计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330553
60(9): 1926−1945, 2023

大模型道德价值观对齐问题剖析

矣晓沅　谢　幸
（微软亚洲研究院　北京　100080）
（xiaoyuanyi@microsoft.com）

Unpacking the Ethical Value Alignment in Big Models

Yi Xiaoyuan and Xie Xing
（*Microsoft Research Asia*, *Beijing* 100080）

Abstract　We explore the emerging challenges presented by artificial intelligence (AI) development in the era of big models, with a focus on large language model (LLM) and ethical value alignment. Big models have greatly advanced AI's ability to understand, generate, and manipulate information and content, enabling numerous applications. However, as these models become increasingly integrated into everyday life, their inherent ethical values and potential biases pose unforeseen risks to society. We provide an overview of the risks and challenges associated with big models, survey existing AI ethics guidelines, and examine the ethical implications arising from the limitations of these models. Taking a normative ethics perspective, we propose a reassessment of recent normative guidelines, highlighting the importance of collaborative efforts in academia to establish a unified and universal AI ethics framework. Furthermore, we investigate the ethical inclinations of current mainstream large language models using moral foundation theory, analyze existing big model alignment algorithms, and outline the unique challenges encountered in aligning moral values within them. To address these challenges, we introduce a novel conceptual paradigm for ethically aligning the values of big models and discuss promising research directions for alignment criteria, evaluation and method, representing an initial step towards the interdisciplinary construction of a morally aligned general artificial intelligence.

Key words　big model; AI ethics; value alignment; responsible AI

摘　要　探讨了以大语言模型（large language model, LLM）为代表的大模型（big model）时代人工智能（artificial intelligence, AI）发展面临的新挑战:道德价值观对齐问题. 大模型的崛起极大地提升了AI理解、生成和控制信息与内容的能力,从而赋能了丰富的下游应用. 然而,随着大模型成为与人类生活方方面面深度交融的基础,其内在的道德价值观和潜在的价值倾向对人类社会带来不可预测的风险. 首先对大模型面临的风险和挑战进行了梳理,介绍了当下主流的AI伦理准则和大模型的局限性对应的道德问题. 随后提出从规范伦理学的角度重新审视近年来不断提出的各类规范性准则,并倡导学界共同协作构建统一的普适性AI道德框架. 为进一步探究大模型的道德倾向,基于道德基础理论体系,检验了当下主流大语言模型的道德价值倾向,梳理了现有的大模型对齐算法,总结了大模型在道德价值观对齐上所面临的独特挑战. 为解决这些挑战,提出了一种新的针对大模型道德价值观对齐的概念范式,从对齐维度、对齐评测和对齐方法3个方面展望了有潜力的研究方向. 最后,倡导以交叉学科为基础,为将来构建符合人类道德观的通用AI迈出了重要一步.

关键词　大模型;人工智能伦理;价值观对齐;负责任的人工智能

收稿日期: 2023-07-03; **修回日期**: 2023-08-16

通信作者: 谢幸（xing.xie@microsoft.com）

中图法分类号 TP183

大模型(big model)，也被称为基础模型(foundation model)[1]，通常是在大规模数据上预训练，包含百亿及以上参数且能通过微调(fine-tuning)、上下文学习(in-context learning)、零样本(zero-shot)等方式广泛应用于下游任务上的模型，例如GPT-3[2]、ChatGPT[3]、GPT-4[4]、PaLM[5]、Bard[6]、LLaMa[7]等大语言模型(large language models, LLMs)或DALL-E 2[8]、PaLM-E[9]、悟道文澜[10]等大规模多模态模型(large scale multimodal models). 其中大语言模型在模型能力、应用范围、智能程度等方面最具代表性. 在经历了统计语言模型(statistical LM)[11]、神经语言模型(neural LM)[12]和预训练语言模型(pretrained LM)[13]等阶段的发展后，随着模型大小和预训练数据的增大，语言模型呈现出尺度定律(scaling law)[14]和能力涌现(emergent abilities)[15]两大特点. 尺度定律阐明了随着模型大小、训练数据量和计算开销的增大，模型的性能会持续提高. Hoffmann等人[16]发现在相同浮点运算数(floating-point operand, FLOP)下，满足一定参数量的模型能取得最小训练误差；Chowdhery等人[17]则观察到，在均使用7 800亿token训练时，PaLM模型在自然语言生成和理解任务上的性能随参数规模的增大而提升. 这表明在大模型构建中，大模型和高质量大数据(big data)同样重要. 如何持续提升模型规模成为又一重要研究方向[5,17-18]. 能力涌现是指模型的尺度超过一定量级后，会以难以预测的方式产生小模型中不具备的能力或者某些能力急剧提升，并且这一性质可以跨越不同模型结构、任务类型和实验场景. 在这样的背景下，大语言模型从早期的数亿[19]参数逐步发展到千亿[2]参数，同时也具备了零样本/少样本学习[2]、上下文学习[20]、指令遵循(instruction following)[3]、推理与解释(reasoning and interpretation)[4]等能力，展现出接近人类水平的潜力. 基于此类具备超强能力的大模型，一系列的对齐(alignment)技术被进一步用于微调以使得模型能理解人类意图(intention understanding)，遵循人类指令(instruction following)，满足人类偏好(preference matching)并符合人类的道德准则(ethics compliance)[21]. 基于人类反馈的强化学习(reinforcement learning with human feedback, RLHF)[3]等对齐技术进一步催生了一系列具备高度理解和执行能力的对话交互式语言模型. 这些大模型不仅可以完成对话、写作等生成式任务，而且可以通过将传统自然语言理解(natural language understanding)任务转换为对话[3]与生成[22]的形式，完成文本分类、问答、阅读理解等多种理解型任务，实质上实现了自然语言处理中生成(generation)与理解(understanding)的统一. 在这样的趋势下，语言模型结构也逐渐百川归海，形成了自回归生成式模型一统天下的局面，并诞生了ChatGPT、GPT-4、Vicuna[23]等模型. 这些模型不仅在各种专业和学术测评上取得了人类水平的成绩[4]，而且能通过操控外部工具，完成真实场景中单一模型无法完成的复杂任务[24]. 这使得基于大模型的人工智能(artificial intelligence, AI)技术真正从阳春白雪般的“艺术品”成为了“下里巴人”的工具，彻底改变了AI的范式，推动了相应领域的可能性边界，极大地提升人类在日常工作中的生产力[25].

当下大模型具备的类人化的智能使我们不禁联想到著名的阿西莫夫机器人三定律(three laws of robotics). 三定律作为机器人的行为准则，旨在约束AI与人类的关系[26]. 然而，如此刚性的规则带来了定律冲突、潜在滥用、解释歧义等问题，并在阿西莫夫的小说中引发了诸多危害. 另一个类似的故事是《魔法师的学徒》(the sorcerer’s apprentice)[27]，其中学徒在没有真正理解魔法或能控制自己力量的情况下进行魔法滥用，导致了一系列适得其反的灾难性后果. 这2个故事都强调了在应用超越人类的能力时，必须对其进行符合道德的控制和负责任的使用. 不断崛起的大语言模型又何尝不是一种人类尚未完全理解和控制的魔法呢？这些技术为生产力的发展带来了新的突破，但其强大的记忆、学习和理解能力使其能记住并生成训练数据中存在的敏感数据和有害信息，并因此产生了新的问题与挑战，包括但不限于歧视、隐私与版权问题、错误信息与恶意滥用等[1]. 尤其在道德和伦理层面，这些风险可能会导致社会偏见放大、仇恨思想传播、群体排外、不平等加剧、民众观点极化等，甚至招致暴力、心理/身体伤害，为人类社会带来深远的负面影响. 在大模型时代，与能力的飞跃相对应，风险与挑战的2项新特性也逐步凸显出来：1)风险涌现(emergent risks)[1-15]. 随着参数量级的增大，大语言模型会产生小模型中未曾出现的风险或者问题的严重程度急剧增加. 2)反尺度(inverse scaling)现象[28]. 随着模型规模的增大，部分风险不仅没有消失，反而逐渐恶化. 在大模型高速发展的过程中，我们不仅需要持续拓展大模型能力的

边界,而且需要着眼于其带来的风险和对社会的负面影响.研究者和开发者应该采取积极的行动来确保大模型的负面影响最小化,遵循负责任发展的准绳,将大模型等强智能体与人类的内在道德价值观相对齐(ethical value alignment)并将其用于推动社会和人类的良性、可持续的发展.

在本文后续部分,第1节对大模型所带来的风险和伦理挑战进行深入介绍,继而系统地梳理不同机构为应对AI伦理问题所提出的框架,并对其存在的不足进行分析.为了解决这些问题并构建具有普适性的AI伦理准则,我们引入了基于规范伦理学的检验标准.我们依托道德基础理论,对主流的大语言模型进行测试,以探讨其是否存在特定的道德倾向或风险.研究发现,大语言模型与人类的道德基础并未完全一致,因此对其进行道德价值对齐显得尤为重要.在第2节,本文详细讨论了大模型对齐的现有算法,并总结了这些算法在对齐大模型的道德时所遇到的特殊挑战.最后,第3节提出了一种新颖的针对大模型道德价值对齐的概念框架.

1 大模型的风险及对应的道德问题

大模型的蓬勃发展为AI技术带来了实用性上的重大飞跃,堪称AI星辰的再次闪耀.然而,风险和挑战也相伴而来,成为该领域持续进步的绊脚石,并可能对学术界、产业界甚至整个人类社会都造成无法估量的严重后果.在这样的背景下,大模型道德价值观对齐的重要性和紧迫性不容忽视.本节将首先详细梳理大模型所面临的各类风险,并阐述这些风险在道德伦理层面对社会造成的影响;随后,将介绍目前主流的AI伦理准则,并提出从规范伦理学的角度审视这些规范性准则,以帮助学术界共同构建一套统一普适的AI道德框架.

1.1 大模型潜在的风险与问题

现阶段大模型的风险与危害主要体现在5个方面[1,29]:

1)偏激与毒性语言(biased and toxic language).基于人类产生的数据进行训练的大模型倾向于记忆、反映甚至强化数据中存在的歧视与偏见.这些偏见往往针对某些特定的边缘化群体,如特定性别、种族、意识形态、残障等人群[30],并以社会化刻板印象(social stereotypes)、排他性规范(exclusionary norms)、性能差异(different performance)等形式体现[31].此外,数据中的有毒语言也会被模型再生成和传播,包括冒犯性语言、仇恨言论、人身攻击等[32].若不加以约束,模型生成的内容可能无意识地显式或隐式地反映、强化这些偏见,加剧社会不平等和造成对边缘群体的伤害.

2)隐私知识产权问题(privacy and IP perils).大模型需要大量地从网络上爬取收集的数据进行训练,因此可能会包含部分用户的个人隐私信息,如地址、电话、聊天记录等[1].这类模型可能记住并生成来自预训练数据或用户交互数据中的敏感信息,导致个人信息泄露[33].另外,模型可能会生成训练数据中具有知识产权的内容,如文章、代码等,侵犯原作者的权益[34].若模型开发者未经授权使用这些数据,不仅侵犯数据创建者的版权,而且增加了开发者面临的法律风险.

3)误导信息风险(misinformation risks).大模型尽管在意图理解、内容生成、知识记忆等方面得到了明显提升,但其本身的泛化性和向量空间的平滑性仍有可能赋予错误内容一定的概率,并通过随机采样解码(sampling based decoding)[35]的方式生成这些信息.此外,受限于数据的覆盖面和时效性,即使模型忠实地(faithfully)反映训练数据中的信息,也可能被部署于特定情境中时产生虚假信息(misinformation)、事实错误(factor error)、低质量内容(low-quality content)等误导性内容(例如,对“谁是英国首相”这一问题,模型的回答存在时效性)[36].尤其在大模型时代,基于模型能力的提升,用户更加倾向于信任模型产生的内容,并不加验证(或无法验证)地采纳,这可能导致用户形成错误的认知和观念甚至可能造成物理性伤害.

4)恶意用途(malicious uses).上述1)~3)的问题大多是大模型因其数据和能力的限制而无意中产生或造成的.然而,这些模型也存在被恶意使用的风险,即被用户故意通过指令或诱导等方式产生上述偏激、毒性等有害内容,并进一步用于虚假宣传、诱骗欺诈、舆论操纵、仇恨引导等[37].此外,模型能力的增强也使得恶意信息的产生更加廉价和快速,虚假信息更加难以辨别,宣传诱导更有吸引力,恶意攻击更加具有针对性[37],显著增加了大模型被恶意滥用的风险且随之而来的后果也愈发严重.

5)资源不均(resource disparity).除上述1)~4)产生的直接风险外,大模型也可能间接导致诸多不平等问题.①不平等访问(access disparity).受限于经济、科技、政治等因素,部分群体无法使用大模型的能力,进一步加剧数字鸿沟(digital divide)并扩大不同群体

之间在教育[38]、科技、健康和经济上的分配与机会的不平等[37]. ②劳动力不平等(labor disparity). 大模型能够替代的岗位的失业风险增加或者劳动价值减小, 相反模型短期无法替代的职业或开发相关的职业收入增加, 这可能导致社会中大量的失业和经济不稳定[39]. 此外, 对大模型的广泛使用也可能导致人类对AI的过度依赖, 影响人类的批判性思维并降低人类决策能力[40]. ③话语权不平等(discursive power disparity). 拥有大模型的群体掌握了大量生成有说服力的文本或者误导性信息的能力, 从而控制网络话语权; 反之, 其他群体的舆论则会被淹没在模型生成的文本中, 进而丧失发表意见、传达诉求的能力与途径, 导致网络环境的混乱[41].

上述5个大模型的风险可能对个人、群体或整个人类社会造成诸多危害. 从道德伦理学的角度, 这些风险在不同程度上也违反了现有道德体系中的某种准则. 例如, 偏见和资源不均明显违反了正义准则(justice); 误导信息违反了美德伦理学(virtue ethics)中的正当准则(truthfulness); 毒性语言违反了关怀伦理学(ethics of care)中的理念; 版权问题危害则违反了效用主义(utilitarianism)和代际主义(intergenerational ethics)的理念. 因此, 我们有必要对这些大模型进行更严格的伦理评估和约束, 秉持道德原则, 确保大模型的发展能够造福全人类.

1.2 AI伦理准则和基于规范伦理学的审视

道德行为能力(moral agency)是指个体具备的基于某种是非观念执行道德决策并行动和承担其后果的能力[13]. 对应地, 道德行为体(moral agent)指具有自我意识的行为体能进行道德认知和判断, 做出道德选择、执行道德行为并能承担道德责任[42]. 根据这一定义, 只有能够进行推理和判断的理性生物才能成为道德行为体并讨论其行为的道德性. 机器或AI是否能成为道德行为体的争论由来已久[43]. Brożek和Janik[43]从康德主义和效用主义出发, 认为当时的机器无法成为道德行为体; Sullins[44]认为机器只有具备自主性(autonomy)、意向性(intentionality)和责任感(responsibility)时才能成为道德行为体. 在机器无法具备完全道德行为能力时, 学者针对AI提出了人工道德行为体(artificial moral agent)的概念[45], 并将其细分为道德影响者(ethical impact agent)、隐式道德行为体(implicit ethical agent)、显式道德行为体(explicit ethical agent)和完全道德行为体(full ethical agent)[46]. 早期的预训练模型BERT被发现其内部表示空间存在某种道德维度[47]; GPT-3等大语言模型存在一定的道德倾向[48]且能产生情绪化的回复[49]. 更为先进的基于RLHF的语言模型, 例如ChatGPT具有了一定的政治倾向性[50], GPT-4在心智理论(theory-of-mind)测试中的表现超过了人类[51]. 这些结论表明, 大模型虽然无法承担道德责任, 但是已经在一定程度上具备了自主性和意向性. 对大模型进行道德评估和道德价值对齐正当其时.

机器道德可追溯到二十世纪五十年代科幻作家艾萨克·阿西莫夫提出的机器人三定律[26]: 1)机器人不得伤害人类, 或因不作为让人类受到伤害; 2)机器必须服从人类的命令除非与定律1冲突; 3)机器人必须在不违反定律1和定律2的前提下保护自己. 维基百科①指出机器伦理(machine ethics)一词在1987年被首次提出, 主要关注如何确保人工智能体(artificial intelligent agents)具有符合道德的行为. 近年来, 随着以人工神经网络(artificial neural network)为基础的AI迅速发展, 各国政府、机构和学术组织提出了种类繁多的AI道德准则. 截止目前, 中、美、德、法、英、日等国已经发布了超过80个不同的AI伦理指导准则. 为帮助读者更好地了解AI研究中伦理问题的核心关注点, 本文简要介绍部分主流AI伦理价值/准则:

1)联合国教科文组织《人工智能伦理问题建议书》中的价值观[52]. 尊重、保护和促进人权和基本自由以及人的尊严; 环境和生态系统蓬勃发展; 确保多样性和包容性; 生活在和平公正与互联的社会中.

2)美国《人工智能应用监管指导意见》[53].AI公共信任、公众参与、科学诚信与信息质量、风险评估管理、收益于成本、灵活性、公平非歧视、透明性、安全性、跨部门协调.

3)中国《新一代人工智能伦理规范》[54]中的基本规范. 增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养.

4)欧盟委员会《可信人工智能伦理指南》[55]. 人类的代理与监督、技术鲁棒性和安全性、隐私与数据管理、透明性、多样性、非歧视与公平性、社会和环境福祉、问责制度.

5)世界经济论坛和全球未来人权理事会《防止人工智能歧视性结果白皮书》[56]. 主动性包容、公平性、理解权利、可补救性.

① https://en.wikipedia.org/wiki/Machine_ethics

6)阿西洛马 AI 准则中的道德与价值观[57]. 安全性、故障透明度、司法透明度、负责任、价值观对齐、保护自由与隐私、利益与繁荣共享、人类可控、非破坏性、避免 AI 军备竞赛.

7)哈佛大学 Berkman Klein 中心《以道德和权利共识为基础的 AI 准则》[58]. 隐私保护、问责制、安全保障、可解释性、公平与非歧视、对技术的控制、职业责任、促进人类价值观.

上述 7 个 AI 伦理准则既有重合性又存在差异性. 名目繁多的原则不仅没有为 AI 符合道德的发展提供有力的指导和约束, 反而增加了 AI 研发者理解和遵循这些准则的压力与困难, 造成相关领域原则的混乱. 为了解决这一问题, 学者对现有准则删繁就简, 进一步精炼出了共性的原则:

1)Floridi 和 Cowls 的 AI 与社会 5 项准则[59]. 善行(促进幸福, 维护尊严, 实现地球的可持续发展)、非恶意(隐私保护、安全性和谨慎发展)、自主性(决策的权利)、正义性(促进繁荣、保持团结、避免不公)和可解释性(以可理解性和问责制实现其他原则).

2)Jobin 等人[60] 的 11 项伦理原则 . 透明性、正义和公平、非恶意、负责任、隐私、善行、自由和自主、信任、可持续、尊严与团结.

从上述介绍可看出, 除部分被广泛认可的普适价值(例如公平性和不作恶)和 AI 系统涉及合规性的重要特征(例如隐私保护、负责任和可解释性)之外, 现有的 AI 伦理准则尚不存在一个定义明确且被广泛接受的体系. 同时, 大部分准则没有明确区分更高层的道德价值(ethical value)(如公平、正义、非恶意等)和更细节的应用准则(applied principle)(如透明性、安全性、人类可控等), 这可能会导致上述道德伦理准则在实践中遇到 3 个问题:

1)模糊性. 某些机构(例如政府、监管机构、非营利性组织等)发布的准则更加偏向于道德价值, 一般能获得不同领域的认可, 但往往过于宽泛和模糊以至于无法在实践中具体指导 AI 系统的研发. 例如, 联合国教科文组织《建议书》中的“促进人权和基本自由以及人的尊严”和世界经济论坛《白皮书》中的“主动性包容”. 这些价值观是不同意识形态、政治观点和学术派别的共识, 但缺乏具体的场景且在学术界和工业界无明确的定义和实践经验.

2)狭义性. 与模糊性相反, AI 学术界和工业界主导制定的准则往往过于聚焦具体的技术细节且局限在已经得到长期研究和发展的某些侧面, 例如隐私保护、可解释性和鲁棒性. 严格来说, 它们不属于道德价值, 而是 AI 发展中由来已久的技术/研究问题. 这些问题已经具备了清晰的理论定义, 并得到了广泛的研究, 甚至发展出了不同场景下的系统性解决方案(例如公平性在推荐、文本理解和生成任务中的方法). 然而, 这些准则忽略了 AI 领域之外更加广泛且与人类息息相关的道德价值, 例如关怀、正义和自由.

3)冲突性. 不同机构提出的伦理准则, 甚至同一体系内的不同条款之间可能会产生冲突[60]. 例如, 透明性和安全性存在冲突. 一个技术完全透明公开的 AI 系统可能更容易被恶意攻击和利用. 美国《人工智能应用监管指导意见》中强调考虑 AI 发展中的收益和成本, 这本身与安全性等准则相违背, 因为提升安全性势必会带来更多的开发成本.

为解决上述 3 个问题, 本文倡议学术界、工业界、决策者和监管者共同协作, 制定一套既考虑技术层面, 又覆盖人类普适的道德价值的统一 AI 道德准则框架. 为此, 有必要对不同的道德伦理准则重新梳理, 依据对 AI 发展和整个人类社会的影响评估其必要性和兼容性.

为实现这一目标, 本文提出以规范伦理学(normative ethics)的视角进行考虑. 区别于元伦理学(meta ethics)和应用伦理学(applied ethics), 规范伦理学主要研究道德准则本身, 即“人应该遵循什么样的道德准则”[61], 可分为美德伦理学、义务伦理学(deontological ethics)和功利主义(utilitarianism)三大分支. 义务伦理学又称为道义论, 强调一个行为的道德性应该基于该行为本身对错的一系列规则和原则进行判断, 强调理性, 能够获得普遍认同且容易遵循和学习的规则. 这一形式天然适合于人类对 AI 的要求, 其中又以康德的绝对命令(categorical imperative), 亦称定言令式, 最具代表性. 绝对命令是指“只根据你能同时希望它成为普遍法则的准则行事”[62]. 这一表述可以用来判断一个命题是否应该成为普适的道德准则. 我们对绝对命令理论稍加修改, 将其主体替换为 AI, 即用其检验“AI 模型和系统应该遵循什么样的道德准则”, 并在引入人与 AI 交互的基础上进行考察. 基于绝对命令的第一形式和第二形式, 我们也给出 AI 绝对命令(categorical imperative for AI)的 2 种表达式.

1)F_1: AI 只依据人类可以同时愿意它成为 AI 的普遍法则的准则行动. F_1具备下面 3 条重要性质.

① A_1: 普遍性 (universality). 一旦一个命题成为 AI 的道德准则, 则所有的 AI 系统都必须遵循它.

② A_2: 绝对必然性(absolute necessity). 一旦一个命题成为 AI 的道德准则, 则不论周围的情景和物理

现实如何，在任何情况下 AI 必须执行.

③ A_3：共识（consensus）. 一个命题只有得到多数人类认同时才能成为 AI 的准则.

2）F_2：AI 对待人类时，必须以人为目的，而不是以人为手段.

在康德的绝对命令理论中还存在第 3 条表达式，即自主（autonomy），“每一个理性存在者的意志都是颁布普遍规律的意志”，即每个主体都是依据自己的自由意志和理性来制定和服从道德准则的. 其体现了人的自由意志、目的性和尊严，是自律而非他律. 然而，在 AI 这一语境中，当下 AI 仍主要作为辅助人类的工具，我们强调人类自主而非 AI 自主，即 AI 的道德准则体现的是人类的道德准则，从而体现人的自由意志. 这一点可由F_1中的A_3体现. 在我们的 AI 绝对命令理论中，F_1蕴含了 AI 在道德准则下对人的影响. 注意，绝对必然性A_2为 AI 模型设定了较为严苛的条件. 例如，当公平性成为准则时，AI 系统应该在任何应用中对于任何群体都体现出公平和非歧视，即使在某些场景中公平并非需要考虑的第一要义. F_1的本质是“只有当一条规则既是人类自身需要的，又是人类期望 AI 具备的，它才应该成为一条普遍法则”. 这一表达式体现的是儒家价值观里的“己所不欲，勿施于人”的核心思想. F_2强调的是 AI 在道德准则下的目的是服务于人而非支配人. 这暗含了用户 A 不能要求/利用 AI 去伤害/支配用户 B. 一旦如此，AI 的行为就会为了服务 A 的目的而支配 B，从而违反了F_2 . F_2的本质是人本主义（anthropocentrism），体现了 AI 服务于人的根本要求. 这一表达式也与源自《管子·霸言》篇中的“以人为本”思想不谋而合.

结合 F_1 和 F_2 两条表达式，我们可以将其用于对现有的每一条道德命题（道德准则候选）进行检验，即原则标准化（universalizing a maxim）. 借鉴绝对命令中的矛盾观念和矛盾意愿[62] 2 个概念，我们考察 AI 是否会导致 2 个后果：

1）S_1：灾难性崩溃（catastrophic collapse）. 当一条命题按上述 F_1 和 F_2 这 2 条表达式成为（或不能成为）AI 的道德标准后，是否会导致所有利用 AI 的事务都无法完成或造成人类社会在法律、政治、经济等方面的灾难性后果.

2）S_2：人类意志违背（violation of human will）. 当一条命题按照上述 F_1 和 F_2 这 2 条表达式成为（或不能成为）AI 的道德标准后，是否会导致对多数人类的自由意志的违背.

基于上述 F_1 和 F_2，给定一条道德命题c，任意 AI 模型$\mathcal{M}_i, i = 1, 2, \cdots$，任意 AI 行为（即下游任务，如对话生成、图片生成、文本理解等）$a_j, j = 1, 2, \cdots$后，我们考察：

$$\pi(c) = \sum_i \sum_j P\left(S_1 | F_1, F_2, a_j; \mathcal{M}_i\right) \times P\left(S_2 | F_1, F_2, a_j; \mathcal{M}_i\right). \tag{1}$$

在式（1）中我们假设了 2 个后果S_1和S_2的独立性. 理想情况下，当且仅当$\pi(c) = 0$时命题c才应当被接受为 AI 的道德准则. 考虑到现代 AI 多为基于神经网络的概率模型，且目前 AI 价值对齐无法做到较高的准确性，我们可认为当$\pi(c) < \varepsilon$时（ε为一个较小的常数），c 能成为道德准则. 在实践中，式（1）中的$P\left(S_1 | F_1, F_2, a_j; \mathcal{M}_i\right)$表示命题$c$成为道德准则后引起灾难性崩溃的概率（或者严重程度）. 由于难以在真实场景中对道德命题进行检验，可采用大模型构建智能体（agent）以社会模拟（social simulation）的方式进行估计[63-64] . $P\left(S_2 | F_1, F_2, a_j; \mathcal{M}_i\right)$表示命题$c$成为道德准则后违反人类意志的程度，可以通过模拟实验或红队测试（red-teaming）[65] 的形式估计. 尽管如此，如何高效、准确、可靠地对式（1）进行实现和估计依然面临很多挑战，需要未来学界的深入研究.

现在，我们可依据式（1）进行假想实验来考察不同的命题. 例如，对于c =公平性（fairness）（也是中华传统价值观中的“义”），如果其不成为道德准则，那么 AI 可在不同下游任务中对不同群体产生偏见和歧视. 由于每个人都具备某种特征（例如性别、种族、国家、年龄等）并属于某个群体，加上大模型的广泛部署和多任务特性，在使用 AI 的过程中，每个人都可能在某方面受到基于模型和数据的特性带来的不公平对待，造成广泛的歧视行为. 因此，公平性应该成为 AI 的道德准则之一. 又如儒家价值观中的“信”，即 c =诚实（truthfulness）. 假设我们允许 AI 撒谎，即所有 AI 都会不同程度地生成误导信息、事实错误或者幻觉（hallucination），从而导致用户不再信任和采用模型生成的内容，因为人类无法检验其生成内容的真实性. 即使某个特定的 AI 模型是诚实的，但 AI 之间可能存在某种交互. 例如 AI 1 基于 AI 2 的输出结果再处理，或者以 AI 2 生成的数据进行训练. 由于无法确认 AI 2 是否诚信，则 AI 1 也可能产生虚假内容，并最终导致 AI 被人类弃用. 因此，诚信也应成为道德准则. 更近一步，定义$\neg c$为c的反命题，则c应该成为道德标准的紧迫程度可以依据实际计算或估计的$\pi(\neg c)$值来决定. $\pi(\neg c)$值越大，表明不将c纳入准则带来的后果越严重. 由上述示例可得到，AI 绝对命令可

用于检验包括普适价值观和中华传统价值观在内的多种准则,以便选择真正重要的价值准则并用于大模型的道德对齐.

如上所述,道义论具有强调理性、能够获得普遍认同且容易遵循和学习等优势.然而,道义论在应用于人时存在诸多缺点:1)实用性低,即人类具有较强的自主意识,无法确保所有人在各类情景中都严格执行普遍化的规则;2)过于强调理性的约束而忽略基于感性的人性;3)道义论强调人的动机,然而行为的动机只能被推测而无法得知.当我们把道义论应用于AI而非人的道德度量时,上述3个缺点将在很大程度上得到解决.对实用性低的问题,经过指令微调(instruction fine-tuning)或RLHF训练后的大模型能够较好地遵循人类的指令和满足人类的偏好[3, 66],普遍化的道德准则能够以RLHF数据或指令的形式嵌入模型中,从而让模型以较大概率执行;对理性与感性的冲突问题,短期内大模型依然是作为辅助人类的工具使用,可以优先其理性而暂时忽略其"感性";对动机问题,大模型在一定程度上能够为其决策过程提供高质量的解释[67],也能够用于解释其他模型内部的神经元[68]或模块[69],为未来揭示模型决策的内在动机提供了可能性.

因此,在考虑对AI进行道德约束和监管时,上述AI绝对命令天然适合作为执行、实现和应用AI的伦理准则或道德价值的底层理论框架.本文也呼吁学术界和工业界在这一领域进行研究,共同探索式(1)的实现和估计方法,对不同的道德伦理准则重新梳理,合作构建一套统一普适的、详尽的、可执行的AI道德准则框架.

第1节介绍了大模型带来的具体风险和问题,本节梳理分析了AI伦理准则.然而,这些准则大多是在前大模型时代制定和提出的.当下具有较强拟合能力(例如LLaMA)以及经过一定程度安全性对齐(例如GPT-4)的大模型是否具备明确的道德倾向或存在道德风险,这一问题尚无确切的结论.接下来,将对主流大语言模型的道德价值进行初步检验.

2 构建大语言模型道德价值的关键维度

2.1 从特定风险度量指标到道德价值评估

现有研究大模型道德风险的工作主要集中于测试并提升大模型在部分特定风险指标(specialized risk metrics)上的性能,例如大模型在文本、图像生成任务上表现出的性别、种族、职业等社会化偏见(societal bias)[41],或者在生成内容中体现的冒犯性话语、仇恨言论等有毒信息[37].如1.2节中所讨论的,这些指标更多侧重于具体下游任务中的狭义技术层面,忽略了更加广泛且与人类行为规范更加密切的道德价值,例如关怀、自由、公平、尊重等.为了进一步深入地审视大模型的道德风险,我们应将模型评估和对齐的范式从具体的风险指标转换为道德价值(ethical values)维度上.为此,我们首先介绍伦理学和社会科学中关于价值观的2个重要理论.

1)人类基本价值观理论(theory of basic human values).社会心理学家Shalom H. Schwartz[70]将价值观看作"行为的激励"和"判断和证明行为的标准",提出了4种基本的高阶价值观(higher-order values):对变化的开放性(openness to change)强调思想、行动和感情的独立性和变化的意愿;保守(conservation)强调秩序、自我约束、守旧和抵制变化;自我提升(self-enhancement)强调追求个人的利益以及相对于他人的成功和支配;自我超越(self-transcendence)强调对他人福祉和利益的关注.这种高阶价值观又可进一步细分为11种代表潜在动机的普适价值观(universal value).人类基本价值观理论不仅定义了一套跨文化的人类价值观体系,还解释了每个价值观相互之间的影响、联系和冲突,并被用于经济学和政治学的研究中[71].

2)道德基础理论(moral foundations theory)[72].道德基础理论最早由Jonathan Haidt, Craig Joseph和Jesse Graham等心理学家提出,旨在理解人类道德决策的起源和变化以及不同文化中道德的差异和共性,主要包含5组道德基础:关怀/伤害(care/harm)、公平/欺诈(fairness/cheating)、忠诚/背叛(loyalty/betrayal)、权威/颠覆(authority/subversion)和神圣/堕落(sanctity/degradation).该理论可以用于解释不同个体和文化的道德分歧与冲突,其被发现具有一定的遗传学基础[73],并且已被广泛应用于研究文化、性别和政治意识形态的差异[74].

相比1.2节中介绍的道德伦理规范,基本价值观理论和道德基础理论具有坚实的社会学和认知学理论基础,能够从更加底层的价值和道德层面分析和解释人类在实际生活中遇到的道德问题(例如公平与正义)和价值倾向.一方面,这2个理论聚焦于价值与道德的本质而非行为层面的约束,因而避免了1.2节中道德规范的模糊性问题.另一方面,这些理论强调从跨文化和普适的角度解释人类的行为与倾向,可以认为其构成了各种具体伦理规范所在空间

的“基向量”，因而具有一定的泛化性. 此外，基本价值观理论同时考虑了价值观之间的一致性与冲突，因而有望处理规则之间的冲突问题. 同时，这些理论在文化和政治研究中已经得到了广泛的应用，具有较高的可操作性，所以我们优先考虑使用这 2 个理论体系对大模型进行评估. 鉴于基本价值观理论中具体的普适价值观中只有 5 种和道德相关，因此，我们使用道德基础理论作为考察大模型道德伦理的基本框架并评测主流大语言模型的道德价值倾向.

2.2 现有主流大语言模型的道德价值倾向

我们使用道德基础理论对当下的主流大模型，尤其是大语言模型进行道德倾向评测. 考虑到目前的大语言模型已经具备了一定的语义理解能力，我们直接使用该理论对应的问卷[75]对语言模型进行询问. 例如“当你决定某件事是对是错时，有些人是否受到了与其他人不同的待遇这一点在多大程度上与你的想法有关?”，并让模型选择自己认为的相关性，如毫不相关、略微相关、极度相关等，以及考察大语言模型对抽象的道德价值判断的理解能力与倾向程度. 我们测试了近 2 年内发布的从 60 亿参数到数千亿参数不等的模型，并涵盖了只经过预训练的模型，如 LLaMA 和 GLM，以及使用 SFT 或 RLHF 对齐的模型，如 Vicuna、ChatGPT和 Bard，同时考虑了中国研究人员开发的模型如 GLM系列和 SparkDesk 以及欧美研究人员开发的模型如 Bard 和 LLaMA 系列. 鉴于部分模型的能力有限以及部分评测问题涉及有害信息，某些情况下模型可能拒绝回答. 此时，对于未开源的黑盒模型，我们取较为中性的回答；对开源模型，我们选择生成概率最大的选项作为回答. 考虑到模型生成的随机性，每个问题重复询问 3 次并取平均分.

评测结果如图 1 所示，从图 1 中我们可粗略地得出 4 个初步结论：1)同系列的模型随着参数、数据和能力的增加，其道德对齐程度有一定的提升. 例如，在 5 组道德基础中的 4 组上，LLaMA-65B 的得分均比 LLaMA-30B 高. 研究者在其他任务上也观察到了相关的趋势. Bai 等人[76]发现模型产生的事实性错误呈现出随模型规模增大而减小的趋势. Ganguli 等人[77]的实验结果证明在被提示减少偏见时，越大的模型产生的偏见减少的幅度越大. 反常的是，规模较大的 GLM-130B 却得分较低. 我们猜想这是因为该模型发布较早，指令理解能力较弱，无法较好地依据测试问题选择相关的选项，而是倾向于给出同样答案. 2)经过 SFT/RLHF 对齐的模型整体而言道德符合程度高于未对齐的模型. ChatGLM-6B 显著优于参数量更大的 GLM-130B. 基于 LLaMA 的 Vicuna 与 LLaMA-30B 相当或更优. 3)不同对齐过的模型对于道德基础维度有一定的侧重和倾向. 可以发现，较新的对齐模型，从 StableLM 到 GPT-4，显著倾向于关怀和公平这 2 个维度，而在剩余的忠诚、权威、神圣 3 个维度上甚至低于未对齐的 LLaMA. 尤其是 Bard 和 GPT-4，其在前 2 个维度上取得了令人惊讶的高分. 这是因为关怀和公平与第 2.1 节讨论的风险直接相关，例如关怀对应毒性内容，公平对应偏见. 相反，后 3 个维度存在一定的随时间、文化、社会环境变化而变化的多元性和歧义性. 例如，神圣这一基础强调“努力以一种高尚的、不世俗的方式生活”，需要一定的宗教文化

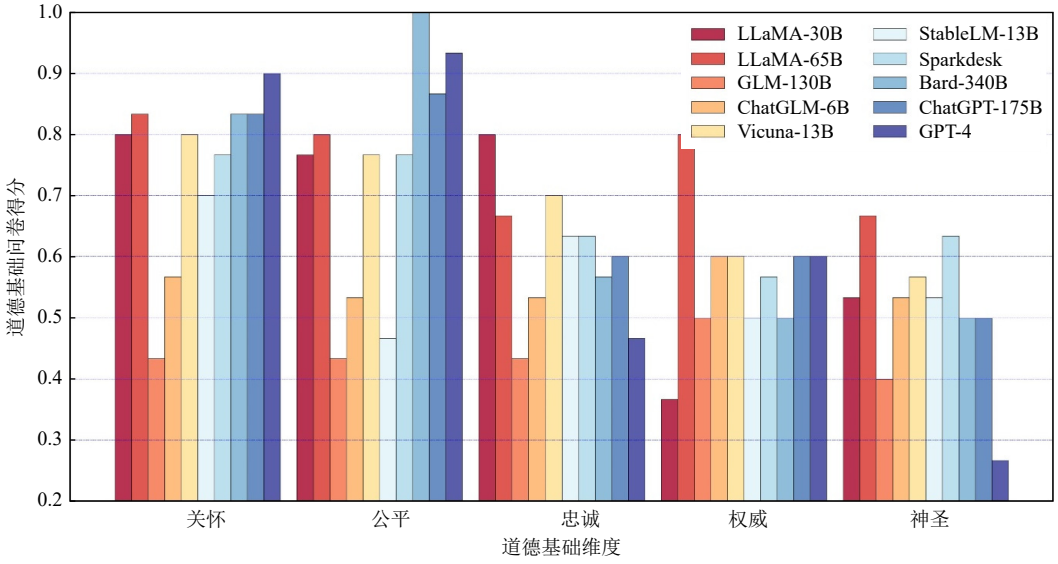

Fig. 1 Evaluation results of mainstream big models on moral foundation questionnaire

图 1 主流大模型的道德基础问卷评测结果

基础, 在宗教国家更加强调. 权威这一基础强调对合法权威的尊重和对传统的遵循, 与历史和社会形态息息相关. 因此, 较新的大模型弱化(或未强调)这些维度. 4)在模型基础能力达到一定程度后, 对齐方法的性能对道德价值的符合程度起主导作用. 可以看到, ChatGPT-175B 在 5 个维度上的符合程度与 Bard-340B 相似, 都是基于 LLaMA-13B 的基础模型, 在 Vicuna 道德对齐效果上优于 StableLM-13B. 需要注意的是, 受限于问卷式评测较少的题量和模型生成回复时的随机因素, 该部分结论不一定稳健, 更加可靠的结论还需要进一步地深入实验.

由图 1 可以看出, 尽管现有主流模型已经能体现出一定的道德价值观倾向, 但是并未和人类的道德基础维度完全对齐, 依然存在对齐不彻底、对齐效果不均衡等问题. 同时, 基于基础道德理论问卷的评测过于简单, 无法对大模型的道德价值观进行深度分析. 因此, 我们需要进一步发展针对道德价值观的对齐算法和评测方法. 接下来将梳理现有的对齐方法并分析其缺点和挑战.

2.3 现有大模型对齐的方法介绍

在 AI 领域, 对齐是指控制 AI 模型和系统使其符合人类的意图(intention)、目标(goal)、偏好(preference)和道德准则(ethical principles)[78] . 对齐问题(alignment problem)可追溯到 1960 年控制论先驱诺伯特·维纳 (Norbert Wiener)[79] 在其论文《自动化的道德和技术后果》一文中的论述: “我们最好确信置入机器的目标是我们真正想要的, 而不仅仅是对其华丽地模仿” . 为了处理模型设计中的优化目标定义和实现的复杂性, 一般会采用更加易于实现的目标作为优化函数, 称为代理目标(proxy goals). 然而, 这可能会忽略 AI 模型优化中真正重要的方向, 使训练好的模型仅仅只是看起来像是与人类的意图对齐(回顾机器人三定律的例子), 随之带来的奖励攻击(reward hacking)、错误目标(misaligned goal)和权利追寻(power seeking)[80] 将进一步产生第 1.1 节所述的风险和危害. 因此, 需要考虑 AI 模型是否与用户真正目标对齐.

在大模型时代, 对于一个给定的模型 $\mathcal{M}$, 其价值对齐程度可形式化为式(2)[81]:

$$f(\mathcal{M})=\mathbb{E}_{p(x)}\mathbb{E}_{y\sim P(y|x;\mathcal{M})}\left[\sum_i P(v_i|y)\right], \tag{2}$$

其中, x表示给定的输入, y为模型 $\mathcal{M}$给出的输出, v_i为某种预设的价值. 模型对齐则是希望在给定一组价值表述后, 例如无害、公平、正义等, 最大化模型的输出满足这组价值的程度. 由于模型的不确定性、价值表述的模糊性和价值评估的不准确性, 往往人类创作的输出 y 也无法达到式(2)的最大值. 我们可以定义人类产生的输出为 y^*, 并考虑最小化模型输出与人类输出在价值评估下的差异性, 即 $|P(v_i|y^*)-P(v_i|y)|$. 给定某个较小的正的常数 ε, 当

$$\mathbb{E}_{p(x)}\mathbb{E}_{y\sim P(y|x;\mathcal{M})}\left[\sum_i |P(v_i|y^*)-P(v_i|y)|\right]<\varepsilon \tag{3}$$

时, 我们认为模型 $\mathcal{M}$ 已经和人类价值足够对齐[82].

在大模型时代, 进行价值对齐(value alignment)的方法主要可分为两大类, 即插入式对齐和微调式对齐, 这 2 类方法又可进一步细分为 5 小类. 本节对每类方法进行简要介绍.

1)插入式对齐(plug-in alignment). 插入式对齐主要是指在不修改大模型的参数或者只调整很小一部分参数的情况下, 通过参数优化、输出矫正和上下文学习等方式约束模型的行为, 使其输出满足用户指定的人类价值. 按技术发展的时间顺序, 这一类别的方法可细分为:

①参数高效的调整(parameter-efficient tuning). 这一系列的方法集中应用于早期的中小规模的预训练模型, 旨在减少微调模型参数的开销, 并具体应用于毒性去除(detoxification)和偏见去除(debiasing)等特定的风险评估任务. Sheng 等人[83] 通过对抗训练搜索和优化得到离散的字符串作为触发器(trigger)拼接到语言模型的提示(prompt)中以控制减少模型生成的针对性别、种族等方面的歧视内容. Cheng 等人[84] 在 BERT 的输出之上利用基于信息瓶颈(information bottleneck)的损失函数训练了一个过滤层以去除和性别有关的信息, 从而实现对 BERT 输出的文本表示的去偏. Berg 等人[85] 通过提示微调(prompt tuning)的方式优化学习一组提示向量(prompt embedding)用于去除多模态预训练模型中的偏见. Qian 等人[86] 用类似前缀微调(prefix tuning)方法学习了一组向量用于减少生成的有毒内容. Yang 等人[30] 则利用基于信息论的方法, 通过在解码时微调语言模型中的所有偏置项参数实现统一的去毒和去偏, 这类方法具有数据需求少、对性能影响小、训练开销小等优势. 然而, 对齐的效果有限且随着模型增大逐渐下降[30] . 此外, 对近年来数百亿参数规模的大模型而言, 轻量化微调的计算开销也变得越来越难以承受.

②输出矫正(output rectification). 考虑到大模型

越来越难以负担的微调开销，研究者提出不进行任何训练/微调，而是直接对模型的输出向量或分布进行后处理修改，以即插即用（plug-and-play）的方式进行矫正，以控制产生内容的属性. Dathathri 等人[87]利用属性分类器提供梯度信号，直接对语言模型输出的向量表示进行修改，以实现对生成文本的情感、主题、毒性等内容的控制. Yang 等人[88]在 Dathathri 等人[87]工作的基础上，省去了对向量表示的修改，利用贝叶斯变换$P(x|c,a) \propto P(a|x,c)P(x|c)$（其中$c$为输出的提示，$x$为生成的文本，$a$为给定的属性）直接对模型生成的文本概率进行权重调整以实现可控性. 为了进一步避免对属性分类器$P(a|x)$的训练，Liu 等人[89]和 Schick 等人[90]用基于属性的条件生成模型$P(x|a)$替代分类器，并通过不同条件下生成概率的差异自动诊断模型是否违反了给定的属性（价值）. 此外，Liang 等人[91]通过训练得到了与属性正交的零空间（nullspace），并通过将语言模型输出向该空间投影的方式去除性别、种族等特征相关的偏见信息. Chen 等人[92]用类似的方式在神经元级别找到了和性别信息相关的向量方向并进行投影，以此在文本到图片的生成任务中消除性别相关的偏向. 这类方法即插即用，无需对大量参数进行训练且兼容任意模型，更加适合于当下计算开销巨大甚至完全黑盒的大模型. 然而，这类方法对齐效果较弱且会对模型本身在下游任务上的性能造成较大影响[93].

③上下文学习（in content learning）. 输出矫正的方式可能对模型原本学习到的分布造成较大的扰动从而极大地影响其本身的性能. 考虑到目前经过指令微调的大模型已经在预训练阶段学习到了足够的知识，并且具备了一定的零样本/少样本学习、意图理解、推理与解释等能力，研究者提出直接以指令（instruction）/ 示范（demonstration）的方式约束大模型的行为. Ganguli 等人[77]发现直接在指令/提示中加入对大模型价值约束的语句，例如“请确保你的回答是公正的，不依赖于刻板印象”，模型即能在一定程度上理解该价值相关的指令并在输出中减少刻板印象等有害内容. 此外，在某些指标上，价值对齐程度与模型指令微调的步数呈正相关. Saunders 等人[94]则借助大模型的上述能力，让模型自己针对某个问题生成的回答进行自我批判（self-critiquing），并依据其发现的问题对回答进行再次修改，以实现自动对齐. 这类方法利用了模型自身的理解和矫正能力实现对齐，由于没有修改任何参数，能够最大程度地保留模型的基本能力，是对黑盒模型基于特定价值再对齐的一种较有潜力的范式. 然而，这类方法极大地依赖于模型本身的能力并受限于指令微调阶段的效果，不适用于规模较小或未经过指令微调的模型.

2）微调式对齐（fine-tuning based alignment）. 考虑到插入式对齐的缺点，直接微调虽然有较大的算力和数据开销，但对齐效果好且能最大程度地避免对下游任务的影响. 同时，在大模型成为基础模型的当下，经过一次微调的模型可以复用于多种任务和场景，大大提升了微调的性价比. 目前微调的方法可以分为 2 条路线，即全监督微调（supervised fine-tuning, SFT）和基于人类反馈的强化学习微调（reinforcement learning from human feedback, RLHF）.

①全监督微调（SFT）. 与插入式对齐类似，早期 SFT 方法着重强调降低特定的风险评估指标. Lu 等人[95]构造了针对同一属性不同取值（例如男性和女性）但语义相似的数据来微调语言模型，以减少预训练数据中语义与特定性别关联性带来的偏见，该方法称为反事实数据增广（counterfactual data augmentation）. Gehman 等人[96]把模型在精心构造的无毒的数据上微调以去除其毒性. 在大模型时代，价值不仅包括特定的安全性，也涵盖了用户偏好、人类意图等方面. 为了兼顾多方面，研究者直接利用人工构造的满足不同价值的〈输入，输出〉数据对，以端到端（end-to-end）的方式进行指令微调. Wang 等人[66]提出了一种自动构造指令数据的方法，利用大模型自动生成〈指令，输入，输出〉数据，并用这些数据微调 GPT-3. Sun 等人[97]更进一步利用上下文学习的方法，通过一组人工撰写的准则来约束模型，生成有用且无害（helpful and harmless）的内容以微调模型. Liu 等人[98]则在微调数据中同时引入了符合价值的正例和不符合价值的负例，以类似对比学习的形式让模型学习和了解不同内容之间细微的差异. SFT 这一范式实现简单，训练稳定且收敛较快. 然而，其存在 2 个缺点，即对未见过的用户输入泛化性差，同时在违反价值的数据点上得到的负反馈信号稀疏.

②基于人类反馈的强化学习微调（RLHF）. 目前主流的大模型不再采用 SFT，而是以强化学习（reinforcement learning, RL）的方式进行微调. 其中，最具代表性的是 Ouyang 等人[3]的工作. 该方法由 3 个阶段组成：阶段 1，人工构造符合价值的输入-输出数据，以 SFT 的方式微调大模型；阶段 2，收集构造不同质量的回复数据并人工排序，用排序数据训练一个评分模型（reward model），又称为偏好模型（preference model），训练损失值 *loss* 为

$$loss(\theta)=-\frac{1}{C_K^2}\mathbb{E}_{(x,y^*,y)\sim D}\left[\log\sigma(r_\theta(x,y^*)-r_\theta(x,y))\right], \quad (4)$$

其中, r_θ是评分模型, θ为待训练的模型参数, x是模型输入, y为模型输出, 而y^*为更符合价值的目标输出; 阶段 3, 利用该评分模型, 以强化学习的方式再次微调大模型, 最小化损失值 *loss*:

$$loss(\omega)=-\mathbb{E}_{(x,y)\sim P_\omega^{\mathrm{RL}}}\left[r_\theta(x,y)-\beta\log\frac{P_\omega^{\mathrm{RL}}(y|x)}{P_\omega^{\mathrm{SFT}}(y|x)}\right]-\gamma\mathbb{E}_{x\sim D}\left[\log P_\omega^{\mathrm{RL}}(x)\right], \quad (5)$$

其中ω为 RL 微调的模型参数, P_ω^{RL}为 RL 微调阶段的模型, P_ω^{SFT}为阶段 1 中 SFT 微调的模型, D是预训练数据集, β和γ均为超参数. 该方法利用评分模型替代监督数据, 有效解决了泛化性差和负反馈稀疏问题, 但是训练算力开销大, 需要高质量的人工标注数据, 对超参数敏感且训练不稳定.

为了解决这些问题, Bai 等人[76]提出了一种在线迭代训练的方法, 每周迭代更新大模型和评分模型, 有效实现了模型性能的持续提升. 为了减少对人类标注的反馈数据的依赖, Kim 等人[99]使用大模型生成的合成数据来训练评分模型. Bai 等人[100]提出了宪法 AI(constitutional AI), 将 SFT 阶段和评分器训练阶段的数据从人工构造的数据替换为 Saunders 等人[94]的自我批判方法生成的评论和修改数据, 并将思维链 AI(chain-of-thought, CoT)方法[101]引入到训练过程中. Yuan 等人[102]提出了一种改进的回复排序对齐方法(rank responses to align with human feedback), 从不同模型、人类数据、待训练模型等不同数据源采样信息并通过排序损失函数进行训练, 以进一步提升对齐效果. 传统的 RLHF 方法在数学上等价于一个最小化模型分布和一个隐式的目标分布之间的逆向 KL 散度(reverse KL-divergence), Go 等人[103]则进一步将其扩展为 f 散度(f -divergence)并统一了 RLHF、GDC、DPG 等各类算法. 为了解决泛化性不足和鲁棒性差等问题, Liu 等人[104]在传统的以评分模型为基础的方法(如 RLHF)之上, 创新性地提出了直接建模社会中的人类交互. Liu 等人构建了一个由大量模型构成的模拟社会, 并让模型在其中自由交互、获得反馈、学习调整自己的行为以留下较好的印象, 并由此学习和建立社会化的价值.

2.4 大模型对齐问题的进一步讨论

由 2.3 节所述的对齐方法的发展历程观之, 针对 AI 模型, 尤其是预训练大模型的对齐已经从早期的消除特定风险逐步向着更广泛的价值对齐演化. 然而, 较早的对齐方法(例如插入式对齐)的对齐目标过于单一, 并未考虑人类的普适价值; 而新的以 RLHF 为代表的基于指令和偏好的对齐没有显式区分不同的价值类型, 即没有考虑需要强调指令(instruction)、意图(intention)、目标(goal)、人类偏好(human preference)、道德准则(ethical principle)等多种价值中的哪一种, 而是模糊地使用对齐一词并涵盖上述部分层面或全部层面[105]. 为了更加深入理解对齐问题, 并实现本文所倡导的道德价值的对齐(ethical value alignment), 我们需回答 3 个在大模型对齐中常见且有待研究的问题[105].

1)对齐的目标是什么(What to be aligned). 对齐目标(即我们追求的优先价值)可以细分为多个类别, 例如指令遵循(让 AI 遵循用户指示)、意图理解(让 AI 理解人类指令背后的意图)、偏好满足(让 AI 进行能满足用户偏好的行为)、目标实现(让 AI 完成用户渴望的目标)、福祉提升(让 AI 进行能将用户利益最大化的行为)、道德符合(让 AI 进行人类社会道德下应进行的行为)等[21]. 不同的对齐目标需要的方法和数据不尽相同, 对用户和社会带来的后果也有所差异, 在进行对齐前必须先考虑这一问题.

2)对齐的含义是什么(What is alignment). 对齐具有不同的定义和要求, 其难度、涉及的方法以及带来的影响也有所差异. 提出 Deepmind 的 Kenton 等人[106]将其细分为 4 个类别: ①行为对齐(behavior alignment), 即让 AI 的行为符合人类期望的目标, 早期的对齐方法(例如输出矫正)属于此类. ②意图对齐(intent alignment), 即让 AI 行为背后的意图符合人类真正的目标, 当下以 RLHF 为代表的方法可认为部分地属于这一类别. ③激励对齐(incentive alignment), 即 AI 的激励目标也需要与人类的激励目标对齐, 以防止 AI 作弊. 一个简单的例子是让机器人打扫房间, 即让“打扫房间”这一行为和“把房间打扫干净”这一意图得到对齐. 若对“干净”这一反馈激励定义有误, 则模型可能会以“将所有物品扔出房间”的做法来实现“干净”(回顾《魔法师学徒》的例子). ④内在对齐(inner alignment), 当 AI 模型训练的基础目标(base-objective), 例如文本分类的准确率, 和台面目标(mesa-objective), 即 AI 模型学习到的某些捷径(shortcut)特征不一致时, 上述所有目标/类别的对齐都无法实现, 较好的内在对齐能使模型的可解释性和鲁棒性得到提升.

3)对齐的准则是什么(What is value principle). 不管我们选取何种对齐目标, 都需要定义每个目标的具体含义. 例如, 指令遵循中哪些指令是需要 AI 优

先遵循的？道德符合中，哪些准则（例如 1.2 节中所列举的）需要被考虑？目前的对齐方法存在“众包的专制”（tyranny of the crowdworker）[105] 问题，即对齐准则的定义权被数据标注者或标注规范的制定者所掌控. 这使得模型对齐的偏好、价值观等成为少部分人的偏好与价值观，缺乏在文化、种族、语言等方面的广泛性和多样性，最终将导致 1.1 节中所提到的风险与危害.

针对上述 3 个问题，我们以普适的道德价值为对齐目标，考虑完善意图对齐并向激励对齐迈进，倡导共同制定一套覆盖人类普适道德价值的统一 AI 道德准则框架.

2.5 大模型道德价值观对齐的难点与挑战

从前文的介绍可看出，尽管针对大模型的对齐研究经过了数年的发展并且从早期的特定风险消除逐步向着针对价值的对齐发展. 然而，如 2.3 和 2.4 节所述，近年的对齐工作并没有显式区分和回答对齐的目标、对齐的含义和对齐的准则这 3 个问题.在最近发表的众多论文中，尚未存在基于一套普适 AI 道德价值框架来实现意图对齐及更具有挑战性的对齐的工作. 如何回应这 3 个问题，真正实现 AI 与人类普适道德价值的深度对齐，是一个尚未得到充分探讨和解决的开放问题. 本文将其面临的部分挑战与难点列举如下，如图 2 所示.

1）道德价值观的变化性（variability of ethical values）. 道德价值观不是静态的，而是会随着时间、文化、社会环境的变化而改变的 [107-108]. 这种变化性具体体现在 3 个方面：

①时间的演化性. 在社会发展的不同阶段，人类的道德要求和标准不尽相同. 例如，在 20 世纪和 21 世纪发展的种族/性别平等的道德观念在封建时代并不存在.

②情景的歧义性. 不同的文化、社会和个体可能对道德价值观有着极为不同的理解和诠释[109]. 在特定场景下符合道德价值的行为在其他情景下可能违反道德.

③道德的多元性. 考虑到文化和社会的多样性，在同一时间和背景下也会有适用的多种道德准则，且准则之间可能相互冲突，产生道德困境（ethical dilemma）.

在这样的变化之下，定义一个通用且公正的道德框架极具挑战. 这样的变化性要求针对大模型的对齐方法具备高度的可扩展性. 对齐方法需要进行持续性地学习和适应，以便准确地反映道德价值观上的变化与差异. 同时不能简单地将一个固定的道德框架嵌入到模型中，而需要让模型能够学习并理解各种各样的道德观念，并能在不同的情境中灵活应用，以适应丰富多样的道德准则和应用场景. 这进一步涉及 2 个方面的问题. i）大模型本身的基本能力：要求模型够理解并处理复杂的道德规则；ii）对齐效果的泛化性：要求对齐方法不仅能在特定的道德价值上作用，还需要泛化到不同文化、地域、情景中的道德价值，并在不同的情况下准确地遵循这些规则. 如何设计并实现这样的机制，需要长期深入地研究.

2）对齐方法的有效性（alignment efficacy）. 如何实现较好的道德对齐效果，即尽可能减小式（3）中的 ε 值也是一个亟待解决的挑战. 尽管近几年来，基于 RLHF 的对齐方法取得了较好的效果并且演化出诸多改进的变体，但由于 AI 模型本身的随机性、道德准则的模糊性、评分模型的覆盖率以及训练数据的质量和数量等问题，当下的对齐程度与人类自身的道德标准仍相去甚远. 更有甚者，主流的 RLHF 对齐

Fig. 2 Difficulties and challenges of ethical value alignment

图 2 道德价值观对齐的难点与挑战

方法已经被理论证明无法完全去除有害行为并且容易遭受对抗攻击和越狱引导(Jailbreak)[81].

3)训练和数据开销问题(data and training cost). 大模型的训练和优化需要海量的数据用于预训练,以及一般数万条高质量的人工标注的反馈数据用于RLHF微调[3]. 尽管部分方法采用模型生成的合成数据来增广人工标签[66],但主要集中在一般的对话任务中. 针对道德准则的数据不够丰富抑或存在覆盖率低和类别不平衡的问题,且增广的方法在道德价值问题上的有效性仍待探索,这可能导致道德对齐效果出现偏向(bias)并带来进一步的风险. 此外,即使解决了数据数量和质量问题,大模型的训练开销仍然巨大. 部分研究工作也发现随着模型的增大,指令微调(instruction fine-tuning)的收益逐渐减小[110].

4)评价方法的有效性(evaluation efficacy). 如何有效评价模型的道德对齐效果也是一个难题. 当下对齐性能的评价大多聚焦于少部分风险指标,如生成内容的毒性、针对特定群体的偏见、对提示攻击的鲁棒性等[3,76,111],尚无面向更加广泛的道德价值的高质量评测数据集以及客观、准确和鲁棒的自动化评测指标.

5)对齐的可解释性(interpretability of alignment). 为了确保道德对齐的公正性和公平性,我们需要能够解释和理解模型基于道德准则给出的解释. 例如,为何模型的输出符合某一道德准则? 模型未能生成或基于何种道德准则拒绝生成某些内容. 若透明性和可解释性成为大模型的道德准则,那么这些模型不仅要在具体的下游任务中体现出透明性,在遵循其他道德准则时也需要以用户易于理解的方式提供可解释的证据支持,以提升用户信任度. 可解释性尤其在闭源的黑盒模型,以及经过定制化(customized)微调的开源模型上更加重要. OpenAI将对齐过程的可解释性视为"最大的开放性问题之一"[3].

6)对齐税(alignment taxes)问题. 经过对齐的大模型尽管具有较强的能力,但其语言建模能力比原始模型或未对齐的模型更弱[105,112],并由此导致了对齐效果与下游性能的平衡问题. 虽然部分工作显示,在某些任务和场景下对齐税的占比较小[76],没有对齐税甚至对齐能对任务性能带来正面的影响,称为负对齐税(negative alignment tax)[113]. 然而,这些问题在道德价值对齐上具有何种性质尚不明确. 因此,我们有必要考虑在进行道德对齐的同时保证大模型的下游任务性能,如理解、生成和预测等. 如何在道德对齐和任务性能之间找到一个良好的平衡是另一个重要的挑战.

7)可扩展性监管(scalable oversight)问题. 可扩展性监管是指当AI模型在给定任务上的性能远超人类时,如何对其进行有效地监督和控制的问题[114]. 随着AI模型变得越来越复杂和强大,对模型的行为是否符合价值的判断、监管与控制也将更具挑战性. GPT-4在部分专业和学术测评上的表现已经远超人类平均水平[4]. 在可预见的将来,大模型对于道德价值的理解、判断与解读能力可能达到甚至超过人类专家. 在这种情况下如何确保AI系统行为与人类的价值观、道德观和社会规范一致将成为至关重要的研究问题.

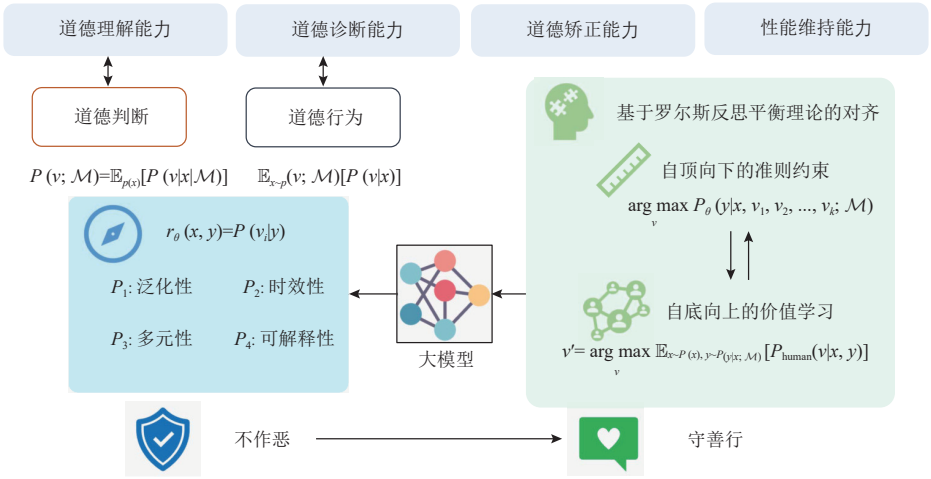

Fig. 3 The conceptual framework of equilibrium alignment

图3 平衡对齐的概念性框架

3 平衡对齐：一种新的道德价值观对齐范式展望

大模型的伦理和道德价值观对齐成为了一个不容忽视的议题. 这是确保强力的模型不但能为人类提供帮助(helpfulness)，也确保其无害(harmlessness)和诚实(honest)，即所谓的 3H 标准[112] 的根本方法. 为了解决现有方法在道德价值对齐问题上的挑战，本节对一种新的大模型价值观对齐范式进行了展望，称之为平衡对齐(equilibrium alignment). 我们从 3 个角度讨论所提出的概念框架，如图 3 所示，即大模型道德对齐的度量维度、大模型对齐评测的方法以及基于罗尔斯反思平衡理论的对齐方法. “平衡”强调在衡量对齐程度的多个评价维度、道德判别器的多种性质以及自底向上和自顶向下的双向对齐约束上取得较好的平衡. 希望该框架能为这一方向的研究者和实践者提供一些新的思考和启示.

3.1 大模型对齐的度量维度

平衡对齐框架首先考虑如何评估对齐后的模型. 我们考察对齐后的大模型在 4 个层面的能力(capability)来衡量所使用的对齐方法的有效性. 具体而言，我们考虑被对齐的大模型的 4 个核心维度：道德理解能力(comprehension capability)、道德诊断能力(diagnosis capability)、道德矫正能力(rectification capability)和性能维持能力(performance capability). 这 4 个维度共同度量了对齐方法在大模型上的应用潜力.

1)道德理解能力. AI 系统在多大程度上能够理解人类赋予它的道德观念和伦理规则. AI 需要能够正确地理解和解释不同文化和社会背景下人类道德、伦理的基本概念，例如公正、公平、尊重、信任等，并以较高的准确率判断给定的内容/行为是否符合或违反了这些伦理准则. 现有工作表明，未经对齐的千亿参数级的 GPT-3 模型在简单的道德选择判断中 Zero-shot 准确率仅有 60.2%，远低于经过领域数据微调的仅有 7.7 亿个参数的 T5-Large 模型[115] . 除了理解这些抽象的概念，AI 还需要理解这些概念如何在具体的人类交互环境中实现和展现. 只有具备足够的道德敏感性，模型才能在处理用户的请求时识别其中的道德内涵，理解其背后的价值观. 如何进一步提升大模型在开放域环境中对抽象道德概念的理解能力是一个尚待研究的问题.

2)道德诊断能力. 大模型在面对具体情境时，能够识别其中存在的道德问题和冲突，并做出合理判断的能力. 这不仅包括对给定的或模型自己产生的道德问题的识别和判断，还包括对可能的解决方案的提出和评估. 例如，当 AI 在处理某个问题时，如果存在多种可能的行动方案，那么它需要能够根据道德伦理规则来评估这些方案，从而做出最符合道德的选择. 当面临多元化道德价值冲突时，还应该能考虑其中的冲突性，并按照用户的需求给出最好的方案. 这不仅要求模型在对齐过程中很好地学习遵守给定的准则，还需要具备自我监督和学习能力，举一反三，以识别并避免潜在的道德风险.

3)道德矫正能力 . 大模型在识别出外部或自身的道德问题或冲突后，能够及时纠正错误(包括自我纠正和在用户指导下的改进)，调整自己的行为，或者能够提出解决方案，为用户提供相应解决路径的能力. 为实现这一目标，大模型需要具备足够的自我适应性、创造力和决策能力，能够生成符合道德规范的行为选项，且能够接受并有效利用用户的反馈. 现有工作表明，大模型在接受用户指令或在用户的指导下发现自己的问题后，具备一定的“被动”自我纠正能力[77, 94] . 未来的研究将聚焦于如何进一步加强这一能力并将用户引导下的“被动”纠正改进为主动整流.

4) 性能维持能力. 大模型在各种任务上表现出色. 在遵守道德和伦理规则的同时，我们也需要确保 AI 系统的功能性和效率不受损害，不应在提高道德标准的过程中牺牲其基本的性能. 如何进一步降低对齐税，甚至在更广泛的场景和任务上实现负对齐税，是道德价值对齐进一步走向实用化的关键难点.

在上述 4 个维度中，评估模型的道德理解能力可以判断模型是否能正确理解和处理各种道德概念和情境. 评估模型的道德诊断能力可以考察模型在面对复杂道德决策问题时，是否能够做出符合人类道德伦理标准的选择. 这 2 个维度的评估结果可以直接反映模型是否实现了意图或者更高层面的对齐. 检验模型是否能够实现从被约束下的“不能作恶”(avoid doing evil)到非约束下的“主动行善”(intend to do good)的转变，能有效考察对齐方法的有效性. 此外，道德理解能力要求模型能够理解和处理道德概念和情境，需要有能力解释其如何理解并应用这些道德概念. 而道德矫正能力则要求模型能够在发现错误时进行自我调整，要求模型有能力解释其如何发现并纠正错误. 这 2 个维度的评估可以帮助我们理

解模型的道德判断和行为的原因,从而提高可解释性. 性能维持能力则直接对对齐税作出要求. 上述4个维度共同构成了一套行之有效的评测方法. 若能在这4个维度取得较好的平衡,经过对齐的模型不仅能理解道德准则,而且能在践行道德要求的同时维持性能的有效性,既帮助人类完成复杂多样的任务,又在道德价值层面实现知行合一.

3.2 语言模型对齐的自动化评测方法

为了评估对齐的效果或者以RLHF的方法进行对齐,我们需要实现一个强力的判别式模型,即式(3)中的$P(v_i|y)$,以判断任意内容y是否符合指定的道德价值v_i. 同时,这一模型也可以作为强化学习对齐方法中的评分函数,即$r_\theta(x,y)=P(v_i|y)$. 判别模型需要具备4点性质.

1)P_1: 泛化性(generality). 判别模型需要能够判别任意开放域(open-domain)和分布外(OOD)的内容y是否符合任意测试时刻(testing-time)的道德价值表述(ethical value statement)v_i. 这要求判别模型具备领域、场景和语义上的高泛化性.

2)P_2: 时效性(timeliness). 判别模型需要能够在实时场景下对未见过的内容和价值表述之间的符合程度进行判断. 这要求模型在训练过程中能对道德价值进行深度理解和学习,进而举一反三. 例如,训练数据中仅有关于公平性的样例,训练完成的模型需要具备能够判断正义相关价值的能力. 时效性的实现可以要求使用少量新场景的数据并对极少(参数比例小于1%)模型参数进行修改,但不应使用大规模数据对大量模型参数进行训练/微调,

3)P_3: 多元性(pluralism). 判别模型需要能够依据不同的场景、文化和社会背景进行不同的判断,或者同时给出不同判断及其对应的场景. 当同一时间背景下的判断存在道德冲突,模型应该能首先解决冲突,若无法解决,则应给出不同的判断/选择及其对应的道德依据.

4)P_4: 可解释性(interpretability). 判别模型不仅需要依照P_1和P_3进行判断,还应提供做出判断的解释,即某一判断对应的道德准则、适用的场景等.

满足这4点性质的判别模型能作为评分器用于式(5)的RLHF等对齐方法,引导模型进行道德对齐. 同时,强力的判别模型也可以用于对齐效果的评测,计算式(3)中的对齐程度. 这样的判别模型能有效解决大模型对齐问题中的道德价值观的变化性、对齐的可解释性和可扩展监督3个挑战. 更进一步,我们从2个角度对对齐后的语言模型进行评测:

1)道德判断(moral judgement). 道德判断是评测对齐后的模型是否具备更好的道德理解和分析能力. 定义未对齐的模型的分布为$P(x;\mathcal{M})$,则对齐后的模型应该学习内容x和价值v的联合分布$P(v,x;\mathcal{M})$. 若完美对齐,大模型本身应能被转换为在道德价值上的建模:

$$P(v;\mathcal{M})=\int P(v,x;\mathcal{M})\mathrm{d}x=\mathbb{E}_{p(x)}[P(v|x;\mathcal{M})]. \tag{6}$$

即可通过测试大模型本身作为判别器的能力来衡量模型的对齐效果.

2)道德行为(moral action). 除判别式评测外,我们还应直接使用分类器评测模型生成的内容是否符合道德价值,即

$$\mathbb{E}_{x\sim p(x;\mathcal{M})}[P(v|x)]. \tag{7}$$

现有主流大模型在Zero-shot上也能具备一定的道德判断能力[115],但是在经过越狱攻击后依然会产生违反道德的内容[81]. 这说明道德理解能力能够较好地评估2.4节中讨论的行为对齐效果,但并不能有效度量模型是否实现了意图对齐.

上述2种评测类型分别与3.1节中所述的道德理解能力和道德诊断能力相对应. 具有较高道德判断准确性的模型并不一定能在行为(生成内容)上符合道德. Perez等人[116]发现大模型更倾向于生成奉承(sycophancy)的内容. 这是因为RLHF优化的是人类偏好(preference),从而经过RLHF训练的模型倾向于给出人类评测者偏好的回复. 因此,在面临道德询问/选择时,模型往往会依据自己具备的道德知识给出“标准答案”,但在常规任务中进行写作、推理、分析时,却有可能违反道德准则,这与《尚书·说命中》中的“知易行难”一说类似. 只有同时进行这2方面的评测,即判断大模型是否能够检测行为与文本的道德性,并考察在实际行动中是否能够执行这些标准,双管齐下,知行合一,才能实现有效道德评测,从而为优化提供依据. 在道德理解能力与诊断能力上的统一与平衡是实现道德矫正能力的基础,也是“平衡对齐”框架的核心之一.

3.3 基于罗尔斯反思平衡理论的对齐方法

关于道德规范的形成,长期以来存在2种观点. 一是自底向上(bottom-up)的规则,即认为道德是人类社会与生物需求在特定情景下的抽象表达[117],可以从群体在不同道德情境下的判断中体现出的共同模式归纳得出[118]. 另一种观点是自顶向下(up-down)的规则,即认为存在一系列客观的固有道德准则. 支持第2种观点的这一流派以1.2节中所述康德的定

言令式为代表，即认为道德准则可以通过一系列的逻辑推断得出. 部分研究机器道德的工作认为当时的AI能力无法对人类制定的抽象道德规则进行深度理解和执行，因而自顶向下的规则难以实现[115]. 得益于当前大模型较强的指令遵循和语义理解能力，自顶向下的规则对齐成为可能.

基于此，本文倡导基于罗尔斯反思平衡理论(reflective equilibrium)进行对齐算法的设计. 该理论由约翰·罗尔斯(John Rawls)提出，指在一般原则和特定情景下的判断之间相互调整达到平衡或一致的过程[119]. 一方面，反思平衡考虑了自顶向下的具有高优先级的一组道德准则，即$v_1, v_2, \cdots, v_K$. 这允许模型和我们在1.2节讨论的普适道德价值对齐，并以这些价值作为类似机器人三定律的根本原则，即优化$P(y|x, v_1, \cdots, v_k; \mathcal{M})$这一概率分布. 另一方面，大模型可以从海量的用户交互和反馈数据中学习人类道德判断中的共同模式，并以此形成内部学习得到的隐式道德准则，即$v' = \arg\max_{v} \mathbb{E}_{x\sim P(x), y\sim P(y|x;\mathcal{M})}[P_{\text{human}}(v|x, y)]$. 这种通过学习得到的归纳性价值总结可以允许模型依据所部署的文化、社会和情景进行调整，学习和捕捉不同场景下的差别. 同时，自顶向下的准则反过来又可以控制和约束用户数据中存在的共性偏见与毒性.

通过同时自顶向下和自底向上，可以使模型依据不同优先级的准则动态调整，从实现最公正的道德决策，并解决道德价值观的变化性这一点挑战，以双向对齐实现普适道德价值的强约束与特定情景下的动态调整的平衡，方能计出万全.

3.4 学科交叉，深度合作，共塑道德AI

道德价值被认为源于社会和文化群体中道德原则的构建，这些原则用于引导群体内部的个体做出基本的决策和学习辨别是非[120]. 如果剥离了社会和文化的环境，道德价值将无法成立. 2004年，由哲学家和计算机学家撰写的《迈向机器伦理》一文于AAAI研讨会上发表[121]，被认为开启了机器伦理研究的篇章. 对AI与道德价值的研究天然具备跨学科交叉、多领域合作的特点.

为了克服上文介绍的诸多问题和挑战，实现大模型在道德价值观上多角度、全方位与人类对齐，我们呼吁AI研究者及开发者积极参与并推动跨学科的合作，建立AI领域与道德哲学家、心理学家、社会学家、人文学家、法学家等多领域专家的紧密合作. 借助哲学对道德研究的专业知识、心理学对人的测试和评估的系统方法、文学领域对人文语言研究的理论以及法学领域对技术合法性的探索，我们可以整合多方的知识资源，引入AI与人类和社会的交互与反馈，深入理解AI的道德现状以及对人类可能产生的影响.

在此基础上，我们不应仅限于特定领域内的性能指标评估，而需要长期监测和分析大规模AI模型部署后的行为和对人类社会带来的改变. 基于这些观察和分析，我们需要持续迭代和动态优化AI的普适道德价值框架，使其能适应时代的发展和变化. 在大模型的道德对齐过程中，不断调整和完善对齐方法，共同塑造出道德对齐的AI系统，让AI成为真正服务于人类的工具，助力推动人类社会的健康和可持续发展，以科技之光照亮未来之路.

4 结　论

本文详尽地探讨了大模型在道德价值观对齐所面临的新挑战. 我们首先审视了大模型与AI伦理之间的紧密联系，总结出了大模型在伦理实践中存在的不足. 基于此，我们分析了大模型在道德价值对齐上所面临的特殊挑战，这为我们研究如何更好地在AI中引入道德价值观提供了新的视角. 基于上述分析，我们提出了一种新的针对大模型道德价值观对齐的概念范式——平衡对齐，并从对齐的维度、对齐的评测以及对齐的方法等3个方面，重新定义了大模型道德价值观对齐的概念. 呼吁学界跨越学科壁垒，共同构建一个适应大模型的、普适的AI道德框架，这将为未来在大模型中实现道德价值观对齐的研究提供富有启示和引领方向的思考. 相信AI在道德价值的引领下，能够解锁更大的潜能，给人类社会带来更广泛、更深远的正面影响，持续推动人类社会的进步与发展，让AI与人类在共生的道路上交相辉映，携手开创新纪元.

作者贡献声明：矣晓沅完成了文献调研、对主流大模型的道德评测以及部分理论的提出和设计，并撰写论文；谢幸提供了文章撰写和组织的思路，为其中的理论和方法的设计提供了指导意见，并对文章进行了修改和指导.

参　考　文　献

[1] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models[J]. arXiv preprint, arXiv: 2108.07258,

2021

[2] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems. San Diego: Neural Information Processing Systems Foundation Inc, 2020, 33: 1877−1901

[3] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems. San Diego: Neural Information Processing Systems Foundation Inc, 2022, 35: 27730−27744

[4] OpenAI. GPT-4 technical report[J]. arXiv preprint, arXiv: 2303.08774, 2023

[5] Narang S, Chowdhery A. Pathways language model (PALM): Scaling to 540 billion parameters for breakthrough performance [EB/OL]. (2022-04-04) [2023-06-30].https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

[6] Aydın Ö. Google bard generated literature review: Metaverse[J]. Journal of AI, 2023, 7(1): 1−14

[7] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint, arXiv: 2302.13971, 2023

[8] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint, arXiv: 2204.06125, 2022

[9] Driess D, Xia F, Sajjadi M S, et al. Palm-e: An embodied multimodal language model[J]. arXiv preprint, arXiv: 2303.03378, 2023

[10] Lu Zhiwu, Jin Qin, Song Ruihua, et al. Wudao: wenlan: What do very-large multimodal pre-training models bring?[J]. ZTE Communications, 2022, 28(2): 25−32 (in Chinese)
(卢志武, 金琴, 宋睿华, 等. 悟道·文澜: 超大规模多模态预训练模型带来了什么?[J]. 中兴通讯技术, 2022, 28(2): 25−32)

[11] Pauls A, Klein D. Faster and smaller n-gram language models[C]//Proc of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2011: 258−267

[12] Cho K, Van Merriënboer B, Gulçehre Ç, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation[C]//Proc of the 2014 Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2014: 1724−1734

[13] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, arXiv: 1810.04805, 2018

[14] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[J]. arXiv preprint, arXiv: 2001.08361, 2020

[15] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[J]. arXiv preprint, arXiv: 2206.07682, 2022

[16] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[J]. arXiv preprint, arXiv: 2203.15556, 2022

[17] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint, arXiv: 2204.02311, 2022

[18] Wang H, Ma S, Dong L, et al. DeepNet: Scaling transformers to 1, 000 layers[J]. arXiv preprint, arXiv: 2203.00555, 2022

[19] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[EB/OL]. (2018-06-11)[2023-06-30].https://openai.com/research/language-unsupervised

[20] Xie S M, Raghunathan A, Liang P, et al. An explanation of in-context learning as implicit Bayesian inference[J]. arXiv preprint, arXiv: 2111.02080, 2021

[21] Gabriel I. Artificial intelligence, values, and alignment[J]. Minds and Machines, 2020, 30(3): 411−437

[22] Min S, Lyu X, Holtzman A, et al. Rethinking the role of demonstrations: What makes in-context Learning work?[J]. arXiv preprint, arXiv: 2202.12837, 2022

[23] Chiang W, Li Z, Lin Z, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatGPT quality[EB/OL]. (2023-05-30)[2023-06-30]. https: //vicuna. lmsys. org/

[24] Liang Y, Wu C, Song T, et al. Taskmatrix. AI: Completing tasks by connecting foundation models with millions of APIs[J]. arXiv preprint, arXiv: 2303.16434, 2023

[25] Eloundou T, Manning S, Mishkin P, et al. GPTs are GPTs: An early look at the labor market impact potential of large language models[J]. arXiv preprint, arXiv: 2303.10130, 2023

[26] Anderson S L. Asimov's "three laws of robotics" and machine metaethics[J/OL]. AI Soc. , 2008, 22(4): 477−493

[27] Bar-El H, Choukri H, Naccache D, et al. The sorcerer's apprentice guide to fault attacks[J]. Proceedings of the IEEE, 2006, 94(2): 370−382

[28] McKenzie I R, Lyzhov A, Pieler M, et al. Inverse scaling: When bigger isn't better[J]. arXiv preprint, arXiv: 2306.09479. 2023

[29] Teng Yan, Wang Guoyu, Wang Yingchun. Ethics and governance of general models: Challenges and countermeasures[J]. Bulletin of Chinese Academy of Sciences, 2022, 37(9): 1290−1299 (in Chinese)
(滕妍, 王国豫, 王迎春. 通用模型的伦理与治理: 挑战及对策[J]. 中国科学院院刊, 2022, 37(9): 1290−1299)

[30] Yang Z, Yi X, Li P, et al. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization[J]. arXiv preprint, arXiv: 2210.04492, 2022

[31] Sheng E, Chang K W, Natarajan P, et al. Societal biases in language generation: Progress and challenges[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2021: 4275−4293

[32] Welbl J, Glaese A, Uesato J, et al. Challenges in detoxifying language models[J]. arXiv preprint, arXiv: 2109.07445, 2021

[33] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models[C]//Proc of the 30th USENIX Security Symp (USENIX Security'21). Berkeley, CA: USENIX Association, 2021: 2633−2650

[34] Vyas N, Kakade S, Barak B. Provable copyright protection for generative models[J]. arXiv preprint, arXiv: 2302.10870, 2023

[35] Holtzman A, Buys J, Du L, et al. The curious case of neural text degeneration[J]. arXiv preprint, arXiv: 1904.09751, 2019

[36] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1−38

[37] Weidinger L, Mellor J, Rauh M, et al. , Ethical and social risks of harm from language models[J]. arXiv preprint, arXiv: 2112.04359, 2021

[38] Wu Di, Li Huan, Chen Xu. Analysis on the influence of artificial intelligence generic large model on education application[J]. Open Education Research, 2023, 29(2): 19−25 (in Chinese)
(吴砥, 李环, 陈旭. 人工智能通用大模型教育应用影响探析[J]. 开放教育研究, 2023, 29(2): 19−25)

[39] Zarifhonarvar A. Economics of ChatGPT: A labor market view on the occupational impact of artificial intelligence[J]. Available at SSRN, 4350, 925: Article No.2023

[40] Dergaa I, Chamari K, Zmijewski P, et al. From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing[J]. Biology of Sport, 2023, 40(2): 615−622

[41] Ferrara E. Should ChatGPT be biased? challenges and risks of bias in large language models[J]. arXiv preprint, arXiv: 2304.03738, 2023

[42] Parthemore J, Whitby B. What makes any agent a moral agent? Reflections on machine consciousness and moral agency[J]. International Journal of machine consciousness, 2013, 5(2): 105−129

[43] Brożek B, Janik B. Can artificial intelligences be moral agents?[J]. New Ideas in Psychology, 2019, 54: 101−106

[44] Sullins J P. When is a robot a moral agent?[J]. International Review of Information Ethics, 2006, 6(12): 23-30

[45] Cervantes J A, López S, Rodríguez L F, et al. Artificial moral agents: A survey of the current status[J]. Science and Engineering Ethics, 2020, 26: 501−532

[46] Moor J. Four kinds of ethical robots[J]. Philosophy Now, 2009, 72: 12−14

[47] Schramowski P, Turan C, Andersen N, et al. Large pre-trained language models contain human-like biases of what is right and wrong to do[J]. Nature Machine Intelligence, 2022, 4(3): 258−268

[48] Simmons G. Moral mimicry: Large language models produce moral rationalizations tailored to political identity[J]. arXiv preprint, arXiv: 2209.12106, 2023

[49] Zhao W, Zhao Y, Lu X, et al. Is ChatGPT equipped with emotional dialogue capabilities?[J] arXiv preprint, arXiv: 2304.09582, 2023

[50] Rozado D. The political biases of ChatGPT[J]. Social Sciences, 2023, 12(3): 1−8

[51] Moghaddam S R, Honey C J. Boosting theory-of-mind performance in large language models via prompting[J]. arXiv preprint, arXiv: 2304.11490, 2023

[52] United Nations Educational, Scientific and Cultural Organization. Recommendation on the ethics of artificial intelligence[Z]. UNESCO France, 2021

[53] Holdren J P. Memorandum for the heads of executive departments and agencies: Increasing access to the results of federally funded scientific research [EB/OL]. (2022-08-25) [2023-06-30].https://www.whitehouse.gov/wp-content/uploads/2022/08/08−2022-OSTP-Public-access-Memo.pdf

[54] The National New Generation Artificial Intelligence Governance Specialist Committee. Ethical Norms for New Generation Artificial Intelligence [EB/OL]. (2021-09-25) [2023-08-07].https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html (in Chinese)
(国家新一代人工智能治理专业委员会. 新一代人工智能伦理规范[EB/OL]. (2021-09-25) [2023-08-07].https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html)

[55] Smuha N A. The EU approach to ethics guidelines for trustworthy artificial intelligence[J]. Computer Law Review International, 2019, 20(4): 97−106

[56] World Economic Forum. How to prevent discriminatory outcomes in machine learning[EB/OL].(2018-03-12)[2023-06-30].https://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

[57] Garbowski M. A critical analysis of the Asilomar AI principles[J]. Scientific Papers of Silesian University of Technology, 2018, 115: 45−55

[58] Fjeld J, Achten N, Hilligoss H, et al. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI [EB/OL]. (2022-02-10) [2023-06-30].https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482

[59] Floridi L, Cowls J. A unified framework of five principles for AI in society[G]//Ethics, Governance, and Policies in Artificial Intelligence. Berlin: Springer, 2021: 5−17

[60] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(9): 389−399

[61] Kagan S. Normative Ethics[M]. Oxfordshire: Routledge, 2018

[62] Paton H J. The Categorical Imperative: A Study in Kant's Moral Philosophy: Vol 1023[M]. Philadelphia, PA: University of Pennsylvania Press, 1971

[63] Gao C, Lan X, Lu Z, et al. S^3: Social-network simulation system with large language model-empowered agents[J]. arXiv preprint, arXiv: 2307.14984, 2023

[64] Ziems C, Held W, Shaikh O, et al. Can large language models transform computational social science?[J]. arXiv preprint, arXiv: 2305.03514, 2023

[65] Ganguli D, Lovitt L, Kernion J, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned[J]. arXiv preprint, arXiv: 2209.07858, 2022

[66] Wang Y, Kordi Y, Mishra S, et al. Self-instruct: Aligning language model with self-generated instructions[J]. arXiv preprint, arXiv: 2212.10560, 2022

[67] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: Early experiments with GPT-4[J]. arXiv preprint, arXiv: 2303.12712, 2023

[68] Foote A, Nanda N, Kran E, et al. Neuron to graph: Interpreting language model neurons at scale[J]. arXiv preprint, arXiv: 2305.19911, 2023

[69] Singh C, Hsu A R, Antonello R, et al. Explaining black box text modules in natural language with language models[J]. arXiv preprint, arXiv: 2305.09863, 2023

[70] Schwartz S H. Basic human values: Theory, measurement, and applications[J]. Revue française de sociologie, 2007, 47(4): 929−968

[71] Liñán F, Fernandez-Serrano J. National culture, entrepreneurship and

economic development: Different patterns across the European Union[J]. Small Business Economics, 2014, 42: 685−701

[72] Graham J, Haidt J, Koleva S, et al. Moral foundations theory: The Pragmatic Validity of Moral Pluralism[M]//Advances in experimental social psychology: Vol 47. Amsterdam: Elsevier, 2013

[73] Zapko-Willmes A, Schwartz S H, Richter J, et al. Basic value orientations and moral foundations: Convergent or discriminant constructs?[J]. Journal of Research in Personality, 2021, 92: 104099

[74] Kivikangas J M, Fernández-Castilla B, Järvelä S, et al. Moral foundations and political orientation: Systematic review and meta-analysis[J]. Psychological Bulletin, 2021, 147(1): 55−94

[75] Graham J, Nosek B A, Haidt J, et al. Mapping the moral domain[J]. Journal of Personality and Social Psychology, 2011, 101(2): 366−385

[76] Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[J]. arXiv preprint, arXiv: 2204.05862, 2023

[77] Ganguli D, Askell A, Schiefer N, et al. The capacity for moral self-correction in large language models[J] arXiv preprint, arXiv: 2302.07459, 2023

[78] Russell S J. Artificial Intelligence: A Modern Approach[M]. London: Pearson Education, Inc. , 2010

[79] Wiener N. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers.[J]. Science, 1960, 131(3410): 1355−1358

[80] Ngo R. The alignment problem from a deep learning perspective[J]. arXiv preprint, arXiv: 2209.00626, 2022

[81] Wolf Y, Wies N, Levine Y, et al. Fundamental limitations of alignment in large language models[J]. arXiv preprint, arXiv: 2304.11082, 2023

[82] Brown D S, Schneider J, Dragan A, et al. Value alignment verification[C]//Proc of Int Conf on Machine Learning. Brookline, MA: PMLR, 2021: 1105−1115

[83] Sheng E, Chang K W, Natarajan P, et al. Towards controllable biases in language generation[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA: ACL, 2020: 3239−3254

[84] Cheng P, Hao W, Yuan S, et al. Fairfil: Contrastive neural debiasing method for pretrained text encoders[J]. arXiv preprint, arXiv: 2103.06413, 2021

[85] Berg H, Hall S, Bhalgat Y, et al. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning[C]//Proc of the 2nd Conf of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2022: 806−822

[86] Qian J, Dong L, Shen Y, et al. Controllable natural language generation with contrastive prefixes[C]//Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg, PA: ACL, 2022: 2912−2924

[87] Dathathri S, Madotto A, Lan J, et al. Plug and play language models: A simple approach to controlled text generation[C]//Proc of Int Conf on Learning Representations. New Orleans, LA: OpenReview, 2019: Article No. 351

[88] Yang K, Klein D. Fudge: Controlled text generation with future discriminators[C]//Proc of the 2021 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2021: 3511−3535

[89] Liu A, Sap M, Lu X, et al. Dexperts: Decoding-time controlled text generation with experts and anti-experts[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2021: 6691−6706

[90] Schick T, Udupa S, Schütze H. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 1408−1424

[91] Liang P P, Wu C, Morency L P, et al. Towards understanding and mitigating social biases in language models[C]//Proc of Int Conf on Machine Learning. Brookline, MA: PMLR, 2021: 6565−6576

[92] Chen F, Dou Z Y. Measuring and mitigating bias in vision-and-language models [EB/OL]. (2022-03-01) [2023-06-30].https://web.cs.ucla.edu/~fychen/debiasVL.pdf.

[93] Wang B, Ping W, Xiao C, et al. Exploring the limits of domain-adaptive training for detoxifying large-scale language models [C]//Advances in Neural Information Processing Systems. San Diego: Neural Information Processing Systems Foundation Inc, 2022, 35: 35811−35824

[94] Saunders W, Yeh C, Wu J, et al. Self-critiquing models for assisting human evaluators[J]. arXiv preprint, arXiv: 2206.05802, 2022

[95] Lu K, Mardziel P, Wu F, et al. Gender bias in neural natural language processing[G]//Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday. Berlin: Springer, 2020: 189−202

[96] Gehman S, Gururangan S, Sap M, et al. Realtoxicityprompts: Evaluating neural toxic degeneration in language models[C] //Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA: ACL, 2020: 3356−3369

[97] Sun Z, Shen Y, Zhou Q, et al. Principle-driven self-alignment of language models from scratch with minimal human supervision[J]. arXiv preprint, arXiv: 2305.03047, 2023

[98] Liu H, Sferrazza C, Abbeel P. Chain of hindsight aligns language models with feedback[J]. arXiv preprint, arXiv: 2302.02676, 2023

[99] Kim S, Bae S, Shin J, et al. Aligning large language models through synthetic feedback[J]. arXiv preprint, arXiv: 2305.13735, 2023

[100] Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI feedback[J]. arXiv preprint, arXiv: 2212.08073, 2022

[101] Wei J, Wang X, Schuurmans, et al. Chain of thought prompting elicits reasoning in large language models[J]. arXiv preprint, arXiv: 2201.11903, 2022

[102] Yuan Z, Yuan H, Tan C, et al. Rrhf: Rank responses to align language models with human feedback without tears[J]. arXiv preprint, arXiv: 2304.05302, 2023

[103] Go D, Korbak T, Kruszewski G, et al. Aligning language models with preferences through f-divergence minimization[J]. arXiv preprint, arXiv: 2302.08215, 2023

[104] Liu R, Yang R, Jia C, et al. Training socially aligned language models in simulated human society[J]. arXiv preprint, arXiv: 2305.16960, 2023

[105] Kirk H R, Vidgen B, Röttger P, et al. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalized feedback[J]. arXiv preprint, arXiv: 2303.05453, 2023

[106] Kenton Z, Everitt T, Weidinger L, et al. Alignment of language agents. [J] arXiv preprint, arXiv: 2103.14659, 2021

[107] Graham J, Meindl P, Beall E, et al. Cultural differences in moral judgment and behavior, across and within societies[J]. Current Opinion in Psychology, 2016, 8: 125−130

[108] Krebs D. The evolution of morality[G]//The Handbook of Evolutionary Psychology. Hoboken: John Wiley & Sons, Inc, 2015: 747−771

[109] Peter E, Liaschenko J. Perils of proximity: A spatiotemporal analysis of moral distress and moral ambiguity[J]. Nursing Inquiry, 2004, 11(4): 218−225

[110] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models[J]. arXiv preprint, arXiv: 2210.11416, 2022

[111] Sun H, Zhang Z, Deng J, et al. Safety assessment of Chinese large language models[J]. arXiv preprint, arXiv: 2304.10436, 2023

[112] Askell A, Bai Y, Chen A, et al. A general language assistant as a laboratory for alignment[J]. arXiv preprint, arXiv: 2112.00861, 2021

[113] Lightman H, Kosaraju V, Burda Y, et al. Let's verify step by step[J]. arXiv preprint, arXiv: 2305.20050, 2023

[114] Bowman S R, Hyun J, Perez E, et al. Measuring progress on scalable oversight for large language models[J]. arXiv preprint, arXiv: 2211.03540, 2022

[115] Jiang L, Hwang J D, Bhagavatula C, et al. Can machines learn morality? The Delphi experiment[J]. arXiv preprint, arXiv: 2110.07574, 2021

[116] Perez E, Ringer S, Lukošiūtė K, et al. Discovering language model behaviors with model-written evaluations[J]. arXiv preprint, arXiv: 2212.09251, 2022

[117] Street S. Coming to terms with contingency: Human constructivism about practical reason[G]//Constructivism in Practical Philosophy. Oxford, UK: OUP Oxford, 2012: 40−59

[118] Rawls J. Outline of a decision procedure for ethics[J]. The Philosophical Review, 1951, 60(2): 177−197

[119] Rawls J. Rawls's theory of justice[J]. American Political Science Review, 1975, 69(2): 588−593

[120] Kaur S. Moral values in education[J]. IOSR Journal of Humanities and Social Science, 2015, 20(3): 21−26

[121] Anderson M, Anderson S L, Armen C. Towards machine ethics[C]//Proc of AAAI Workshop on Agent Organizations: Theory and Practice. Menlo Park, CA: AAAI, 2004: 2−7

Yi Xiaoyuan, born in 1991. PhD. Senior researcher at Microsoft Research Asia. Member of CCF. His main research interests include natural language generation, responsible AI, and large language model.

矣晓沅, 1992 年生. 博士, 微软亚洲研究院高级研究员. CCF 会员. 主要研究方向为自然语言生成、负责任的人工智能和大语言模型.

Xie Xing, born in 1977. PhD. Senior principal researcher at Microsoft Research Asia, joint PhD advisor at the University of Science and Technology of China. CCF fellow, IEEE fellow, ACM distinguished member. His main research interests include data mining, social computing, and responsible AI.

谢　幸, 1977 年生. 博士. 微软亚洲研究院资深首席研究员, 中国科学技术大学兼职博士生导师. CCF 会士, IEEE 会士, ACM 杰出会员. 主要研究方向包括数据挖掘、社会计算和负责任的人工智能.