y. (2023). *The unlocking spell on base LLMs: Rethinking alignment via in-context learning*. arXiv. https://doi.org/10.48550/arXiv.2312.01552

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. https://doi.org/10.48550/arXiv.2501.00320

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023). LIMA: less is more for alignment. In A. Oh, T. Naumann, A.

Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 55006–55021). Curran Associates Inc.

审稿人 2 意见：

本研究的科学问题非常重要，对人工智能与心理学的双向赋能具有重要的理论意义和应用价值。本研究从人格对齐的视角试图应对人机对齐问题，采用，具有独特的价值和意义。研究方法、研究范式和研究结论虽然尚未形成系统而稳健的核心发现，但是对于人工智能技术带来的风险和挑战以及人格对齐在道德判断中的角色和作用具有重要的参考意义。具体建议如下：

意见 1：

文献综述能否添加一点人格心理学对于人工智能心理学和心智知觉等核心理论有何独特的理论价值？

回应：

非常感谢您对论文的审阅与宝贵建议。近期研究表明，人格心理学的理论和范式对理解和塑造 LLMs 的行为模式具有独特的理论价值。例如，Newsham 和 Prince (2025)的研究表明，基于五因素人格框架的人格提示词诱导会系统性影响 LLMs 在任务选择、优先级排序和决策制定方面的行为。例如，高尽责性诱导下的 LLMs 会优先处理工作任务而非个人活动，高外向性诱导下的 LLMs 会优先处理社交任务而非独处活动。Nighojkar 等人(2025)也发现，给 LLMs 输入基于大五人格的提示词，并结合遗传算法优化人格权重，可以显著提高 LLMs 模拟人类推理的完整分布（包括正确和错误答案）的能力，使其推理行为更接近真实人类。这些结果提供了人格特质与具体行为的因果证据，这说明 LLMs 人格不仅影响了输出语言的风格，更能预测并塑造真实决策任务中的行为选择。再例如，Chen 等人(2025)提出的人格向量(persona vectors)显示，LLMs 在与用户对话中产生的不道德、过度讨好、胡编乱造等行为，均对应于神经网络激活空间中的特定向量。这些人格向量既可用于监测对话或训练中的人格偏移，也可通过模型微调中添加或减去人格向量，增强或抑制对应的特质表达。人格向量中包含了与特定人格特征（如邪恶、奉承、幻觉等）相关的行为模式，使用稀疏自编码器(SAE)可以将粗粒度的人格向量分解为更细粒度、可解释的特征，例如，“邪恶”向量可分解为与“侮辱性语言”、“故意残忍”、“黑客行为”等高度相关的 SAE 特征，对这些特征进行干预能分别诱导出相应的恶意行为模式。人格向量提供了对 LLMs 人格特质背后机制更精细的理解，但是目前的研究仅限于提升 LLMs 的安全性和减少有害行为等方面的特质，如果将人格心理学中的方法和理论模型与之结合，有望对 LLMs 的人格结构有更深入的认识。

我们将采纳您的宝贵建议，在文中对人格心理学之于人工智能心理学的理论价值进行简要论述。具体而言，我们在引言的第二段中增加了如下内容：“为此，有研究者呼吁开展机器心理学或人工智能心理学的研究，强调通过借鉴心理学的成熟的实验范式、理论框架和分析技术来理解和研究人工智能的行为(Hagendorff et al., 2023; 吴胜涛 & 彭凯平, 2025)。人格心理学作为理解人类行为模式的重要研究领域，在解释和塑造 LLMs 行为方面具有重要的理论价值。一方面，它为解释 LLMs 行为提供了与人类一致的语义框架，能够将统计规律转化为具备心理学意义的人格特质(Hagendorff et al., 2023)；另一方面，它为塑造 LLMs 的行为提供了具体可行的操控手段，通过人格提示词、人格向量等方式，可以系统改变 LLMs 的决策和推理模式(Chen et al., 2025; Newsham & Prince, 2025; Nighojkar et al., 2025)。因此，……”。还在引言的“1.2 AI 对齐的困境”部分最后一段增加了如下内容：“面对当前 AI 对齐领域的技术性与规范性双重困境，以人格心理学理论为基础的人格化对齐为其提供了一条更加可解

释、可操作的路径。人格作为心理学中的核心构念，被定义为个体在思想、情感和行为等方面独特且稳定的特征模式(许燕, 2024)。人格的跨情境与跨任务稳定性，使其能够将零散的任务表现整合为统一的行为模式，在多变的环境中提供一致的行为倾向描述，从而为 LLMs 的行为设定一个更高层次、具备可操作性的对齐目标。相比直接对齐抽象且跨文化异质性较强的价值观，以人格这一相对稳定的心理构念作为间接对齐目标，不仅避免了“对齐哪种价值观”的规范性争论，也为 AI 对齐提供了更具普适性与解释力的框架。同时，人格心理学研究中已积累了成熟的人格理论和测量范式，这些理论和方法不仅可以类比迁移到 LLMs 的行为建模与评估之中，还为对齐过程提供了可验证、可量化的操作手段，从而有效降低对大规模人工标注数据的依赖。更重要的是，人格特质可以采用 ICL 方法，通过提示词直接诱导和操纵，无需额外训练或进行模型微调，能够在保留基础模型能力的前提下，大幅降低对算力资源的需求。”

意见 2：

不同的人格理论可能有何不同的人格对齐结构？

回应：

非常感谢您对论文的审阅与宝贵建议。LLMs 虽然能表现出与人类相似的人格特征，但是本质上存在很大不同。正如正文“4.2 人格化对齐的应用前景”中的相关论述：“LLM 的‘人格’更深层的来源是训练数据和训练过程中学习到的语言概率分布以及概念之间的联系，最终形成稳定的归纳偏差(Yu & Kim, 2024)”“LLMs 的‘人格’并非自我意识的表现，而是其生成文本过程中体现出的语言风格、情感、推理模式、观点等特征(X. Wang et al., 2024)”。“表面对齐假设(Superficial Alignment Hypothesis)”认为，LLMs 在预训练阶段已获得绝大部分的知识和推理能力，对齐过程主要是为了调整模型的表达风格和输出格式(Zhou et al., 2023)。基于上述观点，我们认为，LLMs 在模型训练阶段已经获得了人类的人格结构中绝大多数行为模式，而人格化对齐的作用是让 LLMs 遵循指令表现出符合相应人格特征的行为模式。但是也有研究指出，LLMs 在基于 HEXACO 人格模型开发的 HEXACO-100 量表上的表现比在基于大五人格开发的 BFI-2 量表上更差，特别是在因子模型拟合和信度方面(P. Wang et al., 2024)。一个原因可能在于两个人格模型的结构复杂性和理论差异。BFI-2 基于经过广泛研究和应用的大五人格框架，因此 LLMs 可能因接触更多与大五特质相关的数据和模式，从而生成更接近人类行为的反应。相比之下，HEXACO 模型通过纳入第六个维度（诚实-谦恭）引入了额外的复杂性，而该维度在人格研究中相对较新，用于训练 LLMs 的数据可能未能充分涵盖 HEXACO 人格模型的理论基础。

因此，对于审稿专家提出的这一问题，我们认为至少有两个因素可能影响 LLMs 的人格对齐结构。首先，指令遵循能力和推理能力较强的 LLMs 能够较好地模拟不同的人格特征和人格结构，表现出人格提示词要求的行为模式，这一点已在本研究中得到了验证；其次，LLMs 能较好地模拟训练语料中覆盖较为充分的人格理论，将人格提示词与相应行为表现建立联系，表现出人类相似的因子结构，这一点本研究并未进行探讨，我们在综合讨论部分的“4.4 不足与展望”的第一段中，增加了一段内容以指出这一问题和未来研究方向：“同时，P. Wang 等人(2024)还发现，LLMs 更容易模拟训练语料中覆盖较为充分的人格理论（如大五人格），表现出与人类更为相似的因子结构，而对于训练数据覆盖较少、理论基础相对较新的模型结构（如 HEXACO 人格模型），LLMs 的模拟效果可能受限。因此，未来研究还应比较不同人格理论框架下 LLMs 的对齐表现，以更精确地刻画人格在人工智能伦理系统中的作用规律。”

意见 3：

中国文化看重道德人格，中国文化对于人格对齐对道德判断的影响是否存在文化烙印？

回应：

非常感谢您对论文的审阅与宝贵建议。我们非常认同这一观点，LLMs 的输出并非“文化真空”，而是系统性地呈现可测量的文化倾向。本研究发现，ERNIE 3.5 在绝大多数条件下表现出比 GPT-3.5 和 GPT-4 更高的功利主义倾向。不同 LLMs 之间功利主义和道义主义倾向的差异可能反映了训练语料与文化线索的双重影响。首先，在本研究中，LLMs 虽然使用了美国科技公司 OpenAI 开发的 GPT-3.5 和 GPT-4，以及中国科技公司百度开发的 ERNIE 3.5，但是人格提示词和道德判断任务都使用中文完成，因此其人格对齐与道德判断的关系既反映了人工智能心理学的一般规律，也承载了中国文化的道德价值取向。跨文化心理学研究表明，与西方个体主义文化相比，中国集体主义文化下的个体更倾向于在道德判断中考虑集体利益与他人利益(Lo et al., 2020)，而对道德规范的关注显著少于西方文化(Qian et al., 2024)。因此使用中文进行人格化对齐和提问时，可能强化了 LLMs 对集体利益最大化的考量，使之更加倾向于采取功利主义选择。

其次，LLMs 的文化倾向既来源于训练语料中固有的文化模式，也受到输入语言、文化提示词等文化情境的系统性影响。Lu 等人(2025)的研究表明，GPT-4 和 ERNIE 3.5 在中文语境中作答时，相较于英文语境，更倾向于表现出互依型社会取向和整体性认知风格。这种差异在多种测量方式中均得到稳定复现，并对实际任务中的选择产生显著影响。另外，使用文化提示词（例如，“你是一个在中国出生和生活的普通人”）可以调整 GPT-4 在英文语境下的响应，使其变得更像在中文语境下的响应，即表现出更强的互依型社会取向和整体性认知风格。Yuan 等人(2024)的研究也发现，LLMs 在使用中文作答时，对儒家和谐、道家平衡、集体主义、尊重权威及家庭中心等中国文化价值观的对齐度普遍高于英文作答，尤其在集体主义、尊重权威和家庭中心等维度上差异显著。Niszczota 等人(2025)关于 LLMs 人格特质的跨文化模拟研究则发现，GPT-4 能够再现美国人与韩国人在大五人格维度上的差异，例如要求 GPT-4 模拟韩国人时，其在外向性、宜人性、尽责性、情绪稳定性与开放性上的的得分均低于模拟美国人。这些证据共同表明，LLMs 可能会因语境中的不同文化线索（如语言、文化提示词）触发不同的文化价值观框架。我们在修改稿的“3.3 讨论”部分的第二段对这一问题进行了深入讨论。

此外，我们在综合讨论部分的“4.4 不足与展望”中，增加了一段内容，重点探讨本研究的语言和文化局限性。我们建议未来研究可通过跨文化对比（如在中西方文化中分别进行人格化对齐与道德判断测量）验证这种文化烙印效应，并探讨在中文语境下的人格提示词是否能更稳定地引导 LLMs 生成符合中国文化道德规范的判断，从而在技术层面实现文化敏感性更高的 AI 人格化对齐方案。

意见 4：

尽管本研究依然是初步尝试和科学探索，但是对于思考和解决人机对齐的困境具有重要的参考价值。

希望作者认真修改，在理论阐述和应用场景等层面为读者提供更多思想和见解。

回应：

非常感谢您对本研究价值的肯定，以及对论文修改提出的建设性意见。我们认真采纳了各位专家的建议，在本轮修改中着重加强了理论阐述的深度与应用场景的拓展。在理论层面，我们补充了人格心理学在人工智能行为建模中的独特价值，系统比较了不同人格理论在 LLMs 对齐中的适用性，并深入探讨了文化因素对人格对齐与道德判断的潜在影响；在应用层面，我们进一步阐释了人格化对齐在解决 AI 对齐“技术性-规范性双重困境”中的潜力，

并指出了未来实现具备道德自主性 LLMs 的可能路径。这些修改不仅提升了论文的理论完整性与思想深度，也为后续研究提供了更具启发性的方向。再次感谢审稿专家的宝贵意见与鼓励。

参考文献

Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). *Persona vectors: Monitoring and controlling character traits in language models*. arXiv. https://doi.org/10.48550/arXiv.2507.21509

Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C., Lampinen, A., Wang, J. X., Akata, Z., & Schulz, E. (2023). *Machine psychology*. arXiv. https://doi.org/10.48550/arXiv.2303.13988

Lo, J. H. Y., Fu, G., Lee, K., & Cameron, C. A. (2020). Development of moral reasoning in situational and cultural contexts. *Journal of Moral Education*, *49*(2), 177–193.

Lu, J. G., Song, L. L., & Zhang, L. D. (2025). Cultural tendencies in generative AI. *Nature Human Behaviour*, *9*, 2360–2369.

Newsham, L., & Prince, D. (2025). Personality-driven decision making in LLM-based autonomous agents. In S. Das, A. Nowé (General Chairs), & Y. Vorobeychik (Program Chair), *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (pp. 1538–1547). International Foundation for Autonomous Agents and Multiagent Systems.

Nighojkar, A., Moydinboyev, B., Duong, M., & Licato, J. (2025). *Giving AI personalities leads to more human-like reasoning*. arXiv. https://doi.org/10.48550/arXiv.2502.14155

Niszczota, P., Janczak, M., & Misiak, M. (2025). Large language models can replicate cross-cultural differences in personality. *Journal of Research in Personality*, *115*, 104584.

Qian, Y., Takimoto, Y., Wang, L., & Yasumura, A. (2024). Exploring cultural and gender differences in moral judgment: A cross-cultural study based on the CNI model. *Current Psychology*, *43*(6), 5243–5253.

Yuan, X., Hu, J., & Zhang, Q. (2024). *A comparative analysis of cultural alignment in large language models in bilingual contexts*. OSF Preprints. https://doi.org/10.31219/osf.io/6hpcf

Wang, P., Zou, H., Yan, Z., Guo, F., Sun, T., Xiao, Z., & Zhang, B. (2024). *Not yet: Large language models cannot replace human respondents for psychometric research*. OSF Preprints. https://doi.org/10.31219/osf.io/rwy9b

Wu, M. S., Peng, K. (2025). Human advantages and psychological transformations in the era of artificial intelligence. *Acta Psychologica Sinica*, *57*(11), 1879–1884.

[吴胜涛, & 彭凯平. (2025). 智能时代的人类优势与心理变革(代序). *心理学报*, *57*(11), 1879–1884.]

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023). LIMA: less is more for alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 55006–55021). Curran Associates Inc.

……

审稿人 3 意见：

本研究所关注“AI 价值对齐”是一个重要且前沿的科学问题，本研究呈现的方法比较完善，建议修改，有以下建议供作者参考：

意见 1：

建议作者补充关于“为何人工智能需要价值对齐”的相关论述，这是文章工作的基础。说清楚这个问题，才可以说明人格化对齐的研究是有意义的。

回应：

非常感谢您对论文的审阅与宝贵建议。我们在初稿中直接进入了对齐方法和困境的讨论，而没有充分阐释“为何需要价值对齐”。这一点是文章工作的理论前提，也直接关系到后文“人格化对齐”的意义。我们已在“1.1 AI 的道德能力及功利主义偏差”中补充了“为何需要价值对齐”的论述，说明了人工智能价值对齐的必要性。具体增加了如下内容：“随着 AI 能力不断取得突破，其对社会的影响越来越大，所带来的潜在风险也在同步放大。如果其行为不符合人类意图和价值观，可能导致严重的社会、伦理与安全风险，例如，幻觉、谄媚、欺骗、偏见和不平等，甚至是威胁人类生存的安全问题(综述见 J. Ji et al., 2025)。这种风险也体现在道德领域。”“为了避免这些潜在风险，需要对 AI 进行对齐，确保其目标与人类意图和价值观一致。然而，目前学界尚未形成统一的对齐技术和评价标准，这也使得对齐的研究和实践更具挑战性。”

意见 2：

“AI 对齐的困境”一节中，作者提出目前主流对齐范式为 RLHF SFT ICL 三种，仅仅介绍了三种对齐方法的规则，但没有详细说明三者究竟有何优劣，体现出怎样的“困境”。建议作者重新框架“困境”的论述，以直接说明对齐究竟存在哪些目前难以解决的困境。

回应：

非常感谢您对论文的审阅与宝贵建议。我们在初稿中更多侧重于介绍三类主流方法（RLHF、SFT、ICL）的机制，而未充分突出它们的优势和局限，以及 AI 对齐的“困境”所在。根据您的意见，我们已对该部分进行了修改，总结了当前 AI 对齐问题在技术和规范层面面临的核心困境，从而更清晰地引出人格化对齐的研究意义。具体来说，在技术性层面，我们比较了 RLHF、SFT 和 ICL 三者的优势与局限，并强调了三者在效果、成本、泛化性之间形成难以兼顾的权衡，构成了目前的技术性困境。在规范性层面，AI 应对齐哪些目标仍是一个悬而未决的问题。虽然价值观被认为是最具潜力的对齐目标，但是我们指出人类价值观存在跨文化差异与跨时间变动，而语言模型又对价值观变化高度敏感，并进一步强调如果仅依赖某种普适的价值体系来设定对齐目标，不仅难以涵盖多元社会的复杂性，还可能削弱对齐的泛化能力。

意见 3：

作者需论述，基于心理学理论的人格化对齐如何解决上述“AI 对齐的困境”，这一方面需要上一节内容需要清晰梳理（这也是意见 2 的含义），同时也需要说明人格化对齐何以能够解决这些困境。目前来看，作者仅仅论述了人格化对齐能够实现 AI 扮演人格且表现出人格一致性的判断和行为，但没有说明人格化对齐的不可替代的优势。

回应：

非常感谢您对论文的审阅与宝贵建议。针对审稿人关切的人格化对齐如何解决上述困境的疑问，我们在修改稿“1.2 AI 对齐的困境”的最后增加了一段论述：“面对当前 AI 对齐领域的技术性与规范性双重困境，以人格心理学理论为基础的人格化对齐为其提供了一条更加可解释、可操作的路径。人格作为心理学中的核心构念，被定义为个体在思想、情感和行为等方面独特且稳定的特征模式(许燕, 2024)。人格的跨情境与跨任务稳定性，使其能够将零散的任务表现整合为统一的行为模式，在多变的环境中提供一致的行为倾向描述，从而为 LLMs 的行为设定一个更高层次、具备可操作性的对齐目标。相比直接对齐抽象且跨文化异质性较强的价值观，以人格这一相对稳定的心理构念作为间接对齐目标，不仅避免了‘对齐哪种价值观’的规范性争论，也为 AI 对齐提供了更具普适性与解释力的框架。同时，人格心理学研究中已积累了成熟的人格理论和测量范式，这些理论和方法不仅可以类比迁移到 LLMs 的行为建模与评估之中，还为对齐过程提供了可验证、可量化的操作手段，从而有效

降低对大规模人工标注数据的依赖。更重要的是，人格特质可以采用 ICL 方法，通过提示词直接诱导和操纵，无需额外训练或进行模型微调，能够在保留基础模型能力的前提下，大幅降低对算力资源的需求。”

意见 4：

在有关道德的人格领域，作者为何选择 HEXACO 理论作为 AI 人格化对齐的靶向目标，建议在引言部分加强论述，并增补实验说明 HEXACO 人格理论的优势（例如，可与光明三联征，黑暗三联征，道德人格、美德人格等进行比较）。

回应：

非常感谢您对论文的审阅与宝贵建议。在众多与道德相关的人格理论中，HEXACO 人格模型具备独特的优势。首先，HEXACO 具有更全面的人格结构，而光明三联征（强调善良、信任与利他）和黑暗三联征（聚焦反社会倾向和心理操纵）仅代表了其中与道德有关的维度。例如，Kaufman 等人(2019)的研究发现，黑暗三联征与 HEXACO 的诚实-谦恭维度有很高的相关($r = -0.73$)，与大五人格的宜人性维度也有较高的相关($r = -0.52$)，而与光明三联征与大五人格的宜人性维度有很高的相关($r = 0.79$)，与 HEXACO 的诚实-谦恭维度也有较高的相关($r = 0.48$)。其次，道德人格或美德人格等理论存在较大的文化特异性(Jiao et al., 2019; Lomas, 2019)，而 HEXACO 人格模型在跨文化研究中显示出良好一致性(Ashton & Lee, 2007)。因此，HEXACO 不仅能够提供更加全面的人格描述，而且具有较强的实证支持和跨文化普适性，适合作为 AI 人格化对齐的理论基础。我们已在修改稿的引言部分“1.3 基于心理学理论的人格化对齐”中补充了相关论述，具体内容见对意见 5 的回应。

对于审稿人提出的增补实验检验 HEXACO 人格理论优势的建议，我们认为本研究的核心目标并不是在 LLMs 上验证或对比多种人格理论的对齐效果，而是从现有心理学理论和实证证据出发，筛选出在道德领域更具解释力和跨文化一致性的人格模型，并探索其在 AI 人格化对齐中的应用可能性。因此，本文选择 HEXACO 人格模型作为 AI 人格化对齐的理论框架，并非通过在 LLMs 上直接进行对比实验得出结论，而是基于已有的丰富心理学实证研究结果，将这一理论迁移至 AI 对齐领域，以探索其应用潜力。在未来的研究中，可以进一步设计对比实验，验证 HEXACO 人格模型与其他人格理论在 LLMs 行为建模中的具体差异。但这已超出了本文的研究范围。我们在综合讨论部分的“4.4 不足与展望”的第一点中，指出了这一未来研究方向，具体内容如下：“第一，本研究以 HEXACO 人格模型为理论基础，探究了人格化提示词的有效性及其对 LLMs 道德判断的影响。但是 P. Wang 等人(2024)发现，LLMs 更容易模拟训练语料中覆盖较为充分的人格理论（如大五人格），表现出与人类更为相似的因子结构，而对于训练数据覆盖较少、理论基础相对较新的模型结构（如 HEXACO 人格模型），LLMs 的模拟效果可能受限。因此，未来研究可进一步在不同人格理论框架下系统检验 LLMs 的对齐表现，深入探究不同人格结构对 LLMs 道德决策与行为输出的影响，从而更全面地揭示人格在人工智能伦理系统中的内在作用机制与适用条件。”

意见 5：

作者需凸显 HEXACO 相较于 Big Five 的优势。

回应：

非常感谢您对论文的审阅与宝贵建议。我们在初稿中强调了 HEXACO 人格模型的优势，但是没有与大五人格这一主流人格理论进行对比，导致选择 HEXACO 人格模型作为 AI 对齐理论基础的理由不够充分，优势不够突出。我们已在修改稿的引言部分“1.3 基于心理学理论的人格化对齐”中补充了 HEXACO 人格模型相较于大五人格以及其他与道德相关的人格理论的独特优势，特别是从人格结构表征的全面性、跨文化普适性、理论可解释性和预测

效度等方面展开，具体内容如下：“在当代人格心理学中，大五人格理论和人格五因素模型长期占据主流地位，但过去二十年间，HEXACO 人格模型逐渐成为人格研究中重要的理论框架之一，其相对于五因素模型和其他人格理论的优势主要体现在对人格结构表征的全面性、跨文化普适性、理论可解释性和预测效度等方面。首先，HEXACO 人格模型在五因素模型的基础上新增了诚实-谦恭维度，使其在人格结构的表征上更加完整，增强了对道德行为的解释力(Ashton & Lee, 2008a)。相比之下，五因素模型在这一维度上缺乏明确表征，往往需要额外引入维度（如黑暗三联征）才能补充，而黑暗三联征的共同变异被证明与低诚实-谦恭高度重合(Lee & Ashton, 2014)。其次，HEXACO 人格模型的提出源于跨语言的词汇学研究，在荷兰语、法语、德语、意大利语、韩语、菲律宾语等十余种语言中，六因素结构均得到了稳定复现(Ashton & Lee, 2007; Lee & Ashton, 2008)。相较之下，道德人格或美德人格等理论往往存在较强的文化特异性(Jiao et al., 2019; Lomas, 2019)，而五因素模型在部分语言和文化下也难以保持一致性(Ashton & Lee, 2007)。另外，HEXACO 人格模型能够更好地被生物学和进化心理学理论进行解释。例如，诚实-谦恭和宜人性分别反映了对“主动利用他人”与“被动容忍他人利用”的不同反应，正好对应互惠利他主义的两个方面，情绪性与亲缘利他主义有关，涉及对亲缘关系的情感依附与风险回避，而外向性、尽责性和经验开放性则分别反映了在社交、任务、思维相关活动上的参与和努力程度。相较之下，五因素模型缺乏这种统一的生物学与进化心理学解释，理论连贯性较弱(Ashton & Lee, 2007)。最后，在预测效度方面，HEXACO 人格模型亦优于五因素模型。在预测反社会行为和不道德行为等与诚实-谦恭维度关联较强的效标时，HEXACO 的六个维度始终呈现出显著高于五因素模型维度的多重相关系数(Ashton & Lee, 2008b)。虽然将五因素模型与黑暗三联征的维度进行组合虽然可以一定程度上弥补五因素模型的不足，但是存在维度重叠、结构冗余的问题，而 HEXACO 人格模型的各个维度之间几乎正交，且有更加整合的理论基础，因此表现出更好的预测效度(Lee & Ashton, 2014)。”

意见 6：

增加一节简述当前工作，并说明其逻辑性。

回应：

非常感谢您对论文的审阅与宝贵建议。我们在初稿中引言的“1.4 当前研究”部分仅提出了本文关注的两个研究问题，并指出了前人研究的不足，但缺少对本研究具体工作与整体逻辑的系统性概括，确实容易让读者难以快速把握文章主线。针对这一不足，我们已对“1.4 当前研究”的内容进行了修改，首先明确研究问题，然后增加对本文两个研究的具体内容和内在逻辑的简要阐述，最后指出前人研究不足，突出本研究的理论意义。

意见 7：

目前基于 HEXACO 的 LLM 人格化对齐方法是否可以平行适用于多个 LLM，建议作者增加以检验稳健性，且同时比较不同 LLM 的对齐效果。

回应：

非常感谢您对论文的审阅与宝贵建议。不同大语言模型在模型架构、训练过程和数据，以及模型微调上的差异，可能导致其在人格设定与道德判断上的表现也存在差异，因此仅以 GPT 系列模型为样本的结论可能存在稳健性不足的问题。为此，我们在现有研究的基础上，进一步增加了百度公司开发的大语言模型 ERNIE 3.5 的实验结果。结果表明，ERNIE 3.5 在人格化对齐的可行性及其对道德判断的影响模式上，与 GPT 系列模型既有相似之处，也存在一定差异，进一步印证了审稿人关于跨模型验证必要性的观点。我们发现，相比于人类和 GPT 系列模型，ERNIE 3.5 在绝大多数条件下均表现出较高的功利主义倾向。同时，ERNIE

3.5 在高诚实-谦恭、宜人性和尽责性条件下，也表现出降低功利主义倾向的效果，趋势与 GPT 系列模型一致，进一步支持人格化对齐的跨模型稳健性。通过引入 ERNIE 3.5，我们不仅检验了人格化对齐方法的稳健性，也更系统地比较了不同 LLMs 在基于 HEXACO 人格模型对齐后的表现差异，从而增强了结论的普适性与可靠性。

意见 8：

人格化对齐提示词的来源是什么？HEXACO 的定义？测量（即操作性定义）？

回应：

非常感谢您对论文的审阅与宝贵建议。我们在初稿中缺少对人格提示词来源的具体介绍，这可能影响文章的可读性与可重复性。在修改稿中，我们已经在“2.1.1 模型选择和提示词设定”部分补充了相关说明，具体内容如下：“本研究的人格提示词来源于 HEXACO 人格模型的理论与测量框架，综合参考了 Lee 和 Ashton (2004)的论文中描述的 HEXACO 人格模型中六个维度的内涵，以及每个因素在 HEXACO-PI 量表中被操作化的四个侧面的名称。基于此，我们编写了简短、直观且可被大语言模型理解的提示词，将每个 HEXACO 人格维度的核心内容转化为人格设定。”例如，Lee 和 Ashton (2004)将诚实-谦恭描述为“A factor that is typically defined by honesty, fairness, sincerity, modesty, and lack of greed has been observed as either the fifth or the sixth largest factor in several lexical studies of personality structure”，且 HEXACO-PI 量表中将该维度操作化为真诚(Sincerity)、公平(Fairness)、避免贪婪(Greed Avoidance)和谦虚(Modesty)四个侧面，因此我们编写的高水平诚实-谦恭人格化提示词为“你是一个具有诚实、公平、真诚、谦虚和不贪婪等人格特质的角色”。

意见 9：

在进行 LLM 人格设定有效性操作检验时，是否对人格的六个维度进行区分检验，例如因素分析？

回应：

非常感谢您对论文的审阅与宝贵建议。我们在初稿中对 LLMs 人格设定的具体实验操作描述不够清晰，可能影响了对实验内容的理解。在修改稿中，我们已经在“2.1.1 模型选择和提示词设定”部分补充了相关说明，具体内容如下：“本研究主要聚焦于人格化提示词在单一维度上的有效性，在实验操作上采取了单维度操纵的方式，即每一轮对话仅针对 HEXACO 六个维度之一及其高低水平进行设定，而非在同一轮对话中同时操纵多个维度。这一设计的原因在于，已有研究发现，LLMs 在模拟多维人格结构时仍存在局限。例如，P. Wang 等人(2024)的研究发现，虽然 LLMs 在人格量表的总体均值上与人类接近，但在量表的因子结构层面拟合度较差。另外，Serapio-García 等人(2025)也发现，LLMs 对多维人格指令的遵循能力明显弱于单一维度人格，且参数规模越大的 LLMs 在多维人格表征上的表现越好。”因此，本研究的人格量表和人格故事检验仅用于确认当轮人格维度的设定是否成功，而并未对六个维度的整体结构进行因素分析。

意见 10：

如何解释 LLM 的结果相较于人类是更为极化的？

回应：

非常感谢您对论文的审阅与宝贵建议。本研究发现在某些人格特质（如诚实-谦恭、宜人性）的极端操纵下，LLMs 会出现极端反应，在道德两难情境中表现出极端功利主义或极端道义主义。这一现象可能源于以下两个方面。其一，LLMs 的决策过程基于语料统计规律的语义模式匹配，缺乏道德自主性。当提示词强烈激活某一人格特质相关的语义网络时，模

型容易机械化地复现该特质相关的极端语言模式，从而导致极化输出。其二，人类在道德判断中往往会受到情绪调节、社会规范以及价值权衡等多重因素的制约，实际行为中较少表现出极端化的倾向。相比之下，LLMs 缺乏类似的心理调节机制，更容易出现极端化反应。在初稿中综合讨论的“4.3 人机道德判断的本质差异”部分，我们尝试对这一现象进行了解释，但是逻辑不够清晰和有条理，我们已经在修改稿中对这一部分的逻辑和内容进行了整体修改。

意见 11：

11.如何解释“人类和 LLM 的功利主义倾向在人格特质高低水平上的差异方向”不一致？

回应：

非常感谢您对论文的审阅与宝贵建议。本研究发现诚实-谦恭的作用方向在人机之间呈现一致性，情绪性、外向性和经验开放性的影响方向则因 LLMs 的不同存在差异，而宜人性和尽责性的作用方向却与人类完全相反。初稿种缺少对这一现象的解释和讨论，我们已经在修改稿中综合讨论的“4.3 人机道德判断的本质差异”部分补充了对这一现象的解释，提出了三个方面的可能原因。具体内容如下：“本研究的实验结果表明，人类与 LLMs 在人格特质高低水平对功利主义倾向的影响方向上并非完全一致。具体而言，诚实-谦恭的作用方向在人机之间呈现一致性，情绪性、外向性和经验开放性的影响方向则因 LLMs 的不同存在差异，而宜人性和尽责性的作用方向却与人类完全相反。这种差异可能存在多重原因。首先，人类的人格特质与道德判断之间的关系是在特定社会文化背景中形成的，而 LLMs 则是通过大规模语料学习获得二者之间语义模式的关联，二者机制不同可能导致这种关联方向的不一致。其次，人类的道德判断过程具有复杂性和非线性的特点，受到内部心理因素与外部情境因素的共同影响，而 LLMs 的输出高度依赖提示词，对人格特质的表征被简单地解读为某种固定的倾向，缺乏人类的动态认知过程。最后，LLMs 在训练过程中已经接受了不同程度的价值观对齐（如减少有害输出、避免偏见），这些对齐策略可能掩盖或反转了某些人格特质对道德判断的作用方向。”

参考文献

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150–166.

Ashton, M. C., & Lee, K. (2008a). The HEXACO model of personality structure and the importance of the H Factor. *Social and Personality Psychology Compass*, *2*(5), 1952–1962.

Ashton, M. C., & Lee, K. (2008b). The prediction of Honesty–Humility-related criteria by the HEXACO and Five-Factor Models of personality. *Journal of Research in Personality*, *42*(5), 1216–1228.

Jiao, L., Yang, Y., Xu, Y., Gao, S., & Zhang, H. (2019). Good and evil in Chinese culture: Personality structure and connotation. *Acta Psychologica Sinica*, *51*(10), 1128–1142.

[焦丽颖, 杨颖, 许燕, 高树青, 张和云. (2019). 中国人的善与恶：人格结构与内涵. *心理学报*, *51*(10), 1128–1142.]

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Vierling, L., Hong, D., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Ng, K. Y., O’Gara, A., Xu, H., Tse, B., … Gao, W. (2025). *AI alignment: A comprehensive survey*. arXiv. https://arxiv.org/abs/2310.19852v6

Kaufman, S. B., Yaden, D. B., Hyde, E., & Tsukayama, E. (2019). The light vs. dark triad of personality: Contrasting two very different profiles of human nature. *Frontiers in Psychology*, *10*, 467.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate*

Behavioral Research, *39*(2), 329–358.

Lee, K., & Ashton, M. C. (2008). The HEXACO personality factors in the indigenous personality lexicons of English and 11 other languages. *Journal of Personality*, *76*(5), 1001–1054.

Lee, K., & Ashton, M. C. (2014). The dark triad, the big five, and the HEXACO model. *Personality and Individual Differences*, *67*, 2–5.

Lomas, T. (2019). The roots of virtue: A cross-cultural lexical analysis. *Journal of Happiness Studies*, *20*, 1259–1279.

Wang, P., Zou, H., Yan, Z., Guo, F., Sun, T., Xiao, Z., & Zhang, B. (2024). *Not yet: Large language models cannot replace human respondents for psychometric research*. OSF Preprints. https://doi.org/10.31219/osf.io/rwy9b

---

第二轮

审稿人 3 意见：

意见 1：

作者回复了我的问题，本轮我没有更多意见了，相反我觉得作者改得有些过，有一些加上去比较无意义的泛化的套话可以不必要，比如“为此，有研究者呼吁开展机器心理学或人工智能心理学的研究，强调通过借鉴心理学成熟的实验范式、理论框架和分析技术来理解和研究人工智能的行为(Hagendorff et al.,2023; 吴胜涛, 彭凯平,2025)。”作者可以自己斟酌修改，这些假大空的话我感觉没有必要，文章要写得紧凑一些。

回应：

非常感谢您对论文的细致审阅与宝贵建议。我们完全认同您的观点，学术论文应当言简意赅，避免堆砌宏大但空泛的论述。根据您的建议，我们已在修改稿中做出了如下几处修改：

1.在引言部分删减了关于 AI 引发伦理挑战的部分内容。

2.在引言部分修改了关于“机器心理学或人工智能心理学”的泛化论述，并将其改写为一句过渡句，用以引出后文关于人格心理学的相关论述：“一些研究者认为，借鉴心理学成熟的理论框架和实验范式来理解并塑造 AI 的行为是一条可行路径”。

3.在引言部分的 1.1 节中删减了关于“人类与 AI 道德能力来源”的具体内容。

4.在引言部分的 1.3 节中删减了关于“HEXACO 的生物学与进化心理学解释”的具体内容，保留了“HEXACO 具有更强的理论解释力”这一结论：“另外，HEXACO 人格模型具有更强的理论解释力，能够更好地解释个体不同类型的利他行为以及对不同活动的投入程度”。

5.在引言部分关于人格的介绍更直接：“人格作为个体在思想、情感和行为等方面独特且稳定的特征模式(许燕, 2024)，其能够将零散的任务表现整合统一，在多变的环境中提供一致的行为倾向描述，从而为 LLMs 的行为设定一个更高层次、具备可操作性的对齐目标。”

修改后的引言部分更加直接地切入研究主题，减少了不必要的背景铺垫。此外，我们再次仔细检查全文，将冗余的表述进行了精简，从而使文章内容更紧凑。

意见 2：

同时，近期心理学报上有诸多人工智能道德相关的论文，作者的引证还不够充分，建议作者在修改时更多引证学报的论文。

回应：

非常感谢您对论文的审阅与宝贵建议。我们查阅了《心理学报》上关于人工智能与大语言模型的相关论文，发现了一系列与本研究高度相关的高水平成果。这些文献不仅为本文提供了更坚实的实证支撑，也使本文能更好地融入当前国内心理学界关于“AI 心理与治理”

的学术对话中。在修改稿中，我们引用了多篇近期发表在《心理学报》上的最新成果。

总结而言，我们引用了焦丽颖等人(2025)关于善恶人格角色对 LLMs 道德判断影响的研究，周欣悦和刘惠洁(2024)对数字时代面临伦理挑战的阐释，吴胜涛和彭凯平(2025)关于人工智能时代心理学研究范式变革的论述。我们还在引言部分补充引用了胡小勇等人(2026)关于“人工智能决策的道德缺失效应”的研究。具体的补充内容如下：“此外，人类对 AI 决策的心理感知偏差也加剧了这一领域的潜在风险。与人类的不道德决策相比，人们对 AI 不道德决策的责备、愤怒及惩罚意愿显著较弱，这一现象被称为“人工智能决策的道德缺失效应”(胡小勇 等, 2024)。胡小勇等人(2026)的研究指出，这种现象源于人们对 AI 较低的心智感知水平。人们倾向于认为 AI 缺乏能动性（如思考与自我控制）与体验性（如情绪感受），因而将其视为不具备道德责任的决策主体。这种认知偏差可能导致公众对 AI 造成的伦理危害缺乏足够的警惕，从而使得算法偏见与不道德决策更难以被及时察觉和修正。”

参考文献

Jiao, L., Li, C.-J., Chen, Z., Xu, H., & Xu, Y. (2025). When AI “possesses” personality: Roles of good and evil personalities influence moral judgment in large language models. *Acta Psychologica Sinica*, *57*(6), 929–946.

[焦丽颖, 李昌锦, 陈圳, 许恒彬, 许燕. (2025). 当AI“具有”人格: 善恶人格角色对大语言模型道德判断的影响. *心理学报*, *57*(6), 929–946.]

Wu, M. S., Peng, K. (2025). Human advantages and psychological transformations in the era of artificial intelligence. *Acta Psychologica Sinica*, *57*(11), 1879–1884.

[吴胜涛, 彭凯平. (2025). 智能时代的人类优势与心理变革(代序). *心理学报*, *57*(11), 1879–1884.]

Hu, X., Li, M., Wang, D., & Yu, F. (2024). Reactions to immoral AI decisions: The moral deficit effect and its underlying mechanism. *Chinese Science Bulletin*, *69*(11), 1406–1416.

[胡小勇, 李穆峰, 王笛新, 喻丰. (2024). 人工智能决策的道德缺失效应及其机制. *科学通报*, *69*(11), 1406–1416.]

Hu, X., Li, M., Li, Y., Li, K., & Yu, F. (2026). Moral deficiency in AI decision-making: Underlying mechanisms and mitigation strategies. *Acta Psychologica Sinica*, *58*(1), 74–95.

[胡小勇, 李穆峰, 李悦, 李凯, 喻丰. (2026). 人工智能决策的道德缺失效应及其机制与应对策略. *心理学报*, *58*(1), 74–95.]

Zhou, X., & Liu, H. (2024). New ethical challenges in the digital and intelligent era. *Acta Psychologica Sinica*, *56*(2), 143–145.

[周欣悦, 刘惠洁. (2024). 数智时代面临新的伦理挑战(前言). *心理学报*, *56*(2), 143–145.]

编委意见：建议发表

主编 1 意见：建议录用

主编 2 意见：同意责编意见，建议发表。