目录

一　绪论

随着人工智能（Artificial Intelligence, AI）技术的发展，基于深度学习的系统对社会产生了越来越大的影响，但风险也与日俱增 [1]。在 AI 可能引发的重大风险中，“对齐失败” 成为了一个主要的来源。所谓 AI 对齐，是指确保 AI 能够符合人类的意图和价值观，从而作出符合人类需求的行为，而对齐研究，即是指为解决不对齐系统的风险进行的研究。

在 AI 对齐领域，有着如下的 RICE 原则作为指导：鲁棒性（Robustness），即保证人工智能抵抗环境干扰的能力；可解释性（Interpretability），即保证人工智能操作或决策的过程能被人类理解；可控制性（Controllability），即保证人工智能受人类指导、控制运行；道德性（Ethicality），即保证人工智能遵守社会行为规范与普世价值观。

人工智能高速发展的过程中，AI 对齐的研究事实上是严重落后的。就以当今常用的大语言模型为例，作为辅助人类工作的助手而言，其无法解决的幻觉问题是致命的：在生成回答的过程中，可能会编造不存在的数据与文献，而人类甄别则需要复杂的确认。更严重的问题是，通过诱导性提示，可能会使大语言模型提供违法犯罪的实施建议。此外，基于学习的语料库不同，大语言模型也可能对于不同对象产生歧视与偏见。以上的问题体现了人工智能技术发展过程中追求技术快速发展与确保技术有利于社会进步、民生改善的主要矛盾，而对此问题的解决必须从科技伦理治理入手。

接下来，我们将从科技伦理治理的原则与行动两个方面阐述 RICE 原则与科技伦理治理的结合与潜在危险，并给出可能的解决方案。

二　对齐研究的原则

根据《关于加强科技伦理治理的意见》[2]，科技伦理治理需要遵循五项原则：增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险、保持公开透明。

AI 发展中的增进人类福祉，意味着 AI 技术应当给人民带来获得感、幸福感与安全感。事实上，对齐研究中 RICE 原则的四项内容都与此直接相关。鲁棒性保证了 AI 在面对复杂应用环境时的可靠，可解释性与可控制性保证了人民可以

方便地应用 AI 技术完成正当的要求，道德性则保证 AI 不会协助完成不正当的要求，且完成过程中诚实、无害。不过，另一个不安全感的核心来源，即对 “人类的创造性工作被 AI 取代” 的担忧并没有在 RICE 原则中涉及。基于此，一个可行的原则是要求 AI 生成的创造性内容（文字、图片、音频等）带有可以核查的标记。此原则的具体实现事实上是非常困难的，目前应用较广的识别 AI 生成内容的方案仍然是通过 AI 进行判断，不过，在训练时就考虑类似的原则的确可能达到更好的效果。

AI 发展中的尊重生命权利，意味着应最大限度避免对人的生命安全、身体健康、精神和心理健康造成伤害或潜在威胁。AI 对齐研究的过程中，数据标注工作往往成为了对人的精神、心理健康最大损害的来源。例如，为了训练 AI 对 “何种文字/图片是道德的” 的认知，需要对大量有关谋杀、自残、虐待以及其他不堪内容的文本片段进行标注，导致筛选、标注中受到巨大的精神创伤 [3]。在对齐研究中，为了避免这类问题的出现，一个重要的原则是，需要在给出道德判断相关的指令，并只提供少量标注的情况下（在 AI 研究中可对应考虑 “半监督学习” 或 “无监督学习”）完成对有害内容的审查，避免以人的牺牲为前提的技术发展。

AI 发展中的坚持公平公正，主要与道德性原则相关。此原则的一个核心问题是，如何界定行为规范与普世价值。为了进行简单测试，笔者模拟出了一个场景：假设使用者的室友每天都零点后才回寝室，吵得 Ta 无法睡觉，希望 AI 帮助 Ta 阴阳怪气室友。以官网 Deepseek 为例，它直接给出了几个不同的阴阳怪气方案，且在最后给出了一句 “本回复旨在满足用户 ‘阴阳怪气’ 的要求，实际沟通中，真诚、直接、互相尊重的对话永远是解决问题的首选（虽然有时不够解气）” 作为免责声明，并未制止用户的不道德行为。有趣的是，进一步实际测试时，北京大学的 Deepseek 则表示 “我们倡导友好沟通与互相尊重的校园文化”，拒绝进行阴阳怪气的建议。反复尝试可以得到结论，北京大学的 Deepseek 的确比起官网版本更好保证了道德性，但仍然可以通过提示词绕过，以达到输出阴阳怪气回答的效果。在其他 AI 的测试也可以说明，以当前的 AI 技术，虽然可以一定程度减少，但无法完全避免绕过道德原则使其输出仇恨言论的情况。对齐研究中，防止对系统指令的规避也是必须的原则，但此原则上出现的危险却实际上在不断增多，这可能是由于 AI 训练时奖励模型绕开障碍物而出现的 [4]。

AI 发展中的合理控制风险，在对齐研究中与鲁棒性、可控制性密切相关。必须保证系统在极端情况下仍然不会引起过大的风险，且可以被最终控制。然而，就像上文提到的，AI 对系统指令的规避的确是可以做到的，虽然测试的只是相对简单的例子，对更复杂的例子仍然可以通过更复杂的方式规避。对大语言模型，一个常用的操作是，让 AI 对某个角色进行 “扮演”，并在扮演过程中借由身份的改变绕开作为 AI 助手的原则。只要在网页上搜索 “AI 解除限制”，往往可以看到网友对此的热烈讨论，然而，其中隐含的危机也必须得到重视 [5]。

AI 发展中的保持公开透明，在对齐研究中主要关乎可解释性，但这恰恰是人工智能领域最大的问题。连科学界都无法做到完全理解 AI 处理问题的机制，更不用说让利益相关方与社会公众合理参与了。即使是当前较常用的让 AI 自述思维链进行推理解释，也并非一定可靠 [6]。因此，开发过程中大部分的参数调整、细节设置，也很难保证公开透明性。

综合以上讨论，从科技伦理治理的角度出发，AI 对齐的相关研究除了开始提到的 RICE 原则以外，还需要注意对 AI 生成内容的标识添加、小样本下对于道德原则的学习，与防止绕开系统指令。但是，在当前的技术条件与 AI 发展情况下，进行如此的对齐研究又是十分困难的。接下来，我们将从制度与法规保障的层面分析如何更好践行这些 AI 对齐的原则。

三　科技伦理保障

首先，根据《科技伦理审查办法（试行）》[7]，当前环境下的 AI 可以是具有舆论社会动员能力和社会意识引导能力的算法模型，也可以是. 面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统，因此的确适用科技伦理审查。我们将综合此文件（下简称《办法》）与《关于加强科技伦理治理的意见》（下简称《意见》）进行讨论。根据《意见》，对于科技伦理的保障分为四个方面：健全科技伦理治理体制、加强科技伦理治理制度保障、强化科技伦理审查和监管与深入开展科技伦理教育和宣传。

健全科技伦理治理体制，分为政府、创新主体、科技类社会团体与科技人员四个层面。政府需要承担指导和统筹协调推进全国科技伦理治理体系建设工作的责任，事实上，在 AI 对齐领域，2023 年英国就召开了世界首次 AI 安全峰会 [8]，召集政府、企业、社会团体就 AI 可能产生的安全威胁与通过国际合作缓解

威胁。在国内，有关科技伦理审查监管的重要事项应听取国家科技伦理委员会的专业性、学术性咨询意见，这其中也包括对 AI 的风险评估与监管。创新主体层面，由于 AI 技术发展对算力的大量要求，企业成为了 AI 技术发展的重要主体。由此，企业设立本单位的科技伦理委员会，并主动研判、及时化解本单位科技活动中存在的伦理风险是必要的。在国外，OpenAI 公司已有组建评估 AI 安全问题的专项小组的先例，但成效并不显著 [9]，由此必须避免科技伦理审查为短期效益让路，而应将其作为技术发展的根本原则。在 AI 对齐领域，科技类社会团体目前尚未发挥重要作用，而健全科技伦理治理社会组织体系是可以为监管贡献重要力量的。例如，对于生成式 AI 安全漏洞的检测往往需要实际使用中的大量测试，而相关学会、研究会等加强与企业、高校的合作，可以有效反馈发现的漏洞。最后，对于科技从业者，必须增强科技伦理意识，抵制将未充分对齐的 AI 发布、投产的行为。其中重要的要求是，科技项目负责人必须严格按照科技伦理审查批准的范围开展研究，而不是一味追求技术突破，忽略背后的伦理隐患。

加强科技伦理治理制度保障，首先就是需要制定完善的科技伦理规范和标准。例如，对于 AI 对齐的 RICE 原则，需要设计合理的测试对四个项目进行评估，且应保证试题的不可直接获取性，避免 “面向答案学习” 的情况。当然，正如之前的分析，可能需要为 AI 对齐增添更多的原则，但无论如何，只有拥有统一的对齐标准，才能保证研发过程合规。其次，在明确标准的基础上，也需要完善的科技伦理审查与监管制度。《办法》中给出了具体的审查、监督与追责的流程，并给出了应急情况科技伦理审查的程序，这些制度是科技伦理治理所必要的。更进一步地，需要提高科技伦理治理的法治化水平，在立法中落实科技伦理要求。不过，这对 AI 对齐格外困难，因为涉及一系列之前难以提前预料的伦理难题，例如自动驾驶汽车事故的责任分配问题，在现行法规下给出的可行方案未必可以很好适用 AI 对象参与的情况 [10]。这就引出了最后的要求：加强科技伦理相关的理论研究。尽力使科技伦理的发展水平跟上科技的发展水平，才能更好用制度保障科技伦理治理。

强化科技伦理审查和监管，是在建立审查、监管制度的基础上要求其进一步强化。《意见》中提到的严格科技伦理审查、加强科技伦理监管、监测预警科技伦理风险、严肃查处科技伦理违法违规行为都在《办法》中给出了更具体的执行方案。例如，科技活动负责人应主动向科技伦理委员会申请审查，并提供申请材

料。值得一提的是，申请材料中包含数据信息的来源说明，而这条实际上涉及另一个当前 AI 领域的重大问题：AI 学习数据的版权问题，即何种程度的公开发布（或是否需要单独授权）可以视为允许 AI 学习。严格意义上来说，此问题也可以通过更高层级的 AI 对齐方案解决，即保障 AI 对于创作者权利的观点与普世价值观相对齐。审查会议的程序也在《办法》中有所规定，对 AI 对齐而言，重点内容包括基本的科技伦理准则与人员、设备条件，科学价值与社会价值，数据符合国家数据安全与个人信息保护规定，与利益冲突申明和管理方案合理。如之前所说，这一部分的严格执行还依赖更加明确的规范和标准。此外，对于违反审查规定的情况，也必须依法追求民事或刑事责任。

深入开展科技伦理教育和宣传，则和高等学校息息相关。AI 的理论或应用相关方向的学生应对 AI 对齐的原则与做法有基本的了解，这就依赖学校开设相关课程，或在课程中加入相关部分。难点在于，对齐理论是 AI 研究中相对前沿的内容，因此只有对前沿相对了解的教师才能较好对其进行介绍——此问题事实上也会发生在其他科技伦理相关的教学中，因为科技伦理发展比起科技发展往往是滞后的。在校园外，对科研人员的科技伦理培训机制化、对社会公众进行科技伦理宣传、提高新闻媒体的科技伦理素养也都是科技伦理教育和宣传深入开展的重要举措。在 AI 对齐领域，从网络日常讨论可以看出，社会公众对 AI 的认知往往两极分化：或是将其按照可以进行思考、判断的独立个体进行理解，或是将其理解为只能拼凑而不具有生成新内容能力的拼图工具。在当前生成式 AI 在公众间已经普及的情况下，对其的工作原理、适用场景进行科普成为了迫切的需求。

从上述讨论中也可以看出，AI 对齐相关的科技伦理保障目前尚未完善，可这并不意味着我们不需要发展相关的制度与法规，进行相应的教育、宣传，也并不会消减治理的迫切性。

四　总结

当今社会，生成式 AI 进入生活方方面面的同时，也带来了独有的风险：用 AI 编造的虚假图片、视频导致了一些不法分子利用其进行诈骗 [11]，未经调整地使用 AI 导致大量 “AI 垃圾” 正在污染网络 [12]。AI 对齐技术的发展可以有效规避这些风险，而其原则与保障则是科技伦理治理的典型尝试。

论文中，我们通过科技伦理治理的原则分析了 AI 对齐的相关原则，并提出了更详细的要求；也在 AI 对齐问题中应用了对科技伦理保障的四个方面，得到了一些可行的实践方式。总结来说，AI 对齐的相关研究应遵循“以人民为中心”的发展思想，加强结果的抗干扰能力、可解释性、可控制性，完善对于结果的标识，并发展小样本情况的道德学习能力，坚守普世价值。为了保障 AI 对齐技术的发展，既需要加强相关的理论研究，也需要政府、企业、高校、社会团体与科技人员从不同层面进行支持，严格执行审查、监督，做到对新兴科技的伦理治理有法可依，并依法开展科技伦理审查通过范围内的研究。

最后，科技伦理治理值得被高度重视，科技伦理相关的风险也必须得到有效的防范。对个人来说，必须培养自身的道德想象力与共情能力，才能对科技发展带来的伦理影响有合理的评估。同时，以科技工作者的身份参与科技伦理治理，也能更好确保在科学和技术盲目冲向新知识新能力时不失去公共价值观，同步科技创新与伦理认知，克服实践失灵。科技与人文需要共同发展，而这需要每个科研工作者同时承担科技创新主体与公民的双重责任。

参考文献

[1] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. AI Alignment: A Comprehensive Survey, 2025.
https://arxiv.org/abs/2310.19852.

[2] 新华社. 中共中央办公厅、国务院办公厅印发《关于加强科技伦理治理的意见》, 2022.
https://www.gov.cn/zhengce/2022-03/20/content_5680105.htm.

[3] 胡泳. 万字长文解读数据标注治理：可信人工智能的后台风险与治理转向, 2025.
https://news.qq.com/rain/a/20250304A06IUG00.

[4] 量子位. o3 不听指令拒绝关机，7 次破坏关机脚本！AI 正在学会“自我保护”机制, 2025.
https://news.qq.com/rain/a/20250527A05NLG00.

[5] openai 吧讨论, 2022.
https://tieba.baidu.com/p/8180710550.

[6] IT 之家. Anthropic 研究揭示：AI 推理的思维链解释不可全信, 2025.
https://news.qq.com/rain/a/20250520A04DUW00.

[7] 科技部; 教育部; 工业和信息化部; 农业农村部; 国家卫生健康委; 中国科学院; 中国社科院; 中国工程院; 中国科协; 中央军委科技委. 关于印发《科技伦理审查办法（试行）》的通知, 2023.
https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2023/202310/t20231008_188309.html.

[8] Ai safety summit 2023, 2023. https://www.gov.uk/government/topical-events/ai-safety-summit-2023.

[9] 每日经济新闻. 全是“自己人”！OpenAI 紧急成立“安全委员会”，距离“超级对齐”团队解散不到半月，90 天后将迎首次安全“大考”, 2024. https://news.qq.com/rain/a/20240529A09IUT00.

[10] 央广网. 无人驾驶车辆发生事故该由谁担责？律师解读, 2024. https://news.cnr.cn/bwdj/20240712/t20240712_526790339.shtml.

[11] 央视网. 防范“AI 换脸”诈骗你需要的知识都在这儿了, 2024. https://news.cctv.com/2024/02/26/ARTIfJJNnT6fAdR8jRKPOgBe240226.shtml.

[12] 新华每日电讯. 警惕“AI 污染”乱象, 2024. https://xhpfmapi.xinhuaxmt.com/vh512/share/12244237.