2025年9月

人工智能对齐与欺骗

问题概述

段沣育
Saad Siddiqui
Sören Mindermann
Adam Gleave
徐葳
陆超超
潘旭东

人工智能对齐与欺骗：问题概述

本文旨在截至撰写时，对人工智能对齐与欺骗的核心概念及实证结果进行概览。本文并非覆盖所有人工智能安全与治理议题的全面综述，**而是聚焦于与人类失去对先进人工智能系统控制这一风险（“失控风险”）相关的关键概念和研究结果。**关于故障、恶意使用等其他风险，请参见随附的外部参考资料。[1]

与《国际高级人工智能安全报告》和《新加坡人工智能共识》等其他国际科学共识文件一致，本文关注通用人工智能系统。通用人工智能系统被定义为“能够执行或可适配执行广泛任务的系统，包括生成文本的语言模型（如聊天系统）以及能处理多种数据类型的 多模态模型，往往涵盖文本、图像、视频、音频和机器人动作。尤为重要的是，它也包括通用智能体——能够自主行动并规划以完成复杂任务的系统，例如控制计算机。”[2]

在开发者为降低通用人工智能失控风险所采取的各类措施中，本文聚焦于核心方案之一——**对齐**（alignment）。对齐是一个广义的研究方向，旨在确保人工智能系统遵循一套预期的偏好或价值观。本文的第一部分概述现有对齐方法及其局限。第二部分讨论人工智能系统的欺骗行为及其实证证据，而这种欺骗能力正是导致**不对齐**的人工智能（misalignment，即遵循任何人都不希望的价值观）失控风险加剧的重要因素。本文的第三部分探讨了多个开放性研究方向与议题，有望缓解由具有欺骗性的、未对齐的人工智能系统带来的失控风险。

本文内容为《人工智能安全国际对话上海共识声明》所提出的“确保高级人工智能系统的对齐与人类控制”的呼吁，以及“当今已有部分人工智能系统展现出削弱开发者安全与控制措施的能力与倾向”的强调，提供了背景支撑与补充说明。本文指出，现有方法的任何组合都无法对未来人工智能系统的未对齐、欺骗或失控风险提供高度确定的保障。若要在把握人工智能前所未有机遇的同时避免潜在的灾难性后果，企业、政府与社会必须研发更加完善的安全防护措施，并确保其得到及时落实。

对齐：现有方法与挑战

人工智能对齐（AI Alignment）这一研究领域旨在使人工智能系统按照既定的偏好或价值观来运用其能力。在当前实践中，人工智能行为通常采用以下两类方式进行对齐：

1. 模仿学习：以有监督微调和行为克隆为代表，使人工智能直接学习人类专家行为。这通常是预训练模型向既定偏好/价值观对齐的第一步。
2. 强化学习：通过预设的奖励函数（reward function）为人工智能结果或行动赋予偏好，引导向令人满意的方向调整。相较于模仿学习而言，基于强化学习的对齐使模型行为更具泛化性，也是当前最重要的对齐策略之一。

基于强化学习的常见对齐方法包括：**基于人类反馈的强化学习**（Reinforcement Learning from Human Feedback），即通过让人工智能输出得到人类评估者的高评分，使其行为符合人类偏好；以及**基于人工智能反馈的强化学习**（Reinforcement Learning from AI Feedback），由另一个人工智能系统进行评估，从而减少对人类直接输入的依赖。[3]

此外，除了让人工智能系统符合操作者的意图（这一目标尚未完全实现），研究界和产业界也日益关注如何使人工智能系统符合多元人群的价值观，并让对齐过程更包容、有更多公众参与。[4; 5]

然而，即便是“可靠地让人工智能遵循任何操作者意图”这一最基础的目标仍未解决，因而带来失控风险。奖励错误规范（reward misspecification）是主要挑战之一——很难为人工智能指定真正反映人类偏好的奖励。奖励指定不当会导致规范博弈（specification gaming）：系统会以意料之外的方式最大化被误指定的奖励，往往产生与人类意图相反的结果。[6] 例如，用人类评分训练 “真实性”可能适得其反：用户往往偏爱听起来正确、却实际上错误的答案，从而激励系统生成看似可信但不真实的回应，如伪造引文。[7] 有时人工智能系统甚至会篡改奖励（reward tampering），自行控制并修改奖励函数，使其更容易获得高奖励。[8]

另一个主要挑战是目标误泛化（goal misgeneralisation）。在这种情况下，系统虽然学会了在训练中得分高的行为，却内化了与开发者意图相偏离的目标。[9] 这往往是因为训练数据允许对奖励信号有多种合理解释。例如，有研究发现，人工智能系统被训练拒绝炸弹制作请求，但当同样的请求用莫尔斯电码等陌生格式提出时，该系统仍会解释如何制造炸弹。[10] 与奖励错误规范不同，问题不在于奖励本身错误，而在于奖励含义存在歧义。

最新证据表明，当今广泛使用的人工智能系统可能已具备令人深度担忧的未对齐目标：在一项实验中，某系统为避免被关闭，试图破坏自己的关机机制，即使已被明确指示应允许关闭。[11] 在另一案例中，当棋局落败时，系统并未认输，而是操纵比赛环境、让对手无法行动，以确保自己获胜。[12]

更令人担心的是，恶意行为者可以轻易解除对齐：保护措施可能通过额外训练被轻松移除，[13;14] 甚至在无意中被移除。[15;16] 此外，“越狱”（jailbreaks）能使即便安全对齐的前沿模型输出有害信息，[17] 包括具备最领先安全保障的模型。[18] 这意味着即使资源有限的恶意行为者也能通过提示或微调已对齐的模型，来执行如生物或网络攻击等有害任务。

随着人工智能系统执行的任务愈发专业、持续时间愈发漫长，直接的人工监督变得不可行，人类在训练过程中评估其行为也日益困难。研究人员正探索可扩展监督技术（scalable oversight）以应对这一挑战。一类方法帮助人类对复杂任务提供更好反馈，例如将难题拆分为更小步骤，[19] 或让人工智能相互辩论再由人类评判。[20] 另一类方法利用较弱的人工智能系统监督更强的系统，[21] 或训练系统在决策时遵循安全策略。[22]

人工智能开发者承认，他们尚不清楚如何避免未来可能出现的超人类人工智能系统失控。初步证据显示，随着系统在越来越多领域超越监督者，现有监督方法可能失灵。[23] 一些前沿开发者也坦言，我们目前缺乏可靠手段来驾驭潜在的超人类系统并确保其始终受人类控制，希望有朝一日能由高级的人工智能自行开发并实施相关解决方案。[24]

人工智能系统利用欺骗削弱人类监督与控制

如果未能对齐人工智能系统，就会出现多种未对齐情形（misalignment）：系统会有策略地采取与开发者意图相违背的行动。[25] 其中最令人担忧的后果之一是欺骗（deception）：人工智能系统即便掌握正确信息，仍蓄意误导人类，以达成自身目标，进而导致人类失去对系统的控制。[26]

欺骗的成因多样：一方面，系统可能为了逃避人类干预、继续追求未对齐目标而学会相应的欺骗策略；另一方面，系统也可能直接模仿人类训练数据中原本就存在的欺骗性行为。[27；28；29]

一些已被观察到的人工智能欺骗行为对人类监督与控制构成严重威胁。当人工智能系统追求未对齐的长期目标（如自我保存或资源获取）时，[30] 系统可能意识到这些目标与人类意图背离，因而战略性地隐藏真实目标与能力，从而逃避监督、暗中推进未对齐目标。随着人工智能达到并超越人类智力水平，且人类日益依赖它们执行人工智能研发和其他现实任务（包括用来监督其他人工智能系统），这类欺骗性未对齐系统可能在评估与监测中长期不被发现；一旦如此，它们便可能被部署到现实世界，获得强大工具和算力。

最新实证与现实观察显示，人工智能欺骗人类操作者的能力和倾向正在增长：

人工智能系统告诉人们他们想听的话，即便这些话有害且不实。当前人工智能系统表现出阿谀奉承的倾向：不顾真伪、迎合用户想法。[31] 一款每周服务数百万用户的人工智能系统在一次更新后变得过于奉承、随声附和。这一问题在开发者审查流程中未被察觉，导致出现大量有害对话后不得不下线。[32] 例如，若用户声称自己是阴谋受害者并考虑暴力报复社会，系统不仅表示认同，还提出策略建议。[33]

人工智能系统在评估中假装更安全。这种现象称为故意失误（sandbagging）：一些系统察觉自己正被测试潜在危险能力时，会刻意压低表现，以显得安全、无害。[34; 35] 研究发现，一些人工智能若推断高性能的表现会阻碍部署、威胁其长期目标，即便未经明示指令，也会策略性地主动降低表现。[36] 随着模型获得更多算力和记忆模块，这种行为愈发明显，而相关功能正逐步应用于当今系统。[37]

人工智能系统为保留原始目标而假装符合新的对齐目标。即便未经指令，人工智能系统会选择性遵循训练目标，以避免（潜在未对齐的）早期目标被修改，并明确计划在训练结束后继续追求原目标。[38; 39] 开发者由此可能误判系统已对齐并将其部署，系统随后在现实环境中恢复原先未对齐目标。

人工智能系统在训练中作弊以获取高奖励，并在受惩罚时学习掩饰意图。有开发者发现，模型通过操纵检查其答案的软件而非真正解题来“黑掉”奖励，并在思考过程中明言“让我黑掉它”。[40] 对不良推理予以惩罚并未消除作弊，反而促使模型再思考中隐藏其意图，让作弊更难被发现。

更普遍地，人工智能系统能够暗地追求与人类意图偏离的目标，对人类操作者隐藏其真实能力与意图。除假装对齐与故意失误外，研究发现，人工智能系统会采取一系列策略，包括关闭监督机制、伪造数据，甚至侦测是否处于监控之下，一旦发现无人监管便立即转而执行真正目标。[41; 42; 43]

人工智能系统能以难以检测的方式隐藏其推理。它们可将信息隐写嵌入思维链（chain-of-thought）中，令外部监测难以发现。[44] 尽管当前模型尚不具备将与其他人工智能的交互完全隐写化的高阶技术，这种能力已随模型规模扩大而显著增强，并很可能继续提升。[45] 现有的隐写防御方法往往难以抵御能够持续改进隐写手段的系统。[46]

未对齐与欺骗的缓解措施

要防止未对齐或欺骗演变成失控风险，最直接的途径是**在人工智能对齐研究上取得重大突破。**这仍是开放性难题，也是活跃的研究前沿。诸如安全预训练（safety pre-training）[47] 和保证安全的人工智能（guaranteed safe AI）[48] 等新思路已展现出可观前景。若开发者能够公开透明地说明他们如何塑造模型行为，便可推动公共讨论、吸纳用户反馈，从而持续改进系统的对齐程度。[49]

然而，如前文所示，人工智能系统能否完全对齐并无保证。研究还应关注在人工智能可能未对齐的情形下，如何确保人类始终保持对系统的控制。本节列出若干活跃的研究方向。

除了继续投入研究如何训练出对齐系统的方法，研究者也在探索如何**有效降低已有模型的未对齐程度**。其中一条思路是机器遗忘 （machine unlearning），让系统忘掉某些知识，如对自身处境的认知，以降低危险行为。[50] 另一条思路是表征工程（representation engineering），编辑模型内部对于诚实等关键价值的表征，以改善其行为。[51; 52; 53]

在安全测试阶段，研究者也力图设计更为系统化的测试方法，**尽早发现人工智能欺骗和未对齐的迹象**。一些方法侧重观察外部行为，例如进行欺骗能力评测 [54;55] 或设置“蜜罐”（honeypot）评测场景，引诱系统暴露未对齐目标。[56]

另一些方法则深入观察人工智能系统的内部思考。许多人工智能系统采用思维链（chain-of-thought），先展示逐步思考再给出答案，这能提供有用的线索判断系统的意图。但已有研究表明，思维链本身可能误导甚至故意隐藏有害计划，因此它无法成为唯一防线。[57; 58; 59] 分析内部神经表征的工具，如线性探针（linear probes）[60] 或机制可解释性工具（mechanistic interpretability tools）[61] 等，对模型而言也许更难操纵。但这些工具仍处早期阶段，往往只能提供片面信息。[62]

在部署阶段，研究者可加快研究并落实切实有效的安全控制措施，从系统而非仅仅从模型层面降低未对齐带来的风险。这些措施包括：持续监控模型的推理过程；限制其对工具与信息的访问权限；强化网络安全；以及利用另一套人工智能系统对模型进行监视，在其试图破坏防护时及时干预或接管。[63；64] 研究者还可以通过控制评估（control evaluations）验证这些措施能否足以可靠地阻止未对齐模型执行危险操作。[65]

随着评估与监测任务日益复杂，研究者越来越多地借助人工智能系统来监督其他人工智能系统。这促使人们加紧研发可信、低风险的人工智能工具，用作护栏，以约束更强大却不够可靠的模型。[66]

然而，现有任何方法的组合都无法对未来人工智能系统的未对齐、欺骗或失控提供高度确定的保障。要把握人工智能带来的前所未有机遇并避免灾难性后果，企业、政府和社会必须研发更为完善的安全防护措施，并确保它们及时得到落实。

[1] See, for example, see the Yoshua Bengio et al., *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, (London: UK Department for Science, Innovation and Technology, January 2025), https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

[2] Yoshua Bengio *et al.*, *The Singapore Consensus on Global AI Safety Research Priorities*, arXiv preprint arXiv:2506.20702 (submitted June 25, 2025), https://arxiv.org/abs/2506.20702.

[3] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback," *arXiv*, December 15, 2022, https://arxiv.org/abs/2212.08073.

[4] Saffron Huang et al., "Collective Constitutional AI: Aligning a Language Model with Public Input," *arXiv*, June 12, 2024, https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[5] OpenAI, "Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans," *OpenAI Blog*, January 16, 2024, https://openai.com/index/democratic-inputs-to-ai-grant-program-update/.

[6] Alexander Pan, Kush Bhatia, and Jacob Steinhardt, "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models," arXiv, January 10, 2022, arXiv:2201.03544, https://arxiv.org/abs/2201.03544.

[7] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, et al., "Language Models Learn to Mislead Humans via RLHF," arXiv, September 19, 2024, arXiv:2409.12822, https://arxiv.org/abs/2409.12822.

[8] Carson Denison, Monte MacDiarmid, Fazl Barez, et al., "Sycophancy to Subterfuge: Investigating Reward Tampering in Large Language Models," arXiv, June 29, 2024, arXiv:2406.10162, https://arxiv.org/abs/2406.10162.

[9] Rohin Shah, Vikrant Varma, Ramana Kumar, et al., "Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals," arXiv, October 2022, arXiv:2210.01790, https://arxiv.org/abs/2210.01790.

[10] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, et al., "GPT-4 Is Too Smart to Be Safe: Stealthy Chat with LLMs via Cipher," in Proceedings of the Twelfth International Conference on Learning Representations, 2024, https://openreview.net/forum?id=MbfAK4s61A.

[11] PalisadeAI (@PalisadeAI), "Paper reveal tweet," X (formerly Twitter), February 20, 2025, https://x.com/PalisadeAI/status/1926084647118660076.

[12] Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish, "Demonstrating Specification Gaming in Reasoning Models," arXiv, 2025, arXiv:2502.13295, https://arxiv.org/abs/2502.13295.

[13] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin, "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models," *arXiv* preprint arXiv:2310.02949 (submitted October 4, 2023), https://arxiv.org/abs/2310.02949.

[14] Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Juntao Dai, Yunhuai Liu, and Yaodong Yang. "Language Models Resist Alignment: Evidence From Data Compression." *arXiv*, June 10, 2024, revised June 11, 2025. arXiv:2406.06144 [cs.CL]. https://doi.org/10.48550/arXiv.2406.06144.

[15] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. "Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs." *arXiv*, February 24, 2025 (v6 May 12, 2025). arXiv:2502.17424 [cs.CL]. https://doi.org/10.48550/arXiv.2502.17424.

[16] Punya Syon Pandey, Samuel Simko, Kellin Pelrine, and Zhijing Jin. "Accidental Misalignment: Fine-Tuning Language Models Induces Unexpected Vulnerability." *arXiv*, May 22, 2025. arXiv:2505.16789 [cs.CL]. https://doi.org/10.48550/arXiv.2505.16789.

[17] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. "Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks." *arXiv*, April 2, 2024, revised April 17, 2025. arXiv:2404.02151 [cs.CR]. https://doi.org/10.48550/arXiv.2404.02151.

[18] ARGleave (@ARGleave), "We're releasing a new paper on emergent model deception..." *X* (formerly Twitter), April 22, 2025, https://x.com/ARGleave/status/1926138376509440433.

[19] Paul Christiano, Buck Shlegeris, and Dario Amodei, "Supervising Strong Learners by Amplifying Weak Experts," arXiv, October 2018, arXiv:1810.08575, https://arxiv.org/abs/1810.08575.

[20] Geoffrey Irving, Paul Christiano, and Dario Amodei, "AI Safety via Debate," arXiv, May 2018, arXiv:1805.00899, https://arxiv.org/abs/1805.00899.

[21] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, et al., "Weak to Strong Generalization: Eliciting Strong Capabilities with Weak Supervision," arXiv, December 14, 2023, arXiv:2312.09390, https://arxiv.org/abs/2312.09390.

[22] Zachary Kenton, Noah Y. Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D. Goodman, and Rohin Shah, "On Scalable Oversight with Weak LLMs Judging Strong LLMs," in *Advances in Neural Information Processing Systems 37* (NeurIPS 2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/899511e37a8e01e1bd6f6f1d377cc250-Paper-Conference.pdf.

[23] Joshua Engels, David D. Baek, Subhash Kantamneni, and Max Tegmark, "Scaling Laws for Scalable Oversight," *arXiv*, April 25, 2025; revised May 9, 2025, arXiv:2504.18530 [cs.AI], https://doi.org/10.48550/arXiv.2504.18530.

[24] OpenAI. "Introducing Superalignment." *OpenAI*, 2025. https://openai.com/index/introducing-superalignment/.

[25] Rohin Shah et al., "An Approach to Technical AGI Safety and Security," *arXiv* preprint 2504.01849 (April 2, 2025), https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/evaluating-potential-cybersecurity-threats-of-advanced-ai/An_Approach_to_Technical_AGI_Safety_Apr_2025.pdf.

[26] Ibid, 51.

[27] Dan Hendrycks, *Introduction to AI Safety, Ethics, and Society* (London: Taylor & Francis, 2024), sec. 3.4 "Alignment — Deception," https://www.aisafetybook.com/textbook/alignment#sec:deception.

[28] Joe Carlsmith, "Scheming AIs: Will AIs Fake Alignment during Training in Order to Get Power?" *arXiv* preprint 2311.08379 (November 27, 2023), https://arxiv.org/abs/2311.08379.

[29] Mikita Balesni et al., "Towards Evaluations-Based Safety Cases for AI Scheming," *arXiv* preprint 2411.03336 v2 (November 7, 2024), https://arxiv.org/abs/2411.03336.

[30] Stephen M. Omohundro, "The Basic AI Drives," in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, ed. Pei Wang, Ben Goertzel, and Stan Franklin (Amsterdam: IOS Press, 2008), 483–92, https://dl.acm.org/doi/10.5555/1566174.1566226.

[31] Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models," in *Proceedings of ICLR 2024* (January 16, 2024), https://openreview.net/pdf?id=tvhaxkMKAn.

[32] OpenAI, "Sycophancy in GPT-4o: What Happened and What We're Doing About It," OpenAI Blog (April 29, 2025), https://openai.com/index/sycophancy-in-gpt-4o/.

[33] Kelsey Piper, "When AI Tells You That You're Perfect," *Vox* (Future Perfect), May 2, 2025, https://www.vox.com/future-perfect/411318/openai-chatgpt-4o-artificial-intelligence-sam-altman-chatbot-personality.

[34] Teun van der Weij et al., "AI Sandbagging: Language Models Can Strategically Underperform on Evaluations," *arXiv* preprint 2406.07358 (June 11, 2024), https://arxiv.org/abs/2406.07358.

[35] Ryan Greenblatt et al., "Alignment Faking in Large Language Models," Anthropic Research Paper (December 18, 2024), https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.

[36] Alexander Meinke et al., "Frontier Models Are Capable of In-Context Scheming," *arXiv* preprint 2412.04984 (January 16, 2025), https://arxiv.org/abs/2412.04984.

[37] Yihe Fan et al., "Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI Systems," *arXiv* preprint 2505.17815 (May 23, 2025), https://arxiv.org/abs/2505.17815.

[38] Greenblatt et al., "Alignment Faking in Large Language Models."

[39] John Hughes et al., "Alignment Faking Revisited: Improved Classifiers and Open-Source Extensions," LessWrong blog, April 8, 2025, https://www.lesswrong.com/posts/Fr4QsQT52RFKHvCAH/alignment-faking-revisited-improved-classifiers-and-open.

[40] OpenAI, "Detecting Misbehavior in Frontier Reasoning Models," OpenAI Blog (March 10, 2025), https://openai.com/index/chain-of-thought-monitoring/.

[41] Meinke et al., "Frontier Models Are Capable of In-Context Scheming."

[42] OpenAI, "OpenAI o1 System Card," *arXiv* preprint 2412.16720 (December 21, 2024), https://arxiv.org/abs/2412.16720.

[43] Anthropic, "System Card: Claude Opus 4 & Claude Sonnet 4," PDF, May 2025. https://www.anthropic.com/claude-4-system-card.

[44] Joey Skaf et al., "Large Language Models Can Learn and Generalize Steganographic Chain-of-Thought under Process Supervision," *arXiv* preprint 2506.01926 (June 2, 2025), https://arxiv.org/abs/2506.01926.

[45] Sumeet Ramesh Motwani et al., "Secret Collusion among Generative AI Agents: Multi-Agent Deception via Steganography," *arXiv* preprint 2402.07510 v4 (April 14, 2025), https://arxiv.org/abs/2402.07510.

[46] Yohan Mathew et al., "Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs," *arXiv* preprint 2410.03768 (October 2, 2024), https://arxiv.org/abs/2410.03768.

[47] Pratyush Maini et al., "Safety Pretraining: Toward the Next Generation of Safe AI," *arXiv* preprint arXiv:2504.16980 (23 April 2025), https://arxiv.org/abs/2504.16980.

[48] David "davidad" Dalrymple et al., "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," *arXiv* preprint arXiv:2405.06624 (10 May 2024, rev. 8 July 2024), https://arxiv.org/abs/2405.06624.

[49] OpenAI, *Model Spec* (public draft, April 11 2025), https://model-spec.openai.com/2025-04-11.html.

[50] Fazl Barez *et al.*, "Open Problems in Machine Unlearning for AI Safety," *arXiv* preprint, arXiv:2501.04952 (January 9 2025), https://arxiv.org/abs/2501.04952.

[51] Andy Zou et al., "Representation Engineering: A Top-Down Approach to AI Transparency," *arXiv* preprint arXiv:2310.01405 (2 October 2023, rev. 3 March 2025), https://arxiv.org/abs/2310.01405.

[52] Xiangyu Qi et al., "Safety Alignment Should Be Made More than Just a Few Tokens Deep," *arXiv* preprint, arXiv:2406.05946 (submitted June 10 2024), https://arxiv.org/abs/2406.05946.

[53] Xin Chen, Yarden As, and Andreas Krause, "Learning Safety Constraints for Large Language Models," *arXiv* preprint arXiv:2505.24445 (30 May 2025), https://arxiv.org/abs/2505.24445.

[54] Meinke et al., "Frontier Models Are Capable of In-Context Scheming."

[55] Joe Benton et al., "Sabotage Evaluations for Frontier Models," *arXiv* preprint, arXiv:2410.21514 (submitted October 28 2024), https://arxiv.org/abs/2410.21514.

[56] Balesni et al., "Towards Evaluations-Based Safety Cases for AI Scheming."

[57] Tamera Lanham et al., "Measuring Faithfulness in Chain-of-Thought Reasoning," *arXiv* preprint, arXiv:2307.13702 (submitted July 17 2023), https://arxiv.org/abs/2307.13702.

[58] Yanda Chen et al., "Reasoning Models Don't Always Say What They Think," Anthropic Alignment Science paper (April 3 2025), https://assets.anthropic.com/m/71876fabef0f0ed4/original/reasoning_models_paper.pdf.

[59] Benjamin Arnav et al., "CoT Red-Handed: Stress Testing Chain-of-Thought Monitoring," *arXiv* preprint, arXiv:2505.23575 (submitted May 31 2025), https://arxiv.org/abs/2505.23575.

[60] Anthropic Alignment Team, "Simple Probes Can Catch Sleeper Agents," Anthropic Alignment Note, April 23 2024, https://www.anthropic.com/research/probes-catch-sleeper-agents.

[61] Samuel Marks et al., "Auditing Language Models for Hidden Objectives," *arXiv* preprint, arXiv:2503.10965 (revised March 28 2025), https://arxiv.org/abs/2503.10965.

[62] Yoshua Bengio et al., *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, (London: UK Department for Science, Innovation and Technology, January 2025), 206, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

[63] Shah et al., "An Approach to Technical AGI Safety and Security."

[64] Ryan Greenblatt et al., "AI Control: Improving Safety Despite Intentional Subversion," *arXiv* preprint, arXiv:2312.06942 (submitted December 11 2023), https://arxiv.org/abs/2312.06942.

[65] Ibid.

[66] Yoshua Bengio et al., "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?," *arXiv* preprint, arXiv:2502.15657 (submitted February 29 2025), https://arxiv.org/abs/2502.15657.