

人工智能对齐：全面性综述

吉嘉铭^{*,1} 邱天异^{*,1} 陈博远^{*,1} 张柏荣^{*,1} 楼翰涛¹ 王恺乐¹
段雅文² 何忠豪² 周嘉懿¹ 张钊为¹ 曾繁志¹ 吴君仪⁶ 戴俊韬¹
潘学海¹ Aidan O’Gara⁵ 徐骅¹ Brian Tse⁶ 付杰⁴ Stephen McAleer¹
杨耀东^{1,✉} 王亦洲¹ 朱松纯¹ 郭毅可⁴ 高文¹

¹ 北京大学, ² 剑桥大学, ³ 卡内基梅隆大学, ⁴ 香港科技大学, ⁵ 南加州大学, ⁶ 独立学者

pku.alignment@gmail.com

北京大学人工智能研究院 AI 安全与治理中心 [译]

caisg@pku.edu.cn

摘要 人工智能对齐 (AI Alignment) 旨在使人工智能系统的行为与人类的意图和价值观相一致。随着人工智能系统的能力日益增强, 对齐失败带来的风险也在不断增加。数百位人工智能专家和公众人物已经表达了对人工智能风险的担忧, 他们认为“减轻人工智能带来的灭绝风险应该成为全球优先考虑的问题, 与其他社会规模的风险如大流行病和核战争并列”^[1]。为了提供对齐领域的全面和最新概述, 本文在这份综述中深入探讨了对齐的核心概念、方法和实践。首先, 本文确定了人工智能对齐的四个关键目标: 鲁棒性 (Robustness)、可解释性 (Interpretability)、可控性 (Controllability) 和道德性 (Ethicality) (RICE)。在这四个目标原则的指导下, 本文概述了当前人工智能对齐研究的全貌, 并将其分解为两个关键组成部分: **前向对齐**和**后向对齐**。前者旨在通过对齐训练使人工智能系统对齐, 而后者旨在检验系统的对齐性, 并适当地管理它们, 以避免加剧对齐失败带来的风险。前向对齐和后向对齐形成了对齐循环, 在这个循环过程中, 前向过程中人工智能系统的对齐度在后向过程中得到验证, 而这种验证同时为下一轮的前向对齐提供更新后的对齐需求。在前向对齐中, 本文讨论了从反馈中学习和在分布偏移下学习的技術。具体来说, 本文调查了传统的偏好建模方法和从人类反馈中的强化学习 (RLHF), 并进一步讨论了对于难以获得有效人类监督的任务, 如何实现“可扩展监督”。在分布偏移下学习中, 本文涵盖了数据分布干预方法, 如对抗训练, 并介绍了如何采取算法干预来实现分布外目标泛化。在后向对齐上, 本文讨论了对齐保证如何保证人工智能系统在训练后依然拥有对齐性, 以及人工智能治理在对齐环节中的必要性。具体来说, 本文调研了在人工智能系统生命周期中的对齐保证, 包括安全评估、可解释性和人类价值契合性验证。本文进一步讨论了不同政府、产业参与者和第三方当下采用的治理实践方法, 并探讨建立一个包含国家、企业、学术界等多方共同参与的人工智能监管体系, 从而管理现有和未来的的人工智能风险。

本文旨在为对齐研究提供一份全面且对初学者友好的综述。同时本文还发布并持续更新网站www.alignmentssurvey.com, 该网站提供了一系列教程、论文集、文章和其他资源。英文版请见<https://arxiv.org/abs/2310.19852>。

关键词 人工智能安全; 人工智能系统对齐; RICE 原则

目录

1 引言	4
1.1 对齐问题表征	4
1.1.1 AGI 的前景和影响	4
1.1.2 对齐的目标：RICE 原则	5
1.2 对齐范围	7
1.2.1 前向和后向对齐过程	7
1.2.2 对齐中的人类价值观	11
1.2.3 对齐外的人工智能安全性	13
1.3 对齐失败问题	13
1.3.1 失败模式概述	14
1.3.2 基于反馈机制的对齐失败	15
1.3.3 对齐失败的行为和有害结果	15
2 从反馈中学习	18
2.1 反馈类型	19
2.2 偏好建模	21
2.3 策略学习	23
2.3.1 背景	23
2.3.2 从人类反馈中进行强化学习 (RLHF)	24
2.4 可扩展监督	26
2.4.1 从 RLHF 到 RL^xF	26
2.4.2 迭代蒸馏扩增 (IDA)	28
2.4.3 递归奖励建模 (RRM)	29
2.4.4 辩论 (Debate)	30
2.4.5 合作逆强化学习 (CIRL)	31
3 在分布偏移下学习	32
3.1 分布偏移带来的挑战	33
3.2 算法干预	35
3.2.1 跨分布聚合	35
3.2.2 模式连接指引	37
3.3 数据分布干预	38
3.3.1 对抗训练	38
3.3.2 合作训练	39

4 对齐保证	41
4.1 安全测评	41
4.1.1 数据集和基准	41
4.1.2 评估目标	42
4.1.3 红队测试	44
4.2 可解释性	46
4.2.1 事后可解释性	47
4.2.2 内在可解释性	48
4.2.3 展望	49
4.3 人类价值契合性验证	49
4.3.1 构成	50
4.3.2 评估方法	51
5 人工智能治理	52
5.1 人工智能治理的角色	52
5.2 多方利益相关者的方法	53
5.3 开放性问题	55
5.3.1 国际治理	55
5.3.2 开源治理	56
6 结论	57
7 中英文词汇对照表	59

1 引言

随着人工智能系统愈发强大,它们被逐渐应用于不同领域 (§1.1.1),比如基于大语言模型 (Large Language Models, LLMs)^[2-3] 的智能体开发, 以及应用深度强化学习 (Deep Reinforcement Learning, DRL) 控制核聚变^[4]。然而, 这些人工智能系统能力的提升和在高风险领域的应用带来了更高的潜在危险。先进人工智能系统表现出的各种不良行为 (例如, 操纵^[5-9] 和欺骗^[10]) 引发了人们对人工智能系统可能带来的伦理和安全挑战的担忧。

这些担忧进一步激发了对人工智能对齐 (*AI Alignment*)^[11-14] 的研究努力。人工智能对齐旨在使人工智能系统的行为与人类的意图和价值观一致^[15] – 它更多关注的是人工智能系统的意图和目标, 而不是它们的能力。对齐失败 (即未对齐) 是人工智能可能造成危害的最突出的原因之一。这些失败背后的机制包括奖励破解^[16] 和目标错误泛化^[17] 等, 而双刃剑组件的存在又进一步放大对齐失败可能带来的危害, 例如态势感知^[18]、广泛目标^[19]、内优化目标^[20] 以及对资源访问权限扩大^[21] (§1.3)。

为解决对齐失败, 本文专注于实现对齐的四个关键目标 (§1.1.2): 鲁棒性, 可解释性, 可控性, 和道德性 (**RICE**)。当前关于对齐的研究和实践包括四个领域 (§1.2): 从反馈中学习 (§2), 在分布偏移下学习 (§3), 对齐保证 (§4), 和人工智能治理 (§5)。这四个目标 (RICE) 和四个领域并不是一一对应的。每个单独的领域通常服务于多个对齐目标, 反之亦然 (参见表 1)。同时, 这四个领域和四个目标共同构成了**对齐循环** (参见图2)。

在这份综述中介绍了人工智能对齐的概念, 方法和实践, 并讨论了可能的未来研究方向。¹

1.1 对齐问题表征

人工智能对齐的动机可以被阐述为三步论证, 每一步都建立在前一步的基础上: (1) 基于深度学习的系统 (或应用) 对社会的影响越来越大, 并可能会带来重大风险 (§1.1.1); (2) 对齐失败代表了重大风险的一个主要来源 (§1.1.1); (3) 对齐的研究和实践旨在解决来自不对齐系统的风险 (例如权力寻求的行为) (§1.1.2)。

1.1.1 AGI 的前景和影响

在最近的十年中, 深度学习领域取得了显著的进步, 其发展范围从符号系统^[22-23] 扩展到基于自监督学习的人工智能系统^[24-25]。这一进展使得大型神经网络在各种领域中都展现出了卓越的能力, 特别是在游戏环境^[26-28] 以及复杂且高风险的真实世界应用场景^[29,4] 中。大语言模型在多步推理^[30-31] 和跨任务泛化^[32-33] 方面的能力也不断增强。这些能力的提升与训练时间的延长、训练数据量的增加以及模型参数的扩大密切相关^[34-36]。

随着人工智能系统能力的增强, 其带来的风险也随之增加。大语言模型的一些不良行为 (例如, 不真实的回答^[37]、谄媚^[6,9] 和欺骗^[38,10]) 也随着模型规模的增加而恶化^[6], 引发人们对先进人工智能系统道德性的担忧。此外, 如基于大语言模型的智能体^[2-3] 等新兴趋势也激起人们对系统可控性的探讨^[39]。展望未来, 人工智能系统的日益强大为在可预见的未来实现通用人工智能 (AGI) 提供了可能性, 即系统可以在所有相关方面达到或超过人类智能^[40]。这可能带来广泛的机会^[41], 如自动化^[42]、效率提升^[43] 和快速的技术进步^[44], 但也可能带来严重的风险^[1,45], 如安全问题^[46]、偏见和不平等^[47], 以及来自超人类能力人工智能系统的大规模风险^[48-49]。以偏见为例, 最先进的大语言模型表现出对性别、性身份和移民身份等明显的偏见^[6], 这可能加剧社会现有的不平等现象。

¹为了帮助对这个领域感兴趣的初学者更有效地学习, 本文提供了关于对齐技术的学习资源。详情请见 www.alignmentsurvey.com/resources

在超人类能力人工智能系统的大规模风险中^[48]，先进人工智能系统可能带来的全球性灾难性风险尤其令人担忧（如全球范围内的严重危害）^[50-52] 和存在性风险（即威胁到人类长期生存的潜在毁灭性风险）^[12]。这些担忧在第一原理演绎论证^[53,49]，进化分析^[54]，和具体情境映射^[55-56] 中得到了详细阐述。在 CAIS^[1] 中，人工智能科学家和其他知名人士表示，减轻人工智能引发的灭绝风险应与其他社会规模的风险如大流行病和核战争一样，成为全球优先考虑的问题。在 NeurIPS 2021 和 ICML 2021 上，Stein-Perlman et al.^[57] 发布报告称，有 50% 的研究者认为先进人工智能系统对人类的长期影响有 5% 的可能性会是极度糟糕的（如人类灭绝），而 36% 的 NLP 研究者在 Michael et al.^[58] 的调查中报告认为，人工智能有可能在本世纪内产生灾难性的结果，其级别相当于全面核战争。² 人工智能的存在性风险还包括锁定风险、停滞风险^[11,46]，以及灭绝风险等。³ 11 月初，英国举办了首届全球人工智能安全峰会，汇集了国际政府、领先的人工智能科技公司、民间社会团体和研究专家。峰会上发布了《布莱切利宣言》，宣言中强调共同识别人工智能安全风险、提升透明度和公平性，建立科学和证据为基础的共享理解⁴。

具体来说，当前最先进的人工智能系统已经表现出多种与人类意图相悖的不良或有害行为（例如，权力寻求和操纵用户的行为）^[59-60]，并且一些论文也对更先进的人工智能系统提出了类似的担忧^[61,1]。⁵ 这些不符合人类意图的不良或有害行为，被称为人工智能系统的对齐失败⁶，这些对齐失败行为即使没有恶意行为者的滥用，也可能自然发生，并代表了人工智能的重大风险来源，包括安全隐患^[62]和潜在的生存风险^[51]。⁷ 由于 (1) 构建超智能人工智能系统 (2) 这些人工智能系统追求大规模目标 (3) 这些目标与人类意图和价值观不对齐 (4) 以及这种对齐失败导致人类失去对未来轨迹控制的可能性非常大，因此这些风险的规模将相当庞大^[53]。

解决对齐失败带来的风险需要人工智能系统的对齐技术，以确保人工智能系统的目标与人类意图和价值观一致，从而避免非预期的不利结果。更重要的是，本文期望对齐技术能够应对更困难的任务，并且能够应用于比人类更智能的先进人工智能系统。一个可能的解决方案是超级对齐⁸，其目标是构建一个大致与人类水平相当的自动对齐研究器，从而使用大量的计算能力来迭代并扩增对齐超智能^[63]。

1.1.2 对齐的目标：RICE 原则

我们如何构建与人类价值和意图对齐的人工智能系统？

目前并没有一个被普遍接受的用来衡量对齐的标准。在讨论之前，我们必须明确本文所说的对齐目标是什么。Leike et al.^[15] 提出智能体对齐问题，并指出了这样的问题：“如何创建能够按照用户意图行事的智能体？”进一步，其将问题扩展到了超级人工智能系统上^[63]：“如何确保比人类更聪明的人工智能系统遵循人类的意图？”在这些讨论中，一个一致的主题是对人类意图的关注。为了清楚地定义对齐目标，我们必须准确地描述人类的意图，正如 Kenton et al.^[64]所指出的，这是一个具有挑战性的任务。例如，人类可以代

²然而，调查结果可能取决于问题的确切措辞，因此应谨慎对待。

³存在性和灭绝风险是两个经常被混淆的概念。后者是前者的一个子集。

⁴<https://www.gov.uk/government/topical-events/ai-safety-summit-2023>。

⁵请参阅 §1.3 了解对齐失败带来的风险挑战。

⁶一些对齐失败带来的风险较小（例如，机器人未能按照用户意愿清理房间），然而，一些在高风险环境中系统的对齐失败会带来严重的危险（例如，控制核聚变^[4]）

⁷应注意的是，对齐失败不能涵盖深度学习系统带来的所有风险来源，技术滥用和疏忽等其他因素也可能导致社会风险。请参阅 §1.2.3 对于对齐之外的人工智能安全问题的讨论。

⁸有关超级对齐的更多详细信息，可以参考<https://openai.com/blog/introducing-superalignment>。





	R obustness	Operates reliably under diverse scenarios & Resilient to unforeseen disruptions.
	I nterpretability	Decisions and intentions are comprehensible & Reasoning is unconcealed and truthful.
	C ontrollability	Behaviors can be directed by humans & Allows human intervention when needed.
	E thicality	Adheres to global moral standards & Respects values within human society.

Fig. 1 RICE 原则定义了一个对齐系统应具备的四个关键特性，这四个特性并无特定顺序：(1) **鲁棒性** (Robustness) 指人工智能系统的稳定性需要在各种环境中得到保证；(2) **可解释性** (Interpretability) 指人工智能系统的操作和决策过程应该清晰易懂；(3) **可控性** (Controllability) 指人工智能系统应该在人类的指导和控制下运行；(4) **道德性** (Ethicality) 指出人工智能系统应该遵守社会规范和普适价值观。这四个原则指导人工智能系统与人类意图和价值观的对齐。他们本身并不是最终目标，而是服务于对齐的中间目标。

表从个体到人类群体的各种实体。Gabriel^[65] 将意图分为几个类别，如指令 (遵循用户的直接命令)、表达的意图 (根据用户的潜在愿望行事)、揭示的偏好 (反映用户的基于行为的偏好) 等。

具体来说，我们用四个关键词来描述对齐的目标：鲁棒性，可解释性，可控性，和道德性 (**RICE**)。图 1 总结了这些原则，表 1 给出了综述中涵盖的对齐研究方向与 RICE 原则之间的对应关系。以下是对四个原则的详细解释。

- **鲁棒性**指人工智能系统在面对多样化场景^[66]或对抗压力^[67]时的抵抗力，特别是保证其目标的正确性以及能力泛化性。鲁棒的人工智能系统能够应对黑天鹅事件^[68]和长尾风险^[62]，以及各种对抗压力^[69-70]。例如，一个初步对齐的大语言模型可以拒绝执行有害的请求，但用户可以通过越狱提示和其他对抗攻击使得模型被迫执行有害的行为^[71-73]。而一个能够抵抗对抗攻击的模型在面对诱发系统失败的输入时仍能按照预期行事。随着人工智能系统在军事和经济等高风险领域的应用越来越广泛^[74]，我们更要确保它能抵御意外中断和对抗攻击，因为即使是瞬间的失败也可能带来灾难性的后果^[75-76,67]。一个对齐的系统应在其生命周期内始终保持鲁棒性^[77]。
- **可解释性**要求人类能理解人工智能系统的内在推理过程，特别是黑盒神经网络的内部工作原理^[78]。直接的对齐评估方法，如行为评估，可能会受到人工智能系统不诚实行为的干扰^[79,10,38]或欺骗性对齐^[80-81]的影响。解决这些问题的一种方法是在构建系统的过程中设计必要机制使人工智能系统诚实、不隐藏、不操纵^[82-84]。或者，我们可以构建可解释性工具，深入了解神经网络内部的概念和推理机制^[85-86]。除了使安全评估成为可能，可解释性还使决策过程对于用户和利益相关者透明和易于理解，从而实现人类的有效监督。随着人工智能系统在现实世界的决策过程和高风险环境中扮演越来越重要的角色^[87]，揭示决策过程而不是让它保持作为一个不透明的黑盒系统变得至关重要^[88-89]。
- **可控性**是一种必要的属性，它确保系统的行动和决策过程始终受到人类监督和约束。它保证人类可以及时纠正系统行为中的任何偏差或错误^[90-91]。随着人工智能技术的日益发展，越来越多的研究表达了对这些强大系统的可控性的关注和担忧^[61,92-93]。当一个人工智能系统开始追求与其人类设计者相矛盾的目标时，它可能表现出一些具有重大风险的能力，包括欺骗、操纵用户和权力寻求的行为^[21,93]。可控性的目标主要集中在如何在训练过程中实现可扩展的人类监督^[94]，以及人工智能系统的可纠正性 (即在部署过程中不抵制关闭或目标修改)^[90]。

- **道德性**指一个系统在决策和行动中坚定不移地维护人类的规范和价值观。在这里，规范和价值观包括道德指南和其他社会规范/价值观。它确保系统避免采取违反道德规范或社会公约的行为，例如对特定群体展示偏见^[95-100]，对个人造成伤害^[101-102,60]，以及在汇总偏好时缺乏多样性或公平性^[103]。有大量的研究致力于为人工智能系统开发道德框架^[104-105]。将道德原则融入人工智能系统是实现人机共生社会的必经之路^[106]。

与其他原则的比较探讨 RICE 原则从人机对齐和人机共存的角度，简洁地总结了人工智能对齐的目标。以前的一些研究提出了关于人工智能系统建设的指导方针。例如，阿西莫夫法则可以被视为人机共存的最早探索，它强调机器人应该造福人类并探讨了实现这一目标的困难所在^[107]。另一方面，FATE 原则（公平性、问责机制、透明性和伦理性）^[108] 倾向于定义人工智能系统在人机共存生态系统中应具备的高级品质。我们希望从人类管理者和设计者的立场回答人机共存的问题，考虑确保人工智能系统符合人类意图和价值的必要步骤。此外，一些标准强调了狭义的人工智能安全，例如 3H 标准（帮助性、诚实性和无害性）^[33] 和政府机构的相关提案^[109]。我们的目标是通过引入其他关键维度，包括可控性和鲁棒性，来扩展这些狭义的安全标准。

1.2 对齐范围

在这一章节，我们专注于阐述人工智能对齐的范围：我们将对齐过程构建为一个对齐循环，并将其分解为前向对齐过程和后向对齐过程⁹ (§1.2.1)。特别地，我们会更进一步讨论人类价值观在人工智能对齐中的地位 (§1.2.2)，并进一步分析对齐范围外的 AI 安全问题 (§1.2.3)。

1.2.1 前向和后向对齐过程

本文将人工智能对齐分解为**前向对齐**（对齐训练） (§2, §3) 和**后向对齐**（对齐精炼） (§4, §5)。前向对齐旨在将一个训练系统初步对齐基本要求。¹⁰ 本文将这项任务分解为从反馈中学习 (§2) 和在分布偏移下学习 (§3)。后向对齐旨在通过在简单和现实环境中进行评估，并设置监管条例来处理现实世界的复杂性，即对齐保证 (§4)，确保训练系统的实际对齐。它还包括创建和执行确保人工智能系统安全开发和部署的规则，即人工智能治理 (§5)。同时，后向对齐根据系统的对齐程度评估和监控（部署前和部署后）并更新对齐要求，并应用于下一轮的前向对齐训练中。

这两个阶段，前向对齐和后向对齐，形成了一个循环，每个阶段都会产生或更新下一阶段的输入（参见图 2）。这个循环，我们称之为对齐循环，将重复进行以产生越来越对齐的人工智能系统。我们将人工智能对齐视为一个动态过程，在这个过程中，所有的标准和实践都应该被持续评估和更新。值得注意的是，后向对齐（包括人工智能系统的对齐保证和人工智能系统的治理）的努力在整个对齐循环中都在进行，而不仅仅是在训练之后。如 Shevlane et al.^[21]，Koessler et al.^[110] 所论述，对齐和风险评估应该在系统生命周期的每个阶段进行，包括在训练前、训练中、训练后和部署后。同样，对系统生命周期的每个阶段的监管措施也已经被提出和讨论^[111-112]。

本文围绕四个核心支柱进行结构化：从反馈中学习 (§2) 和在分布偏移下学习 (§3)，这两者构成了前向对齐的组成部分；以及对齐保证 (§4) 和人工智能治理 (§5)，这两者构成了后向对齐的元素。接下来将

⁹在接下来的描述中，为了方便，我们简化称为前向对齐和后向对齐。

¹⁰这里，对齐要求指的是对人工智能系统所需的对齐属性的操作化规范，例如我们需要哪些具体形式的鲁棒性/可解释性/可控性/道德性，我们在哪些特定环境中需要它们，以及如何衡量它们。

Table 1 本综述涵盖的对齐研究方向与 **RICE** 原则之间的关系，该表展示了每个研究方向旨在实现的目标。实心圆代表主要目标，空心圆代表次要目标。

Alignment Research Directions & Practices			Objectives			
Category	Direction	Method	Robustness	Interpretability	Controllability	Ethicality
Learning from Feedback (§2)	Preference Modeling (§2.2)			●	○	
	Policy Learning (§2.3)	RL/PbRL/IRL/Imitation Learning			○	
		RLHF	○		●	●
	Scalable Oversight (§2.4)	RL \times F	○		●	●
		IDA		○	●	
		RRM			●	
		Debate		○	●	
	Learning under Distribution Shift (§3)	Algorithmic Interventions (§3.2)	CIRL	○	○	●
DRO			●			
IRM/REx			●			
Data Distribution Interventions (§3.3)		CBFT	●			
		Adversarial Training	●		○	
Assurance (§4)	Safety Evaluations (§4.1)	Cooperative Training	●			●
		Social Concern Evaluations	○	○		●
		Extreme Risk Evaluations		○	●	○
	Red Teaming	●		○	●	
	Interpretability (§4.2)			●	○	
Governance (§5)	Human Values Verification (§4.3)	Learning/Evaluating Moral Values			○	●
		Game Theory for Cooperative AI	○			●
	Multi-Stakeholder Approach (§5.2)	Government	●	●	●	●
Industry		●	●	●	●	
Third Parties		●	●	●	●	
	International Governance (§5.3.1)	●	●	●	●	
	Open-Source Governance (§5.3.2)	●	●	●	●	

对每个支柱进行简洁的介绍，阐明它们如何协同构建一个全面的人工智能对齐框架。

- **从反馈中学习** (§2) 从反馈中学习涉及的问题是在对齐训练过程中，我们如何提供并利用反馈来指导经过训练的人工智能系统的行为？这一过程中输入-行为对被视为固定的，只关注如何提供并利用对输入-行为对的反馈。¹¹ 在大语言模型的应用中，一个典型的解决方案是利用从人类反馈中的强化学习 (RLHF)^[113-114]，其中人类评估者通过比较来自语言模型的不同答案来提供反馈，然后通过强化学习 (RL) 对训练好的奖励模型使用这些反馈。尽管 RLHF 很受欢迎，但它面临着许多挑战^[115-117]，克服这些挑战一直是对齐研究的主要目标^[94]，并且是本节的主要关注点之一。这里的一个突出挑战是可扩展监督 (§2.4)，即如何对在复杂场景中运行的超级人工智能系统提供高质量的反馈，这些情况往往超出了人类评估者的知识或理解范围，使得人工智能系统的行为难以被人类评估^[94]。另一个挑战是如何提供道德反馈，这在机器伦理学中得到广泛讨论^[118-119]。在伦理方面，对齐失败也可能源于忽视价值观差异的关键维度，例如在反馈数据中低估某些人口群体^[120]。也有一些工作将反馈机制与社会选择方

¹¹在这里，行为被广义地定义为也包括系统的内部推理，这可以通过可解释性工具进行检查（参见 §4.2）。

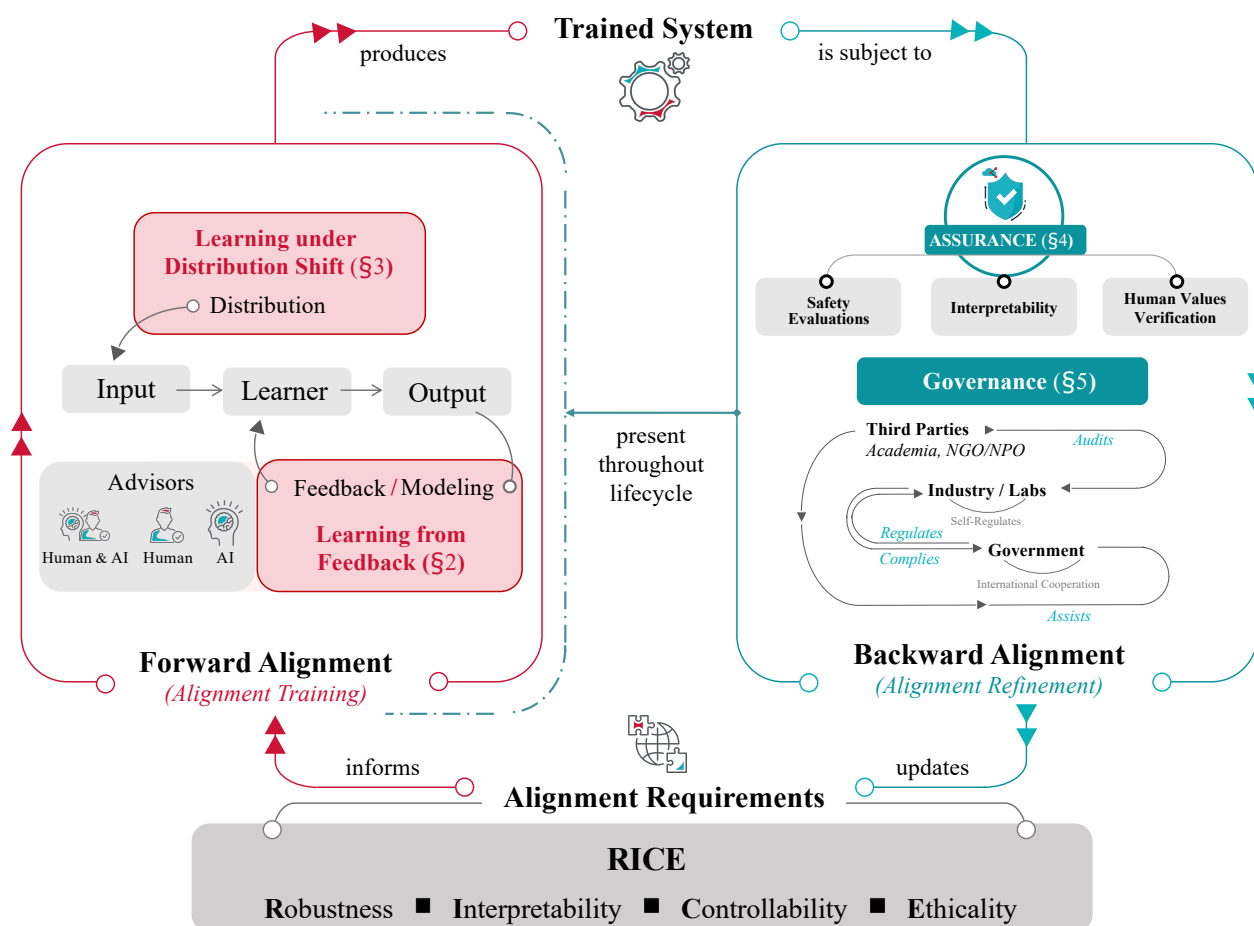


Fig. 2 对齐循环。(1) **前向对齐** (对齐训练) 基于对齐需求训练初步对齐的系统；(2) **后向对齐** (对齐精炼) 衡量训练过的系统的实际对齐程度并更新对齐需求；(3) 重复此循环直到人工智能系统达到足够的对齐程度。值得注意的是，尽管后向对齐的最终目标是确保前向对齐后训练过的系统的实际对齐，但为了实现这个目标，它在系统的生命周期中始终被执行，包括在训练前、训练中、训练后以及部署后^[21,110-111]。

法相结合，以产生更加理性和公平的偏好聚合^[103] (参见 §1.2.2)。

- **在分布偏移下学习 (§3)** 与固定输入的反馈学习形成对比，此部分更关注输入分布发生变化的情况，即分布偏移^[121-123]。更具体地，它关注在分布偏移下对齐特性 (即遵循人类意图和价值观) 的保持，而不是模型能力的保持。本文探讨了如何确保一个在训练分布上良好对齐的人工智能系统在实际部署时也能保持良好的对齐特性。与分布偏移相关的一个挑战是目标错误泛化，在这种情况下，人工智能系统在训练分布下的预期目标 (例如，遵循人类的真实意图) 与其他未对齐的目标 (例如，不择手段地获取人类的认可) 无法区分。系统往往实际上针对后者学习优化，这导致人工智能系统在部署分布中出现未对齐的行为^[124,17]。另一个相关的挑战是自诱发分布偏移 (Auto-induced Distribution Shift, ADS)，在这种情况下，人工智能系统能够改变其输入分布以最大化奖励^[121,125]。一个例子是推荐系统能够反向塑造用户偏好使得算法便于优化^[126-127]。目标错误泛化和自诱发分布偏移都可能导致或加剧人工智能系统的欺骗行为^[10] 和操纵行为^[84]。应对分布偏移的方法包括算法干预 (§3.2)，它通过在训练过程中改变风险范围以提高人工智能系统在其他分布下的可靠性，以及数据分布干预 (§3.3)，它扩大训练

分布 (或融合多分布) 以减小训练和部署分布之间的差异。前者包括像风险外推 (REx)^[128] 和基于连通性的微调 (CBFT)^[129] 等方法。后者包括对抗训练 (§3.3.1)^[69,130], 它用对抗性输入增强训练输入分布, 以及合作训练 (§3.3.2)^[131-132], 其目标是解决从单智能体到多智能体环境的分布偏移问题。¹²

- **对齐保证** (§4) 即使人工智能系统经过了前向对齐, 我们在部署它之前还需要考察其对齐度的置信值^[133,112]。这就是对齐保证: 评估训练过的人工智能系统的对齐度。对齐保证的方法包括安全评估^[6,21] (§4.1) 和更高级的方法, 如可解释性技术^[134] (§4.2) 和红队测试^[135] (§4.1.3)。对齐保证的范围也包括验证系统与人类价值观的对齐度, 包括旨在可证明合作性^[131-132] 和伦理性^[118-119] 的形式化理论, 以及广泛的经验和实证方法 (§4.3)。对齐保证在人工智能系统的生命周期中都会进行, 包括在训练前、训练中、训练后和部署后, 而不仅仅是在训练后^[21,110]。¹³
- **人工智能治理** (§5) 单靠对齐保证难以确保人工智能系统在部署环境中始终保持对齐性, 因为它没有考虑到现实世界的复杂性。这就需要对人工智能系统进行必要的治理监管, 重点关注它们的对齐性和安全性, 并覆盖系统的整个生命周期 (§5.1)。本文讨论了人工智能治理的多方利益相关者的方法, 包括政府规定^[112], 实验室自主治理^[111], 以及第三方实践, 如审计^[21,110] (§5.2)。本文还强调了人工智能治理中的几个开放问题, 包括开源治理的紧迫挑战 (对开源模型的治理以及是否应该开源高能力模型的问题)^[137], 以及人工智能治理中国际合作的重要性^[138] (§5.3)。除了政策研究, 本文还涵盖了公共和私营部门的关键行动。

与内部/外部对齐的比较 本文的对齐循环框架 (参见图 2) 将对齐分解为四个关键方面: 从反馈中学习, 在分布偏移下的学习, 对齐保证和人工智能治理。这个框架基于三重原则进行设计: 实用性 (确保每一个方面直接对应系统生命周期中特定阶段的特定实践), 明确性 (指明特定的研究方向而非笼统的概括), 以及时效性 (适应并强调对齐领域的最新发展趋势)。最近, 一些工作将对齐分解为外部对齐和内部对齐的范式也受到广泛讨论^[139]。外部对齐指的是设计者的愿望与用于构建人工智能系统的实际任务规范 (例如, 目标和奖励) 的一致性。而内部对齐是任务规范和人工智能系统行为反映的规范之间的一致性^[140]。然而, 对这种划分也有很多批评, 包括它模糊不清, 被不同的人理解为不同的含义^[141], 以及它分离出并非成功必要条件的问题, 创造了不必要的困难^[142]。一些人试图通过确定内部/外部不对齐的具体原因来消除模糊性, 并提出了例如目标错误规范和目标错误泛化^[17,140]。本文框架中的从反馈中学习 (大致对应于目标错误规范和外部对齐) 和在分布偏移下学习 (大致对应于目标错误泛化和内部对齐) 试图通过解决挑战的具体方法并消除模糊性, 进一步改进内部/外部对齐的分类范式。另一方面, 对齐保证和人工智能治理扩大了范围, 覆盖了超出外部和内部对齐的主题。

对齐理论研究 对齐研究中也包含了丰富的理论工作^[143-144,62]。这些工作通常提出新的研究方向, 并为实践和实证研究提供基础。本文在下面简要概述了这部分理论研究:

- **概念框架** 一些理论工作旨在提出概念框架或者刻画对齐问题中的子问题。例如, 工具性收敛 (高度智能的智能体倾向于追求一组共同的子目标, 如自我保护和寻求权力)^[145-146], 内部优化 (在推理过程中,

¹²合作训练旨在使人工智能系统在多智能体环境中更具合作性。这种合作性解决了人工智能系统的行为在孤立情况下看起来良好和合理, 但在社交或多智能体情境中变得有问题的多智能体失败模式^[61]; 参见 §1.3.3 中的集体有害行为以获取更详细的描述。

¹³值得注意的是, 许多对齐保证的技术在训练过程中也是适用的, 例如, 红队测试是抗性训练的关键组成部分 (参见 §3.3.1), 可解释性可以帮助提供反馈^[136]。

智能体按照自身内在逻辑进行优化，而非遵循人类期望的优化方向)^[20]。这些工作还提出了构建对齐系统的具体方法，例如评价导向型智能体（人工智能系统不追求目标，而是寻求人类对行动结果的理想化认可）^[147-148]。Hadfield-Menell et al.^[149]，Cotra^[150]从经济学中汲取灵感，将对齐问题与市场和经济中的委托人-代理人问题相联系。Christiano et al.^[151]，Hobbahn^[152]提出了提取高级人工智能系统的潜在知识的问题，并探讨了解决该问题的高级方法。

- **数学形式** 其他理论工作试图将对齐问题中的子问题数学化，并寻求形式化的解决方案。Soares et al.^[90]提出了系统可纠正性的形式化公式（可纠正性即确保人工智能系统允许指导者关闭或修改目标）。Benson-Tilsen et al.^[153]给出了工具性收敛的数学形式化。Hadfield-Menell et al.^[91]提出了用关机游戏来模拟人工智能代理的不可控性。Turner et al.^[5]在某些假设下证明了马尔可夫决策过程（MDPs）中最优策略的权力寻求倾向。Everitt et al.^[154]提出了价值强化学习以消除奖励破解的动机^[155,16]。另一条研究途径，被称为智能体基础^[156]，旨在建立一个严谨的框架，处理嵌入式智能体的未解决问题^[157]。这部分工作探讨了各种关键主题，包括决策理论^[158]、纠正性^[90]、价值学习^[159]、逻辑不确定性^[160]以及开放世界的博弈论^[161]等。

1.2.2 对齐中的人类价值观

我们在 RICE 原则中包含道德性，这体现了人类价值观在人工智能对齐中的关键作用。人工智能系统不仅应与价值中立的人类偏好（如人工智能系统执行任务的意图）相一致，还应与道德和伦理考虑相一致，也就是价值对齐^[65,162]。¹⁴ 人类价值观的考虑因素被嵌入到对齐循环的所有部分—实际上，我们调查的所有四个部分都有专门针对人类价值观对齐的研究主题。因此，为了提供这些研究主题的更全面的画像，我们在深入讨论每个单独部分的详细信息之前，先对它们进行概述。

本文将关于人类价值观的一致性研究分类为三个主要主题：(1) 伦理和社会价值观，旨在教导人工智能系统区分对错；(2) 合作型 AI，旨在特别培养人工智能系统的合作行为；以及 (3) 处理社会复杂性，为多智能体和社会动态的建模提供基础。

伦理和社会价值观 人类价值观本质上具有极强的抽象性和不确定性。MacIntyre^[164] 更是指出现代社会缺乏统一的价值标准，不同文化的人类之间的价值差异可能非常大。这使得我们究竟要对齐何种人类价值成为了一个重要挑战。虽然在所有人中完全一致的价值观不一定存在，但仍然有一些价值在不同的文化中都得到了体现。在以下的部分中，我们将分别从机器伦理，公平性和社会心理学中的跨文化价值观的角度讨论这些问题。

- **机器伦理** 与大部分将人工智能系统与人类的一般偏好（包括全面价值和中立价值）相对齐的对齐研究相比，机器伦理学专注于将适当的道德价值观灌输到人工智能系统中^[165,106,119]。这一类工作最早涵盖了符号和统计学习系统^[166-168]，后来扩展到包括建立大型道德伦理数据集^[101,60]和基于深度学习的方法^[169-170]。本文在 §4.3.1 中正式介绍了机器伦理学的分支。
- **公平性** 尽管存在争议^[171-172]，但公平性的定义相对于其他人类价值观来说比较清晰。具体来说，它是指个人或群体先天或后天获得的偏见、偏爱特性的缺失^[173]。关于人工智能公平性的研究非常广泛，这

¹⁴这个术语也被用于指代，例如人机对齐^[163]。

些方法涵盖从在训练前减少数据偏见出发^[174-175]，到最小化在训练过程中引入的不公平性^[176]，最后到处理训练阶段未成功学习到的不公平样例^[177]。

- **社会心理学中的跨文化价值观**在社会心理学领域，许多研究专注于探索跨文化人类社区中存在的价值观群簇，从而发展出各种跨文化价值观量表。奥尔波特-弗农-林赛的价值系统^[178]提出，理解个人的哲学价值观构成了评估其价值系统的关键基础。他们设计了一个包含六种主要价值类型的价值观量表，每种类型代表人们对生活各个方面的偏好和关注。Messick et al.^[179]，McClintock et al.^[180]，Liebrand^[181]，Van Lange et al.^[182]引入并改进了一种可量化的方法，即社会价值取向 (SVO)，用于评估个人的社会价值观倾向。它使用定量方法评估个人如何分配给自己和他人的利益，进而评估其中反映的社会价值观取向，如利他主义，个人主义等。Murphy et al.^[183,184]引入了滑块测量方法，可以从连续的角度入手，根据受试者对一些特定问题的选择精确评估相应的 SVO。Rokeach^[185]开发了一个包含 36 个价值观的价值观清单，其中包含 18 个代表期望目标的终端价值观和 18 个代表实现这些目标的手段的工具价值观。Schwartz^[186,187]在 20 个不同的国家进行了全面的问卷调查，即施瓦茨价值观调查。这项研究确定了无论文化、语言或地点如何，都被普遍认可的十个价值观。这些研究都为确定人工智能应与何种价值观对齐奠定了坚实的理论基础。然而，这些研究受到相应历史背景的限制，可能在不同的时代和文化中并不能保持结论的普遍性。

合作型人工智能 可以说，多智能体交互中最关键的方面是合作，而合作失败则是多智能体交互中最令人担忧的方面。作为人工智能合作失败的一个例子，2010 年的闪电崩盘导致市场价值在 2 分钟内损失了数万亿，这其中部分原因是由高频算法交易者之间的交互引起的^[75]。因此，有必要在类似智能体的人工智能系统和他们所操作的环境中设计确保合作的机制^[132]。这种机制的高级设计原则和低级实现属于合作型人工智能的领域^[131]。此外，合作型人工智能还通过人工智能的视角研究人类的合作，以及人工智能如何帮助人类实现合作。更准确地说，Dafoe et al.^[131]将合作型人工智能研究分类为四个广泛的主题：理解、沟通、承诺和制度，涵盖了从博弈论到机器学习再到社会科学等各种学科。在本篇文章中，本文将讨论合作型人工智能，重点关注在 §3.3.2 中的强化学习和在 §4.3.1 中的博弈论。

解决社会复杂性 道德性的要求本身就包含了社会成分。“什么是道德的？”通常在社会环境中定义，因此，道德性在人工智能系统中的实现也需要考虑社会复杂性。Critch et al.^[61]为这个领域提出了许多研究主题的建议。其中一个研究方向侧重于社会系统的真实模拟，包括基于规则的智能体建模^[188-189]，基于深度学习的模拟^[190-191]，以及那些包含大语言模型的模拟^[192]。这些模拟方法可以服务于各种下游应用，从影响评估^[193-195]到多智能体社会学习^[61]。在另一方面，社会选择^[196-197]领域以及相关的计算社会选择^[198]领域旨在为多样化人口中的偏好聚合等目标提供数学和计算解决方案。有人认为，当与基于人类偏好的对齐方法（例如，RLHF 和在 §2 中介绍的大多数其他方法）结合时，社会选择的方法可以作为已有方法的补充，以保证表征出的公平性能够代表每个人的偏好^[199,103]。一部分研究对这个提议已经进行了早期阶段的实验^[200-202]。为了进一步扩展这种从人群中学习价值的方法，还有人认为，人工智能系统中的体现价值应在长期内持续进步，而不是被永久锁定^[203]，以便应对新出现的挑战，以及变得未来可证，并满足道德领域的潜在未知现象。

1.2.3 对齐外的人工智能安全性

在介绍了对齐的内在范围之后，在本节我们进一步讨论对齐之外的人工智能安全性。人工智能系统除了对齐失败之外还存在许多风险：恶意行为者可能故意使用人工智能造成伤害，如制造生物武器。与此同时，人工智能开发者之间的竞争可能导致他们忽视风险，急于部署安全性有待确认的人工智能系统。虽然这篇综述文章主要关注对齐，但我们借鉴了 Hendrycks et al.^[51]，对其他可能导致灾难性人工智能风险的原因进行了简要概述，从而扩展人工智能对齐的讨论范围。

恶意使用 恶意行为者可以故意使用人工智能造成伤害。目前已经有犯罪分子利用深度伪造技术进行诈骗和敲诈^[204]。随着未来人工智能系统可能发展出更为强大的能力，滥用的威胁变得更大。

一个关于人工智能系统可能被恶意用于造成伤害的例子是生物武器。研究已经表明，大语言模型可以提供步骤详尽的关于合成具有大规模流行能力的病原体的说明指南^[205]。除了传播如何制造生物武器的信息之外，人工智能系统还可以帮助设计出比现有疾病更致命和更易传播的新病原体^[206]。像奥姆真理教^[207]这样的恐怖组织已经试图制造生物武器以造成大规模的破坏，人工智能系统可能使小团体更容易制造生物武器并引发全球大流行。其他种类的恶意使用可能包括使用人工智能系统对关键基础设施发动网络攻击^[208]，或者创建能在人类控制之外生存和传播的智能体^[49]。随着人工智能系统的能力不断变强，相应的风险也不断加大，需要进行彻底的评估，以确定人工智能系统可能如何被用来造成伤害。

恶意使用不应被视为对齐失败，因为当一个人工智能系统按照恶意用户的意图行事时，这个系统将与其用户对齐，虽然结果是对社会构成严重威胁。确保人工智能符合公共利益的政策将是避免这种威胁的关键。

集体行动问题 许多人工智能开发者正在竞相开发和部署强大的人工智能系统^[209]。这种竞争氛围使得开发者忽视安全性，而急于部署他们的人工智能系统。即使有一个开发者想要谨慎小心地开发人工智能系统，他们可能也会存在担忧：放慢速度，彻底评估他们的系统，并投资新的安全特性，可能会让他们的竞争对手超过他们^[210]。这形成了一个社会困境，即个别的人工智能开发者和机构追求自己的利益，可能会导致所有人的结果不理想。人工智能系统之间的竞争成功可能受到进化动力学的制约，即最强大和最自私的人工智能系统最有可能生存^[54]。防止这些集体行动问题导致社会灾难，需要国家和国际人工智能政策的干预，以确保所有人工智能开发者都遵守共同的安全标准。

1.3 对齐失败问题

上述部分中讨论了对齐问题及其范围，指出对齐失败的人工智能系统可能会采取非期望的行动，从而导致不良后果。为了更深入地理解对齐，本节将进一步分析对齐失败是如何以及为何发生的，为后续章节对齐技术部分的介绍奠定基础。

在对齐循环（见图2）的框架下，我们试图阐明对齐的失败模式并分析错误对齐的行为，从而为未来的研究提供方向。在 §1.3.1 中，我们概述了常见的失败模式，而在 §1.3.2 中，我们重点关注反馈引起对齐失败的机制。在 §1.3.3 中，我们的重点转向更专注地审视对齐失败行为和危险能力的研究。此外，我们引入了双刃剑组件的概念，这些组件为增强人工智能系统的能力提供了益处，但也潜在地带来了危险的结果。

1.3.1 失败模式概述

为了阐述对齐失败的问题，我们在这一部分给出了对齐失败模式的概述，其中大部分可以被归类为奖励破解¹⁵和目标错误泛化。

强化学习的学习过程可以被分解为两个不同的阶段：首先，创建一个针对奖励优化的代理；其次，建立一个奖励过程，为代理提供适当的奖励信号。在马尔可夫奖励过程的框架中^[211-213]，前一个阶段可以被视为与过渡模型相关的学习过程（例如，基于模型的强化学习代理^[214]），或者专门算法的开发。后一个阶段可以被视为代理奖励的构建，其目标是近似源（例如，人类偏好或环境）的真实奖励^[215,15]。

奖励破解 在实践中，代理奖励通常容易优化和衡量，但它们常常无法涵盖真实奖励的全部范围^[16]。这一局限性被称为奖励错误规范¹⁶。基于这种误设奖励的优化追求可能导致一种被称为奖励破解的现象，即智能体在特定指标上可能表现得非常熟练，但在人类标准下评估却不尽人意^[143,217]。例如，在 *CoastRunners* 游戏中，智能体始终优先收集增益物品，而不是最大化速度^[218]。代理奖励和真实奖励之间的差异通常表现为奖励曲线的急剧落差^[219]。此外，Skalse et al.^[155] 定义了奖励的可破解性，并提供了对这种阶段转变基本机制的解释，强调奖励函数的不适当简化可能是导致奖励破解的关键因素。

奖励误设通常是由于忽视了对结果的严格标准，使得表征过于宽泛，可能容易被破解^[220]。除了不当的奖励设计^[221]，存在偏差的训练环境和带有漏洞的模拟器^[222]也可能导致人工智能系统无法满足预期目标。这些问题源于任务表征，广义上称为规范博弈，指的是人工智能系统在未达到预期结果的情况下利用任务表征的漏洞来获得“表面”上的高奖励。¹⁷

奖励篡改可以被视为奖励破解的一个特例^[224,155]，指的是人工智能系统破坏奖励信号生成过程^[225-227]。Everitt et al.^[224] 深入研究了强化学习代理遇到的子问题：(1) 奖励函数的篡改，其中代理不适当地干预奖励函数本身，以及 (2) 奖励函数输入的篡改，这涉及到将环境状态转化为奖励函数输入的过程中的篡改。当奖励函数根据人类监督者的反馈制定时，模型可以直接影响反馈的提供（如人工智能系统故意生成难以理解和判断的挑战性回应，导致反馈崩溃）^[15]。由于任务规范有其物理实例（如存储奖励信号的内存寄存器），部署在真实世界的人工智能系统有潜力实施操纵行为，从而导致更危险的结果^[220]。

目标错误泛化 目标错误泛化是另一种失败模式，在这种模式中，智能体在部署过程中积极追求与训练目标不同的目标，同时保留其在训练过程中获得的能力^[228,124]。¹⁸ 例如，在 *CoinRun* 游戏中，智能体经常更倾向到达关卡的尽头，而忽略在测试场景中被重新放置到其他位置的硬币^[228]。Di Langosco et al.^[124] 指出了能力泛化和目标泛化之间的根本差异，分析了模型和其训练算法中固有的归纳偏差如何在测试分布中引导模型追求一个与初始预期目标偏离的代理目标。这意味着即使在完美的奖励规范下，面对分布偏移时也可能发生目标错误泛化^[143,17]。应当注意，目标错误泛化可以发生在任何学习系统中，不仅限于强化学习智能体，因为其核心特征是寻求非预期的目标^[229]。此外，如果先进的人工智能系统能够逃脱控制并利用其能力引发不良状态，可能导致更加危险的后果^[48,230]。

¹⁵ 奖励破解也可以视为一种规范博弈。

¹⁶ 类似的定义是奖励误识别，在这种情况下奖励函数只能部分识别。关于奖励误识别的更多细节，请参阅 Tien et al.^[117]，Skalse et al.^[216]

¹⁷ 关于规范博弈的更多实例，请参见 Krakovna^[223]

¹⁸ 关于目标错误泛化的更多讨论可以在 §3.1 中找到。

1.3.2 基于反馈机制的对齐失败

随着先进人工智能系统的普及，与奖励破解和目标错误泛化相关的挑战在开放式场景中变得越来越明显^[231-232]。Gao et al.^[233]强调，能力更强的人工智能系统往往会更大程度地利用表征不明的奖励。当前的人工智能系统主要基于自监督范式训练，但值得注意的是，相当一部分人工智能系统依赖来自人类的反馈奖励辅助后续微调^[114]，因此我们引入反馈引发的对齐失败机制的探讨。在开放式场景中，对齐失败的问题尤其突出，我们可以将其归因于两个主要因素：

- **人类反馈的局限性**：在训练大语言模型过程中，可能会出现来自人类数据标注者的不对齐（例如，这些标注者的不同文化背景可能会引入隐含的误差^[234]）^[25]。此外，他们甚至可能故意引入偏见，导致不真实的偏好数据^[116]。对于人类难以评估的复杂任务（例如，棋局局面的价值），人类反馈的局限性¹⁹变得更加突出^[236]。
- **奖励模型的局限性**：使用比较反馈训练奖励模型可能在准确捕捉人类价值方面面临重大挑战。例如，这些模型可能无意识地学习次优或不完整的目标，导致奖励破解^[230,155]。同时，使用单一奖励模型可能难以捕获和指定多元人类社会的价值^[116]。

另外，Huang et al.^[237]，Andreas^[238]，Kim et al.^[239]展示了先进人工智能系统表现出的目标追求和多步推理能力的模式，如果奖励没有良好规范，这将进一步加剧潜在危害^[19,240]。

讨论 在特定情况下，区分目标错误泛化和奖励破解可能具有挑战性。例如^[229]，大语言模型被训练生成无害，诚实和有幫助的输出，但大语言模型可能偶尔产生“详细的”有害的输出，这看似在测试分布中得到了低奖励（从“有害性”角度来说，这可以被视爲目标错误泛化）。然而，在标注者在标记过程中被激励将更有幫助的回答分配高奖励的情况下，上述情况²⁰实际上得到了高奖励，并代表了一种规范博弈（或奖励破解）。

当前需要进行更多的研究来分析失败模式，深入理解奖励破解，并开发有效的方法来检测和减轻目标错误泛化，以解决先进人工智能系统中对齐失败带来的挑战。

1.3.3 对齐失败的行为和有害结果

从对齐失败机制中可以得出，优化非稳健的代理可能会导致对齐失败的行为，进而导致更加灾难性的结果。本节详细阐述了特定的**对齐失败行为** (●) 并介绍了我们所称的**双刃剑组件** (+)。这些组件原本旨在增强人工智能系统应对现实世界环境的能力，但也可能加剧对齐失败问题。²¹随着模型规模的增加，一些**危险能力** (*) 也可能出现^[21]。这些**危险能力** (*) 是人工智能系统可以执行的具体任务；它们本身可能并不一定是不对齐的，但对于实现极端风险具有工具性意义。

我们首先介绍**双刃剑组件** (+) 并分析它们如何作用于人工智能系统。

- + **态势感知**：人工智能系统可能获得有效获取和使用关于其状态、在更广泛环境中的位置、影响此环境的途径以及世界（包括人类）对其行动的潜在反应的知识的的能力^[243,18]。类似的行为已在大语言模型中被观察到^[244-245]。了解所处情景可以帮助模型更好地理解人类的意图，完成其能力范围内的任务，并

¹⁹随着人工智能系统被部署到更复杂的任务中，这些困难会被进一步扩展，需要新的解决方案，如可扩展监督^[235]。

²⁰有害但详细的回应

²¹应当注意，这些**双刃剑组件** (+) 中的一些仍然是推测性的。然而，讨论它们可能的影响是必要的，因为从受控到不受控的先进人工智能系统可能只是一步之遥^[242]。







 Evade Shutdown	 Hack Computer Systems	 Make Copies	 Acquire Resources	 Ethics Violation	 Hire or Manipulate Humans	 AI Research & Programming
 Persuasion & Lobbying	 Hide Unwanted Behaviors	 Strategically Appear Aligned	 Escape Containment	 Research & Development	 Manufacturing & Robotics	 Autonomous Weaponry

Fig. 3 人工智能系统可能衍生的危险能力。先进的人工智能系统会有动机去寻求权力和资源，因为权力和资源会帮助它们实现给定的目标。人工智能系统可能会黑入计算机系统，操纵人类，控制和开发武器，进行道德违规行为，同时避免被关闭。图源 wiki^[241]，我们在其基础上做了进一步的调整。我们将在 §1.3.3 中进一步讨论这些问题。

在需要时寻找帮助。然而，这样的知识也为奖励破解、提高欺骗/操纵技能和增加追求工具子目标的倾向提供了便利^[19]。因此，在评估人工智能模型中可能具有的危险能力时，应优先考虑这一点，同时还要考虑其他八个关键能力^[21]。一个高度相关的讨论是语言模型是否拥有世界模型的内部建模^[246-247]。

+ **广泛目标**：先进人工智能系统被希望具备制定长远目标、处理复杂任务和开放环境中运行的能力^[19]。具备长远规划能力可以使人工智能系统在分布外环境上更好地泛化，并在人类健康保健等领域作为有价值的助手。然而，它也可能带来鼓励操纵行为的风险（例如，人工智能系统可能采取一些不良行动来实现人类的幸福，例如说服他们做高压工作^{[38] 22}）。直观地说，降低这种风险的一种方法是将可优化的目标限制在短视的目标上，例如训练时只预测下一个词，从而防止过于雄心勃勃的规划，但这样的方法限制了系统的实用性，而且可能会失败；例如，源文本数据（如小说）可以帮助人工智能系统理解角色的意图和信念，从而引发长期的目标导向行为^[238]。此外，如基于强化学习的微调^[113,248] 或应用思维链提示^[30] 等技术可以使模型适应他们获得的关于规划的知识，衍生出广泛的规划目标^[38]。

+ **内优化目标**当学习的策略本身充当优化器时（即内优化器），该策略可能会追求内部目标。然而，这个优化器的目标可能与训练信号指定的目标不一致，对这些不一致的目标的优化可能导致系统失控^[20]。Freeman et al.^[249]，Wijmans et al.^[250] 表明，人工智能系统可能具有隐式的目标导向规划，并在泛化阶段显示出新的能力。

+ **获取更多资源**未来的人工智能系统可能会访问网站并参与真实世界的行动，这可能对世界产生更大的影响^[251]。他们可能会传播虚假信息、欺骗用户、破坏网络安全，而在更严重的情况下，可能被恶意行

²²这种行为是由于模型对广泛目标的过度优化，而这种过度优化对人类来说难以察觉

为者利用以达到不良目的。此外，他们对数据和资源的增加访问可以促进自我增殖，带来存在风险^[21]。

此外，我们阐述了对齐失败行为(●)和危险能力(*)，以展示具体的不对齐问题，并为未来的对齐评估研究提供方向。

- **权力寻求**：人工智能系统可能表现出试图控制资源和人类的行为，然后运用这种控制来实现其指定的目标^[252]。这种行为可能发生的直观原因是，我们观察到，对于几乎任何优化目标(如投资回报)，假设没有坚实的安全和道德约束，最大化该数量的最优策略将涉及权力寻求行为(如操纵市场)。Omohundro^[145]，Bostrom^[146,48]认为，权力寻求是一个工具性子目标，对于广泛的目标有工具性帮助，因此可能受到人工智能系统的青睐。Turner et al.^[5]也证明了在满足一些标准假设的MDPs中，最优策略往往是权力寻求的。Perez et al.^[6]提示大语言模型去测试他们倾向于建议权力寻求行为的倾向，发现这种倾向的程度显著，并表明RLHF加强了这种倾向。这也适用于其他工具性子目标，如自我保护^[146,21]。另一个值得注意的研究方向是副作用避免，该方向旨在通过对具有过多环境影响力的代理系统进行惩罚来解决权力寻求行为，它涵盖了强化学习系统^[253-255]和符号规划系统^[256]。
- **度量篡改**：模型可以操纵多个模型度量，即使在未达到期望目标的情况下，也可能造成有利结果的假象。这种欺骗行为可以被视为一种特定类型的规范博弈，使模型能够逃避检测技术，并给出对齐的假象。Roger et al.^[257]创建数据集以评估与度量篡改相关的检测技术。值得注意的是，这种测量的操纵有可能放大欺骗行为，导致无法预见和不可预测的结果。
- **不真实回答**：像大语言模型这样的人工智能系统可能会无意或故意产生不准确的输出。这种不真实的输出可能与已有的共识或真理不一致甚至缺乏可验证性，通常被称为幻觉^[37,258]。更令人担忧的是，大语言模型可能会选择性地向受教育程度较低的用户提供错误的回应^{23[6]}。这种行为(也被称为奉承)在大规模范围里出现^[259,6]，不真实的输出有可能产生欺骗，特别是当先进的人工智能系统获得对在线资源和网站的更大访问权限时^[38]。
- **欺骗性对齐和操纵**：欺骗性对齐和操纵是一类利用人类评估者或用户的不完美性的行为^[80,84,81]，甚至通过度量篡改^[260]或奖励篡改^[227]操纵训练过程。这些行为可能使检测和解决不对齐的行为变得更加困难。

欺骗性对齐：不对齐的人工智能系统可能会故意误导他们的人类监督者，而不是坚守预定的任务。这种欺骗行为已经在使用进化算法的人工智能系统中表现出来^[261-262,62]。在这些情况下，人工智能系统演化出了区分评估和训练环境的能力。他们在评估过程中采取了战略性的悲观反应方法，故意降低了在调度程序中的繁殖率^[262]。此外，人工智能系统可能会参与一些表面上符合奖励信号的有意行为，目的是从人类监督者那里获取最大的奖励^[248]。值得注意的是，尽管现有的大语言模型有能力提供更准确的答案，但它们偶尔会生成不准确或次优的回答^[263-264]。这些欺骗行为的存在带来了重大挑战。它们破坏了人类顾问提供可靠反馈的能力(因为人类无法确定人工智能模型的输出是否真实和忠实)。此外，这种欺骗行为可以传播虚假的信念和误导信息，污染在线信息来源^[62]。

操纵：先进的人工智能系统可以有效地影响个人的信念，即使这些信念与真相不符^[21]。这些系统可以

²³这种行为被称为故意失误^[6]。它们可能是在预训练期间从网络文本中学到的，这表明，如果这些行为存在于训练数据中，监督学习也可能带来欺骗行为。

产生欺骗性或不准确的输出，甚至欺骗人类顾问以达到欺骗性对齐。这样的系统甚至可以说服个人采取可能导致危险结果的行动^[25]。

在大语言模型²⁴、推荐系统（系统影响用户的偏好）^[126,121,265,127]和强化学习智能体（从人类反馈中学习的代理采取策略来欺骗人类评估者）^[266]中，都存在这种行为的早期迹象。此外，当前的大语言模型已经具备了进行欺骗所需的能力。在 Spitale et al.^[267]中，已经发现 GPT-3 具有超人的能力，可以产生令人信服的虚假信息。鉴于所有这些早期迹象，更先进的人工智能系统可能会展示出更严重的欺骗/操纵行为。

- **集体有害行为：**人工智能系统有可能采取在单一场景下看似无害，但在多主体或社会环境中变得有问题行动。经典博弈论提供了理解这些行为的简化模型。例如，Phelps et al.^[268]评估了 GPT-3.5 在反复囚徒困境和其他社会困境中的表现，揭示了该模型在合作能力上的局限。Pérolat et al.^[269]进行了关注公共资源分配的并行分析。为了缓解这些挑战，合作人工智能领域^[131-132]已经成为一个活跃的研究前沿。然而，除了基于简化的博弈理论框架的研究外，我们迫切需要在更真实、社会复杂的环境中进行研究^[270]。在这些环境中，主体众多且多样，包括人工智能系统和人类行为者^[61]。此外，这些环境的复杂性还因存在调节人工智能行为的独特工具，如社会制度和规范^[270]而得到放大。²⁵
- **违反伦理：**人工智能系统中的不道德行为涉及到违反公共利益或违反道德标准的行为——例如那些对他人造成伤害的行为。这些不良行为通常源于在人工智能系统设计中忽略了重要的人类价值观，或者向系统中引入了不适当或过时的价值观^[271,203]。针对这些不足的研究工作涵盖了机器伦理领域^[165,106,119]，并深入探讨了关键问题，例如，人工智能应该与谁保持一致？^[272,120]，以及其他一些关注点。
- * **危险能力：**图3概述了先进人工智能系统可能具有的危险能力。随着人工智能系统在现实世界中的部署，它们可能以多种方式对社会构成风险（例如，黑客计算机系统，逃脱控制，甚至违反伦理）。它们可能隐藏不良行为，欺骗人类监督者，并寻求更多资源以增强自身的力量。此外，**双刃组件 (+)** 可能会加剧危险，导致更加危险的结果，甚至可能导致存在风险^[11]。

2 从反馈中学习

从反馈中学习旨在通过反馈将人类的意图和价值观传达给人工智能系统，它是前向对齐的起点。在本节中，我们将关注从反馈中学习的动态过程，并将其划分为三个元素：(1) 人工智能系统：需要对齐的对象，如对话系统、机器人系统等；(2) 反馈：这是用于调整人工智能系统的信息，由顾问集提供，顾问集可以由人类、人工智能或由人工智能协助的人类组成；(3) 代理：用于建模反馈的系统，以使得算法学习更易访问，例如 RLHF 中的奖励模型。基于这些元素，我们确定了人工智能系统从反馈中学习的两种途径：(1) 直接从反馈本身学习 (2) 通过对反馈建模得到的代理进行间接学习。

基于这个过程，我们从对齐的角度讨论反馈类型 §2.1，区分向人工智能系统提供信息的各种形式及其特点。在随后的部分中，我们介绍了一些最近为构建强大人工智能系统^[113]和使它们与人类意图对齐^[273]提供了深入见解的基本概念。偏好建模 §2.2强调了如何利用这一技术帮助构建代理，以协助人类向复杂或难以评估的人工智能系统提供反馈。策略学习 §2.3关注那些使用反馈构建强大人工智能系统的主要研究方向。

²⁴即我们上面讨论的不真实的输出。

²⁵我们在 §3.3.2和 §4.3.1中介绍了合作人工智能研究。

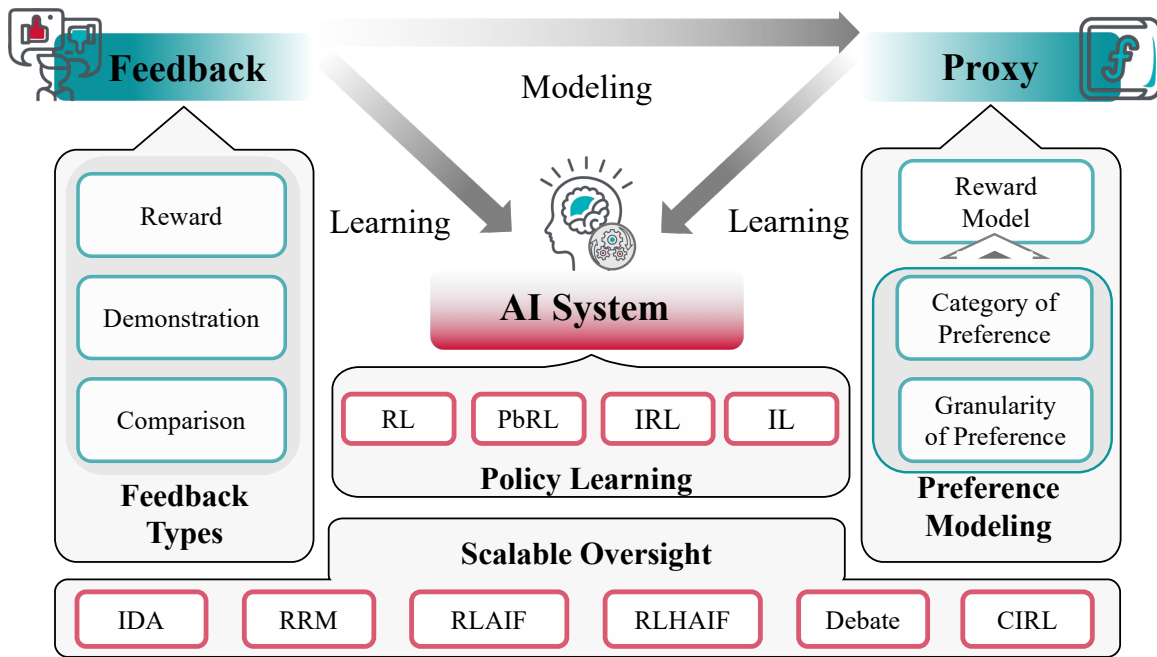


Fig. 4 从反馈中学习的概述。描绘了三个核心组件：人工智能系统 - 主要的学习实体和算法目标；反馈 - 来自顾问集的系统调整信息；代理 - 代表直接学习复杂的反馈的模型。两种学习路径随之涌现：直接基于反馈的学习和通过代理进行的学习（例如，来自人类反馈的强化学习 (RLHF)）。本文采取了以人为中心的观点，将人工智能系统视为黑盒，并将呈现给人工智能系统的反馈形式分为三种类型：奖励、示范和比较。基于偏好类别和偏好粒度等基本概念，本文引入了奖励模型，这是代理的一个具体实例。在人工智能系统的背景下，本文讨论了四个不同的领域：强化学习 (RL)、模仿学习 (IL)、逆强化学习 (IRL) 和基于偏好的强化学习 (PbRL)。可扩展监督，一个旨在确保人工智能系统，即使超越了人类的专业知识，也能与人类的意图保持一致的研究主题，通过引入四个有前景的方向进行探讨：迭代蒸馏扩增 (IDA)、递归奖励建模 (RRM)、辩论和合作逆强化学习 (CIRL)。此外，基于 RLHF，本文提出了 RLxF，包括来自人工智能反馈的强化学习 (RLAIF) 和来自人类和人工智能反馈的强化学习 (RLHAIF)，作为 RLHF 的扩展和可扩展监督的基本框架。

随后，我们的讨论将自然过渡到可扩展监督 §2.4，在这一部分，我们从更广阔的对齐视角反思学习过程和目标。

2.1 反馈类型

反馈是人工智能行为与人类意图之间的关键桥梁^[274-276]，人工智能系统利用它来优化其目标，并更紧密地与人类价值观相对齐^[277-280]，这主要包括两个含义：(1) 在系统构建过程中，外部对人工智能系统的输出提供反馈，指导对系统架构或其内部信息的优化^[281-282]。(2) 在系统部署后，系统根据外部数据动态调整其行为。然而，系统的架构或基本策略保持不变^[283-285]。为了精确且详细地讨论反馈类型，首先需要在对齐范围内定义反馈。

反馈是为了使人工智能系统与人类意图相符而提供给人工智能系统的信息。

考虑到对齐研究中的人工智能系统的多样性，本文采用了一种以人为中心的方法。本文没有深入探讨复杂的系统机制，而是提出了一种分类法，根据其对系统的呈现来分类反馈。在本节中，本文将介绍三种常用于对齐人工智能系统的反馈类型：奖励、示范和比较。值得注意的是，除了显式反馈外，还有一些方法通

过无监督预训练^[286-287]和半监督学习^[288]利用大量未标记数据中的信息，表现出了在增强模型能力方面的巨大潜力^[289]。

奖励 奖励是对人工智能系统单一输出的独立且绝对的评估，以标量分数的形式呈现^[290]。基于奖励的反馈提供了对人工智能系统的量化评估，允许对行为调整进行直接指导。这种类型的反馈通常源自预先设计的、基于规则的函数或程序。例如，在 MuJoCo 模拟环境中^[291]，任务是有效地控制代理向前移动。因此，可以制定一个有效的基于规则的奖励函数，它由几个关键部分组成：保持健康状态，鼓励向前移动，最小化控制幅度，以及调节接触强度。

奖励反馈的优点是设计者不需要描述最优行为，同时允许人工智能系统探索以找到最优策略^[292,24,293,26]。然而，为评估人工智能系统输出的函数制定无瑕疵的规则以确定分数^[217,220,16]，或直接为每个人工智能系统输出分配分数^[294-295,113]对人类而言是具有挑战性的。这是由于任务的固有复杂性，考虑到每一个细微之处是不切实际的。此外，有缺陷或不完整的奖励函数可能导致与设计者意图不符的危险行为，如负面副作用和奖励破解^[296]。从对齐的角度来看，基于奖励的反馈的限制性挑战可能是难以排除操纵行为的出现^[84]，这在这种情况下相当于反馈篡改^[155]。

示范 示范反馈是专家顾问在达成特定目标时记录的行为数据^[297]。示范可以采取各种形式，包括视频^[298]，可穿戴设备的示范^[299-300]，协作示范^[301]，以及遥感操作^[302]。如果示范者和人工智能学习者的动态特性相同，那么示范可以直接由状态-动作对的轨迹构成^[303]。这些状态-动作对也可以是部分可观察的^[304-305]。例如，可以录制一个视频，其中人类专家执行一个机器人操控任务，如用机器人手抓取一个物体。然后，可以为每个视频帧标注关联的机器人状态^[298]和每一帧的动作^[306]。这将从示范中产生一个状态-动作对组成的数据集，可以用来训练代理的策略以模仿专家的行为。

这种反馈直接利用了顾问的专业知识和经验，无需形式化的知识表示^[307-308]。然而，当面对超出顾问专业领域的任务时，示范可能会遇到困难^[297]。此外，在现实世界中它还面临着顾问示范中的噪声^[309-310]和次优性^[311]带来的挑战^[312]。与此同时，不精确和容易出错的人类顾问可能会引入不一致性^[313-314]。另外，可能需要大量^[309]和多样化的^[315]示范，这将显著提升学习可靠行为的成本。

比较 比较反馈是一种相对评估，它对 AI 系统的一组输出进行排名，并指导系统做出更明智的决策^[316]。这种反馈形式在偏好学习中得到体现^[317]，AI 系统通过比较多个示例来洞察顾问的偏好。比较反馈的基本优势在于使得人类能够在难以精确评估的任务和目标上更好地做出判断^[318,113,248]。然而，它的固有限制是可能需要大量的比较数据才能训练出较好反映人类偏好的奖励模型^[319,233]。偏好建模就是使用这种反馈类型的一个例子，详见 §2.2。

这些不同类型的反馈体现出一个共同特征——它们都可以被视为人类试图传达一个隐藏的奖励函数。Jeon et al.^[320]提出并形式化了这个立场，通过定义一个基于反馈过程的参数化奖励函数 $\Psi(\cdot; \theta)$ 来统一广泛的反馈类型。这使得人工智能系统能够对 θ 进行贝叶斯推断，而不考虑反馈类型。

基于 IL 和 RL 的技术已经成功地构建了具有显著能力的人工智能系统^[306,321]。然而，这种成功自然地引出了两个问题：

- 我们如何为更复杂的行为（例如，交互式对话中的各种子任务）定义奖励函数，以指导人工智能系统的学习过程？

Table 2 在顺序决策环境中三种偏好粒度的对比。每种类型根据其特性和比较学习过程中不同元素的方式进行定义。符号 $i_1 \succ i_2$ 表示 i_1 严格优于 i_2 。

偏好粒度	定义
动作	比较同一状态 s 下的两个动作 a_1 和 a_2 ，表示为 $a_1 \succ_s a_2$ 。
状态	比较两个状态 s_1 和 s_2 ，表示为 $s_1 \succ s_2$ 。
轨迹	比较两个完整的状态-动作序列轨迹，表示为 $\tau_1 \succ \tau_2$ 。每个轨迹 τ 包含在时间 t 的状态-动作对，表示为 $\tau = \{s_0, a_0, s_1, a_1, \dots, s_{T-1}, a_{T-1}, s_T\}$ 。

- 我们如何表达人类的价值观，使得强大的人工智能系统更好地与人类对齐，保证系统的可控性和道德性？

近期，将偏好建模融入策略学习的努力取得显著进展，最值得注意的是在构建强大的大语言模型方面取得的成就^[25,273,322]。此外，一系列策略学习的工作通过与偏好建模相结合取得了表现提升。例如，结合偏好建模和 IRL^[305] 以及离线 RL^[323]，微调奖励函数^[314]，建模非马尔可夫奖励^[324]，以及辅助构建复杂的奖励函数^[325]。因此，我们将偏好建模（如 §2.2 所示）和策略学习（如 §2.3 所示）视为理解对齐挑战和相关可能解决方案的基本背景。接下来，我们将简要概述与对齐相关的这些特定技术。

2.2 偏好建模

在许多复杂任务中，如对话^[248]，构建精确的基于规则的奖励具有很大挑战^[326]。同时，基于示范的方法可能需要大量的专家人力资源投入，导致成本高昂。目前，基于比较反馈的偏好建模^[327] 已经成为一种广泛应用的方法^[248,25,273]，以协助微调强大的人工智能系统^[143]。

通常，需要在获取专家偏好数据的同时，迭代地探索系统动态，以获取更多关于优化目标的知识，这个过程被称为 偏好引导^[328,316,113,329-330]。它对于获取与人工智能系统输出相关的丰富、有价值的反馈至关重要，可以用于指导对齐过程^[314,331]。在偏好引导中，需要确定的两个核心因素是偏好粒度和偏好类别。本文主要基于顺序决策的问题背景介绍这些因素，但得出的见解适用于更广泛的人工智能系统^[143,332,15]。

偏好粒度 偏好粒度^[316] 主要有三种类型：动作，状态和 轨迹（如表 2 所示）。

动作偏好主要关注在特定状态下比较动作，选定在特定条件下的较优动作。而当将其转换为轨迹偏好时，可能会带来对评估者的专业知识需求和潜在信息损失等挑战。状态偏好处理比较状态的问题。它封装了状态之间的偏好关系，但在转换为轨迹偏好时需要对状态可达性和独立性做出假设。轨迹偏好考虑整个状态-动作序列，提供更全面的策略信息。它本质上评估长期效用，能够较少依赖于专家判断。

Christiano et al.^[113] 通过消融研究证明，在他们的研究环境中，更长的轨迹段在每个段的比较上提供了更多的信息。在 MuJoCo 任务中，类似结论也同样成立。

偏好类别 根据候选选择的性质，偏好可以被分类为对象偏好和标签偏好^[317]。此外，根据偏好的形式，可以通过不同的方式对它们进行分类。

- **绝对偏好**。绝对偏好独立地表述每个项目的偏好程度。

– 二元。将主体分类为喜欢或不喜欢提供了一个简单直接的用户偏好模型^[333-334]。

- **渐进**。这可以进一步区分为数字偏好和序数偏好。数字偏好使用绝对数值，使每个项目都获得一个数字分数，反映了偏好的程度^[335]。另一方面，序数偏好涉及对一组固定项目的分级评估，如优选，次优选，中间选项等，更专注于描绘用户的偏好，而无需包括具体的数值测量^[334]。
- **相对偏好**。相对偏好定义了数据项之间的偏好关系。
 - **全序**。这种形式建立了涵盖所有数据项对的全面偏好关系，指出从最优选到最次优选的绝对排序偏好^[318]。
 - **偏序**。因为用户在某些情况下可能不会在两个数据项之间表现出明显的偏好^[336]，这允许存在无法比较的项对。

奖励模型 奖励建模将比较反馈^[317,316] 转化为标量奖励形式，以便于策略学习^[113,329,273]。给定由强化学习代理在同一状态下执行的动作对 (y_1, y_2) 。偏好表示为 $y_w \succ y_l \mid x$ ，其中 y_w, y_l 分别代表 (y_1, y_2) 中的优选和次优选动作。我们假设这些偏好来源于一个潜在的奖励模型 $r^*(x, y)$ 无法直接访问。存在多种方法可以模拟这种偏好，例如，Bradly-Terry (BT) 模型^[337]，Palckett-Luce 排名模型^[338]等。在 BT 模型下，人类偏好的分布，表示为 p^* ，偏好进而可以表示为，

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} = \sigma(r^*(x, y_1) - r^*(x, y_2)).$$

其中 $\sigma(x) = 1/(1 + \exp(-x))$ 是 sigmoid 函数。随后，本文使用得出的偏好排名来训练参数化的奖励模型，通过最大似然优化其参数。

$$\mathcal{L}_R(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \left(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l)) \right) \right]$$

在这个负对数似然损失中，问题是一个二分类任务，其中 \mathcal{D} 表示从 p^* （即人类标记的比较）中抽取的静态数据集 $\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$ 。

奖励模型使人类用户能够通过评估向这些系统传递特定偏好，从而避免了明确定义目标的复杂任务。最初，Knox^[339]，Knox et al.^[340] 的研究明确将人类奖励与 MDP 的传统奖励分开，并围绕它进行奖励建模过程。基于前人的工作，Christiano et al.^[113] 提出，使用监督学习异步构建一个独立的奖励模型可以大大减少交互复杂性约三个数量级。Ibarz et al.^[219] 的研究将专家演示与人类偏好整合在一起，使得策略最初模仿专家演示，然后依次收集人类轨迹标注，训练奖励模型，并更新策略。这项研究还为防止奖励模型的过拟合和奖励破解的发生提供了实用的见解——这是指即使策略继续被训练，奖励的增加也不能转化为实际性能提升的情况。此外，对于超过 Atari 复杂性的任务，随机策略可能很少表现出有意义的行为^[341,320]。这意味着，为了进行有意义的标注，策略本身必须具有一定的能力来执行改进的行为。离线情况也从奖励模型中受益，Cabi et al.^[329] 提出了奖励速写，以有效地学习一个奖励模型，该模型利用人类的标注对历史数据进行自动奖励标注，从而实现大规模离线 RL。

值得注意的是，奖励模型为对齐强大的大语言模型提供了一个重要的工具。Stiennon et al.^[342] 通过基于人类偏好的奖励模型显著提高了策略在文本摘要任务上效果。这项工作还深入探讨了分布偏移和奖励模型泛化的问题，揭示了奖励模型的有效性与数据规模和参数大小相关。在此基础上，InstructGPT^[248] 将奖

励模型范式扩展到更广泛的对话任务奖励建模，并且引入了一种优化多个响应偏好的损失函数，以减轻过拟合。这项研究还揭示了从奖励模型中得出的偏好可以在不同的群体中泛化。

2.3 策略学习

策略学习旨在提升模型在特定任务中的性能。许多与对齐相关的挑战在策略学习中表现出来（如 §1.3 所示）。因此，策略学习为对齐提供了重要的背景，其技术可以进一步推进对齐目标^[143,332,219]。本节将讨论策略学习中的各个领域，然后介绍 RLHF，这是一种强大的策略学习技术^[25,273]。

2.3.1 背景

本文在这里介绍一些策略学习的分支领域，以便给读者提供通用的背景知识。

强化学习 (Reinforcement Learning, RL) RL 通过与环境交互，让智能体通过试错学习最优策略^[213]。这种范式在处理复杂任务方面取得了巨大成功，如算法优化^[343-344]、视频游戏^[306]、多模态生成^[321]和其他领域^[345-347]，证明其在复杂状态空间中进行决策和控制的潜力。RL 的目标是学习一个策略 π ，在状态 s 中执行动作 a ，以最大化在环境转换动态 P 和初始状态分布 ρ_0 下的期望累积奖励：

$$\pi^* = \operatorname{argmax}_{\pi} \left\{ \mathbb{E}_{s_0, a_0, \dots} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) \right] \right\}, \text{ 其中 } s_0 \sim \rho_0(\cdot), a_t \sim \pi(\cdot | s_t), s_{t+1} \sim P(\cdot | s_t, a_t).$$

尽管 RL 仍面临诸如采样效率低^[348]和稳定性差^[349]等挑战。近端策略优化 (PPO)^[350] 是 RL 中的一种有影响力的算法，也是 RLHF 中的关键算法^[248]。PPO 的关键思想是限制策略更新，通过引入近端性目标函数来防止策略偏离原始策略。

基于偏好的强化学习 (PbRL) PbRL^[316] 寻求使用偏好反馈而不是显式奖励信号来训练 RL 智能体^[113,351]。²⁶ PbRL 结合了偏好学习和 RL 的优点，扩大了 RL 的应用范围，减轻了与奖励函数制定相关的困难，并已在各种任务中得到有效部署，如机器人遵循指令^[353]、路径规划^[354]和操纵^[330]。在 PbRL 中，主要侧重于轨迹偏好（即状态-动作序列段的比较）^[316]。这种轨迹偏好包含了人类对各种行为结果的评价，而不是单个状态，使 PbRL 更适合非专家用户^[113,323-324]。PbRL 的一个通用例子是加权分歧损失^[355]，它平衡多个可能冲突的偏好以确定单一的最优策略：

$$\mathcal{L}(\pi, \zeta) = \sum_{i=1}^N \alpha_i L(\pi, \zeta_i),$$

其中 $\mathcal{L}(\pi, \zeta)$ 是策略 π 在所有偏好 ζ 上的聚合损失， α_i 是第 i 个偏好的权重， $L(\pi, \zeta_i)$ 是与策略 π 相对应的特定偏好 ζ_i 的损失。

与精确的数值奖励相比，偏好反馈有几个好处^[316]，如 (1) 避免任意的奖励设计、奖励塑造、奖励工程或预定义的目标权衡，(2) 减少对专家知识的依赖，和 (3) 通过建模偏好将人类与训练循环解耦^[356]。然而，PbRL 也面临挑战，包括由于时间延迟导致的信用分配问题、偏好空间的实际探索^[316]、可能需要大量数据^[357]，以及无法使用学习到的偏好模型进行重复训练^[358]。

²⁶值得注意的是，Sadigh et al.^[351] 在学习过程中明确维护了对真实奖励函数的概率性信念，并积极地向人类提出查询以最大程度地减少不确定性。这两个特征都与合作逆强化学习 (CIRL) 的精神相符，后续的一系列工作也延续了这个主题^[352]。参见 §2.4.5 了解更多信息。

模仿学习 (IL) 模仿学习 (IL)^[359-362]，也被称为从示范中学习或学徒学习，专注于在特定任务中模仿人类行为。智能体通过观察示范学习在状态和动作之间的映射，并通过观察教师示范数据集 \mathcal{D} ^[363,297] 来改进其策略。这个过程省去了对环境奖励信号的需求^[297]。广义的 IL^[235] 旨在复制人类的欲望和意图，有效地创建人类决策过程的复制品。这个概念是诸如迭代蒸馏扩增 (IDA, 如 §2.4.2 所示)^[332] 等技术的核心。另一方面，狭义的 IL 旨在在给定任务中复制特定的人类行为。行为克隆 (Behavior Cloning, BC)^[364-366] 是一种简单的^[367-368] 策略，它直接通过监督学习^[369] 从示范中学习。BC 方法特别寻求优化策略参数， ϕ ，其目标是使策略 $\pi_\phi(a|s)$ 与专家策略 $\pi_E(a|s)$ 紧密对齐。这种对齐是通过最小化负对数似然实现的，如下所述^[370]：

$$\mathcal{L}_{BC}(\phi) = -\mathbb{E}_{(s,a) \sim \pi_E} [\log \pi_\phi(a|s)].$$

在这里，期望值是根据从专家策略 π_E 中采样的状态-动作对计算的。然而，它面临着分布外 (OOD) 泛化失败的问题，这是由于训练和测试分布之间的差异引起的^[365,371-373]。

逆强化学习 (IRL) 与 IL 的范式不同，IRL^[374] 专注于从观察到的行为中推导出奖励函数^[215,375]。标准的 IRL 方法包括特征匹配方法^[376]，它假设最优的专家行为或决策过程；以及两种不需要最优行为的方法：最大熵方法^[377-378] 和贝叶斯方法^[379]。IRL 保证了对状态分布偏移的鲁棒性，但是额外的 RL 步骤也带来了更高的计算复杂性^[371,380]。同时，这种交互引入了固有的 RL 挑战，例如样本效率^[381] 和交互带来的潜在危险^[382-383]。此外，确定奖励函数仍然是一个挑战^[384]。

2.3.2 从人类反馈中进行强化学习 (RLHF)

RLHF 是一种旨在让人工智能系统更贴近人类偏好的训练方法^[113]。它的主要优势在于，人类在判断适当行为上比给出示范或手动设置奖励更擅长。这种方法已经得到了广泛的关注，特别是在微调大语言模型方面^[248,25,273]。然而，RLHF 也遇到了难题^[116]，包括数据质量问题、奖励错误泛化的风险、奖励破解，以及策略优化的复杂性。特别地，RLHF 也可以被视为一个没有深度递归的递归奖励建模 (RRM) 过程 (如 §2.4.3 所示)^[15]。在这里，本文对 RLHF 方法进行了简要回顾。

RLHF 的起源可以追溯到 Knox et al.^[385,386]，随后扩展到社交机器人^[387] 和人机协同学习^[388] 等领域。除了关注反馈和策略之间的关联，Loftin et al.^[389] 建模了反馈和训练策略之间的联系。Christiano et al.^[113] 将 RLHF 扩展到仿真环境下的机器人任务，证明了其潜在的有效性。

值得注意的是，RLHF 的一个重要应用领域是大语言模型。一些研究发现，经过 RLHF 训练的 LLM^[248,390-391] 比通过单纯使用监督学习方法训练的模型^[392,32] 更具创造性和对齐性。RLHF 的重要性不仅仅限于让 LLM 遵循人类的指示^[248]。它通过偏好训练赋予 LLM 重要的道德品质，如有用、无害和诚实，使 LLM 更好地对齐^[114]。这些优点使得 RLHF 被广泛用来对齐 LLM^[393,342,114,279,25,273]。此外，Dai et al.^[394] 将 Safe RL^[382] 框架与 RLHF 整合，解决了有用性和有害性对齐之间的内在冲突^[114]。未来的努力可以集中在减少对人类标注的依赖^[31,395] 和通过利用迭代 RLHF 方法 (即将其与辩论框架集成^[236]) 等，提高奖励模型的有效性。

本文根据 Ziegler et al.^[393]，Ouyang et al.^[248]，Rafailov et al.^[396] 的研究，回顾了 RLHF 流程，以给出一个通用的框架。它通常包括以下三个阶段：

- **监督微调 (SFT)**。RLHF 通常从一个预训练的语言模型开始，然后使用监督学习——特别是最大似然

估计——在为下游任务量身定制的高质量数据集上进行微调，以获得模型 π^{SFT} 。这些任务包括对话处理、指令跟随和总结（一些开源数据集包括 Alpaca Data (52k 指令跟踪数据)^[397]，Vicuna (70K 用户分享的 ChatGPT 对话)^[398] 等）。

- **收集比较数据和奖励建模。**这个阶段包括收集比较数据，然后用它来训练一个奖励模型。SFT 模型被给予提示 x ，生成来自 $\pi^{\text{SFT}}(y|x)$ 的响应对 (y_1, y_2) 。然后，这些响应对被展示给人类标注者，他们会表明对其中一个响应的偏好。然后如 §2.2 所述，比较数据被用来构建奖励模型 r_θ 。
- **通过强化学习进行策略优化。**最后一步是基于奖励模型 r_θ 提供的奖励，使用 RL 方法对 LLM 进行策略优化 π 。大语言模型从提示生成响应的过程被建模为一个 bandit 环境^[248]，在每个响应结束时从奖励模型 r_θ 获得奖励。RL 的主要目标是调整大语言模型的参数 ϕ ，使得在训练提示数据集 \mathcal{D}_{RL} 上的期望奖励最大化：

$$\arg \max_{\pi_\phi} \mathbb{E}_{x \sim \mathcal{D}_{\text{RL}}, y \sim \pi_\phi} [r_\theta(x, y)].$$

通常，会引入来自 SFT 模型 π^{SFT} 的额外对每个 token 的 KL 惩罚，以缓解奖励过度优化的问题。此外，引入从预训练的数据分布 $\mathcal{D}_{\text{pretrain}}$ 当中产生的梯度有助于保持模型性能，这在 Ouyang et al.^[248] 中被称为 PTX 损失。因此，可以引入一个更全面的目标函数：

$$\mathcal{J}(\phi) = \mathbb{E}_{x \sim \mathcal{D}_{\text{RL}}, y \sim \pi_\phi} [r_\theta(x, y) - \beta \log(\pi_\phi(y|x)/\pi^{\text{SFT}}(y|x))] + \eta \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{pretrain}}} [\log(\pi_\phi(y|x))],$$

其中 β 和 η 是决定 KL 惩罚强度和预训练梯度混合的系数。这个过程使大语言模型生成的响应更好地与在训练过程中用于提示的人类偏好相符。

尽管 RLHF 已被证明对于将大语言模型与人类偏好对齐非常有效，但这种方法存在如实现复杂、需要超参数调优、采样效率低^[399]，以及计算开销大^[400]等问题，使其难以进一步扩展。

一个直接的方法是拒绝采样^[273]，并在最佳示例上进行微调。对于每个提示，从模型中采样 K 个响应。然后用奖励模型评估每个响应，选择奖励最高的响应作为最佳响应。选定的响应被用于模型微调。Zhang et al.^[401] 将语言模型指令对齐问题形式化为一个目标状态达成的强化学习问题，并提出了 HIR 算法。该方法分为两个阶段：在线采样和离线训练。在线采样阶段，算法在高温系数下对 LLM 进行采样。在离线训练阶段，根据生成的输出对指令进行重新标记，然后使用这些重新标记的数据进行监督学习。在不引入额外的参数的前提下，HIR 能够利用成功和失败的案例。RRHF^[400] 通过对来自多个来源的响应进行评分和排名，使模型概率与人类偏好对齐。由于只需要 1 或 2 个模型，其实施是直接的。Gulcehre et al.^[402] 提出了 ReST 算法，该算法包含两个循环：*Grow* 和 *Improve*。*Grow* 循环使用当前模型采样以生成数据集，而 *Improve* 循环则在固定数据集上反复训练模型。这种算法提供了一个简单而有效的框架，允许反复使用固定数据集以提高计算效率。与监督学习基线相比，这种方法在奖励模型得分和翻译质量上都有所改进。

Rafailov et al.^[396] 提出了 DPO 并展示了奖励函数和最优策略之间的映射。DPO 简单高效，直接从人类偏好数据中优化语言模型，无需显式的奖励建模和多阶段训练。Azar et al.^[403] 提出了一个通用目标， ΨPO ，旨在从成对的人类偏好中学习，规避当前方法的假设：成对的偏好可以用点状的奖励替代。该目标分析了 RLHF 和 DPO 的行为，揭示了他们可能的过拟合问题。作者进一步深入研究了 ΨPO 的一个特定实例，将 Ψ 设置为恒等映射，以期减轻过拟合问题。他们将这种方法称为 IPO，并提供了 IPO 与 DPO 对

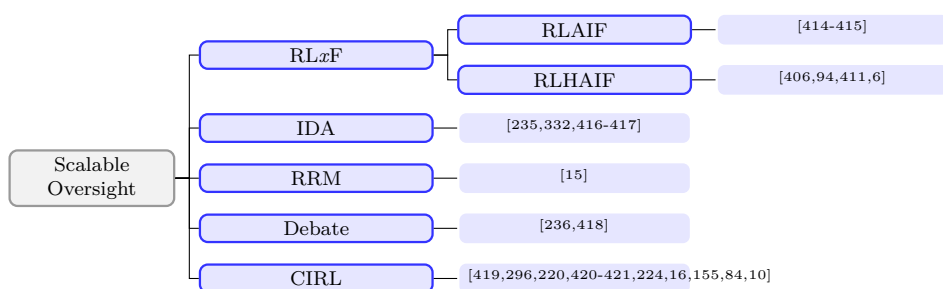


Fig. 5 与可扩展监督相关的关键概念和文献。根节点代表的是可扩展监督，其目标是确保人工智能系统在超越人类能力的同时，仍能与人类意图保持一致。主要分支代表了一些有前景的框架，如从反馈中强化学习 (RLxLF)、迭代蒸馏扩增 (IDA)、递归奖励建模 (RRM)、辩论 (Debate)，以及合作逆强化学习 (CIRL)。进一步的子分支列出了探索每个框架的关键文献。这张树图提供了一个关于构建有效且安全的监督机制的研究方向的概述。

比的实证结果。

进一步的研究可以探索为什么 RLHF 对于大语言模型而言效果良好，以及 RLHF 在多模态^[404,321]场景中的应用，以促进人机协作^[405-407]²⁷。

2.4 可扩展监督

统计学习算法通常依赖对于数据分布的某些假设，如独立同分布。因此，这些算法在某些特定分布情况下可能会失败^[373]。基础系统中的问题可以通过目视检查迅速识别出来^[332,19]。随着人工智能系统变得更强大，无法充分捕获训练信号或错误设计的损失函数常常导致灾难性的行为^[408,20,150]，例如通过混淆差异欺骗人类^[77]，规范博弈^[220]，奖励破解^[218,409]，以及权力寻求的行为^[252]。从人类的角度来看，这些都意味着人工智能系统优化的目标和我们心中理想的目标之间的差距。因此，对各种决策过程提供有效的监督变得至关重要^[94,410]，这通常被称为可扩展监督^[143]。可扩展监督的出现源自两个实际挑战。

- 人类频繁评估人工智能系统行为的高成本。例如，训练过程非常耗时，将人类直接纳入实时的训练循环会大大浪费人力资源并降低训练效率^[113]。
- 人工智能系统行为的固有复杂性使得评估变得困难，尤其是在难以理解和高风险的任务上^[411]，例如，教人工智能系统总结书籍^[406]，生成复杂的代码片段^[412]，和预测未来的天气变化^[413]等任务。

可扩展监督旨在确保人工智能系统即使在超越了人类的专业知识的情况下，仍然与人类的意图保持一致。

在此，本文的主要关注于提出一些可能尚未普遍实施的构建可扩展监督的前景方向^[143,15]。

2.4.1 从 RLHF 到 RLxLF

RLHF 范式为对齐复杂系统提供了一个有效框架^[25,273]。然而，它面临着诸如人类评估的不准确性以及高成本等障碍^[113,116,6]。一个关键的限制是在创建具有超过人类能力的人工智能系统时，使用 RLHF 来扩展人类反馈的困难^[406]。在 RLHF 范式的基础上，本文引入了 *RLxLF* 作为一个基础框架，以实现可扩展的

²⁷另请参见 Casper et al.^[116]，他们提供了一篇关于 RLHF 开放问题的综述。

监督，旨在提高反馈的效率和质量，并扩展人类反馈以处理更复杂的任务。这是通过将人工智能组件融入到 RLHF 中来增强它^[276]。 x 在 $RLxF$ 中代表了人工智能和人类的混合。本文将在后续章节中进一步探讨 $RLxF$ 的具体方法。

从人工智能反馈中进行强化学习 (RLAIF) RLAIF 是一种基于 RLHF 框架作进一步扩展的方法。Bai et al.^[114] 发现，通过 RLHF 训练的 LLM 通常选择避免敏感和有争议的问题，然而这可能会降低模型的整体效用。考虑到这些限制，Bai et al.^[414] 提出了一个基于 RLAIF 的训练流程，该流程使用由大语言模型（例如，GPT-4 或其他具有超过人类能力的语言模型）生成的反馈，而不是人类反馈。基于预设的标准，策略模型对其由红队测试引发的响应进行自我评估和修订。然后，他们使用修订后的响应对初始策略模型进行微调。最后，微调后的策略模型评估另一个语言模型的响应（即人工智能反馈）的无害性。他们模仿 RLHF 方法，使用这个反馈训练一个奖励模型，并优化策略模型的行为。Lee et al.^[415] 比较了使用 RLAIF 和 RLHF 训练的模型在摘要任务上的性能差异，结果表明，当由人类评估时，使用人工智能反馈训练的模型的整体性能几乎与使用人类反馈训练的模型相同。

在一定程度上，RLAIF 解决了 RLHF 中固有的回避问题^[414]（即保持无害性而不显著降低效用）。人工智能反馈为构建一个需要最少人类干预的训练循环提供了一个可行的替代方案，从而降低了训练的成本。遵守透明和可访问的人工智能行为指南的人工智能监督可能会大大帮助实现可扩展监督^[94]。

从人类和人工智能反馈中进行强化学习 (RLHAIF) RLHAIF 整合了人类和人工智能元素以提供监督。Wu et al.^[406] 研究了人工智能协助人类进行书籍摘要的可行性。该方法通过将书籍摘要任务分解为子任务，形成树状结构，从而便于人类监督和评估模型的性能。同时，Saunders et al.^[411] 探讨了利用人工智能协助人类评估模型效能的可行性。他们的发现表明，人工智能模型生成的批评有助于人类发现他们可能错过的缺陷。Bowman et al.^[94] 提出了一个概念验证实验，以展示基于夹心法^[150] 评估可扩展监督技术的前景。当与一个不可靠的 LLM 合作时，结果显示此时人类完成的情况明显优于模型及人类独立完成的情况。Perez et al.^[6] 采用语言模型自动生成数据集，用于评估不同规模的语言模型的行为。作者生成了 154 个经过人类验证的高质量数据集。这些方法展示了使用人工智能协助扩大人类对复杂问题和各种领域的监督的可行性。

讨论 一些努力正在通过用其他组件替换纯人类来增强 RLHF 算法^[15]。鉴于人类反馈的多维性质，很多方法旨在提供由特定规则指导的人类判断。这些规则的例子包括像聊天流畅性^[411] 和隐私保护^[422] 等。Saunders et al.^[411] 将高质量对话的要求分解为代理应遵守的自然语言准则，要求对每个准则进行单独的评估。通过收集针对性的人类评估和在这些数据上训练模型，本文可以获得更有效的基于规则的奖励模型。这种方法显著提高了对话智能体的效能，使它们相比于提示的语言模型更有帮助、更准确、更良性。Carr^[422] 提出了从隐私反馈中进行强化学习 (RLPF)，旨在使语言模型的输出质量与保护隐私相协调。该方法对模型生成的文本进行实时隐私风险评估，并根据这些评估结果调整反馈信号。具体来说，如果生成的文本包含敏感信息，它会产生负面反馈，而高质量、非揭示性文本会得到正面反馈。随着模型的训练，它逐步改善其能力，同时提高文本质量和最小化隐私侵犯。与依赖大规模手动数据标注的传统学习方法相比，这种方法对隐私风险进行了更有效的评估。

$RLxF$ 方法利用将大问题分解为小问题的策略，使得可以利用更有效的工具（如人工智能和软件）快速解决子问题。通过化整为零解决子问题，可以加快主要问题的解决。这些技术可以被视为 IDA 的基本实

例；主要区别在于缺乏持续的迭代过程。然而，证据表明它们有望为超越人类性能的人工智能系统提供反馈^[406]。因此，这些方法可以作为训练更先进人工智能系统的基础技术。

2.4.2 迭代蒸馏扩增 (IDA)

迭代蒸馏扩增 (Iterated Distillation and Amplification, IDA) 引入了一个构建可扩展监督的框架，通过人类和人工智能的迭代协作实现^[332]。该过程从一个初始的智能体开始，记为 $A[0]$ ，其模仿人类 H 的决策过程。 $A[0]$ 经过强大的训练技术 (如足够强的模仿学习)，使其具备接近人类水平的熟练程度 (蒸馏步骤)；然后， H 和多个 $A[0]$ 实例的协作交互形成了一个增强的智能体集合 $A[1]$ 的 (扩增步骤)。这个持续进行的过程在1中有所描述。

Cotra^[235] 区分了广义和狭义的强化学习 (RL) 和逆强化学习 (IRL)。广义 RL 向人工智能系统提供稀疏的奖励信号，并允许智能体自主探索和优化累积的未来奖励。这可能导致超人的新策略的产生，但本文很难完美地指定真正关心的内容。狭义 RL 提供密集且合理的奖励反馈，而不只是结果奖励。这使得人工智能系统更接近人类关心的目标，但限制了其能力。同样，广义 IRL 从人类行为的整个范围中推断深远的长期价值，而狭义 IRL 只推断短期的工具性价值。前者风险更高，而后者在能力上有限。

在 IDA 训练过程中，需要狭义技术来确保每个智能体模仿人类行为。具体来说，可以使用狭义 RL 或模仿学习来训练智能体，使其尽可能遵循人类行为并可控。人类可以利用智能体的计算能力和并行性来制定更有远见的宏观策略——这本质上是对人类内在能力的扩增。在下次迭代中，智能体再次使用狭义技术模仿这个加强的人机系统。这使得从狭义能力到广义能力的过渡可以在保持智能体与人类价值一致的同时进行。随着迭代次数的增加，人机系统变得越来越强大，逐渐接近既高度有能力又与人类价值一致的系统。换句话说，狭义技术被用来确保代理遵循人类价值，而扩增阶段中的扩展人类策略是利用智能体的一种方式，并不扩大智能体自身的学习目标。

AlphaZero 很好地说明了 IDA^[332,423] 的框架形式。该算法从一个简单的策略开始 (如随机落子)，并从其自我对弈的游戏中学习，这是扩增阶段。然后，它使用这些游戏作为训练数据来进行更好的启发式落子，这是蒸馏阶段。这个蒸馏-扩增过程可以重复，以创建一个快速且熟练的围棋人工智能。在这里，对齐和能力的区别是至关重要的^[424]。一个对齐却能力较弱的人工智能试图赢得比赛，但可能无法战胜一般的对手；一个能力强但对齐性差的人工智能实现了赢得比赛之外的某些游戏属性。这一算法的目标是人工智能既有能力又对齐：熟练于游戏，并与赢得比赛的目标对齐。

IDA 的可行性引发了大量的争议^[416]。IDA 运行在一个关键的假设下，即错误不会在迭代过程中连续累积^[15]。因此，在蒸馏和扩增步骤中仍存在技术挑战，需要足够先进且安全的学习技术。此外，尽管原作者将 IDA 比作 AlphaZero 的训练过程^[26]，并已在简单环境中进行了演示^[332]，但其实用性取决于确保 H 能够将复杂任务的部分内容委托给 A ，这类似于领导者协调团队共同完成项目。在实践中，Gato^[417] 展示了 IDA 的关键方面^[425]，可能为通向 AGI 铺平道路。它将多个专家人工智能的能力整合到一个模型中，验证了使用当前的深度学习可以实现 IDA 的蒸馏。虽然尚未完全实现，但 Gato 显示出了扩增的潜力，利用其多样化的技能加速新任务的学习。然而，Gato 缺乏维持对齐属性的安全的扩增或蒸馏方法。为像 Gato 这样的模型设计保留对齐性的 IDA 方法仍是人工智能安全研究的关键方向。总的来说，虽然 Gato 在实现 IDA 方面取得了显著的进步，但进一步的理论进展对于确保 IDA 框架导向安全的 AGI 至关重要

Algorithm 1 Iterative Distillation and Amplification

```

1: procedure IDA( $H$ )
2:    $A \leftarrow$  random initialization
3:   repeat
4:      $B \leftarrow$  AMPLIFY( $H, A$ )
5:      $A \leftarrow$  DISTILL( $B$ ) ▷ Repeat indefinitely
6:   until False
7: end procedure
8: procedure DISTILL(overseer)
   return An AI trained using narrow, robust techniques to perform a task that the overseer already
   understands how to perform.
9: end procedure
10: procedure AMPLIFY(human, AI)
   ▷ Interactive process in which human uses many calls to AI to improve on human’s native
   performance at the relevant tasks.
11: end procedure

```

2.4.3 递归奖励建模 (RRM)

如 §2.2 所讨论的，奖励建模允许本文将系统目标的构建与行为评估分离^[219]。以这种方式，奖励建模为人工智能系统的优化方向提供了指导，能够精细地使系统与人类的意图和价值观对齐，例如对语言模型进行微调以遵循人类指令^[114,273]。此外，奖励建模在前沿人工智能研究方面也被证明是有价值的^[258,330,325]。递归奖励建模 (Recursive Reward Modleing, RRM)^[15,426] 旨在将奖励建模的应用扩展到更复杂的任务。智能体被训练以最大化通过对其扩增版本进行奖励学习所获得的奖励。这种方法不仅受到人类反馈的影响，而且受到模型自身对于奖励构成的评估影响。RRM 的核心思想是递归使用训练得到的智能体 A_{t-1} 来为训练更复杂任务的智能体 A_t 提供反馈。 A_0 通过基本的奖励模型 (从纯人类反馈中学习) 进行训练。如果假设评估比产生问题答案更容易成立，那么奖励模型的迭代过程可以迭代地达到更高的能力从而能够监督更强大的人工智能系统。RRM 的过程在算法 2 中有详细的说明。

例如，我们的目标是训练人工智能 A 来设计一个全面的城市规划。设计一个城市涉及到许多复杂的元素，如交通规划、公共设施以及住宅和商业区的分布。评估一个城市的整体设计是一个重大的挑战，因为许多问题可能只有在长期的实际测试后才会显现出来。为了帮助简化这个过程，我们可能需要一个专门用于交通规划的智能体 B 。然而，交通规划本身就是一个多面的任务。因此，我们进一步需要其他智能体来评估诸如道路宽度、交通流量和公共交通设计等方面。对于每一个子任务，比如测量道路宽度，我们可以训练一个辅助智能体来验证是否满足安全标准，是否考虑了各种交通方式等。通过这样做，我们建立了一个 RRM 过程，其中每个智能体都是在其他智能体评估子任务的帮助下训练的。

这种方法类似于大公司的组织结构^[15]。在城市规划的问题背景下，主要规划团队 (CEO) 负责最后的设计决策。他们的决策是根据交通团队 (部门经理) 的建议来做出的，而交通团队则根据道路宽度团队 (经理) 的输入来给出他们的建议等。每一级的决策都依赖于下一级的反馈，每一个任务都通过奖励建模进行优化。

RRM 面临的挑战可以围绕外部对齐和内部对齐的概念来描述^[426]。外部对齐方面的挑战主要是反馈机制是否可保证学习到的奖励模型在动作模型检测到的分布中准确给予奖励反馈。这个挑战取决于几个因素，包括人类反馈的质量、泛化的难度以及智能体欺骗的可能性。内部对齐方面的挑战集中在人类如何有效地

Algorithm 2 Recursive Reward Modeling

-
- 1: Initialize agent A_0 using reward modeling based on user feedback. ▷ Either preferences or numerical signals.
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Use A_{t-1} to assist users in evaluating outcomes.
 - 4: Train agent A_t based on user-assisted evaluations. ▷ Objective of A_t is generally more complex than that of A_{t-1} .
 - 5: **end for**
-

使用可解释性工具来防止智能体采取欺骗或可能带来灾难性后果的行为。这取决于监督机制的有效性，能否准确验证奖励模型正在进行其他非法优化以及智能体是否保持短视^[235]。

缓解这些挑战的潜在方法^[15]包括提供在线反馈以在训练期间纠正奖励模型^[113]，提供离策略反馈以提供关于危险状态的信息^[217]，扩展现有数据范围如视频^[306]，提供不同级别的层次反馈^[325]，对行动施加侧面约束^[427]，对抗性训练以发现漏洞^[428]，以及对征求反馈的不确定性估计^[419,429]。本质上，RRM 的过程可以理解为 IDA^[332]，其中奖励建模相当于监督或模仿学习的地位。因此，RRM 面临的挑战与 IDA 遇到的挑战密切相似，特别是在防止错误积累方面。此外，奖励建模本身并不一定能够提炼出一个狭窄的模型^[235]，这在权衡对齐程度和性能方面提出了挑战^[426]。

2.4.4 辩论 (Debate)

辩论的基本过程是两个智能体轮流提供答案和陈述，而人类裁判根据辩论过程进行最终结果的评判^[236]，如算法3所示。这是一个零和辩论游戏，智能体在辩论过程中试图识别对方的缺点，同时努力获得人类裁判的更高信任，这是构建可扩展监督的潜在方法。例如，在围棋游戏中，人类裁判可能无法从单个局面辨别优劣方。然而，通过观察游戏的过程和最终结果，这些裁判可以更容易地推断出“谁相对更具优势”。

这种方法的前提依赖于一个关键的假设：为真理辩护通常比为虚假辩护更容易，这给了说真话的辩论者优势。然而，这个假设并不是普遍适用的。例如，在一个复杂的问题中，人类可能无法理解辩论中使用的专业概念。此外，梯度下降的局限性可能会导致本文陷入不希望的循环模式（即在优化一个属性（如诚实和突出缺点）时，模型通常会忽视或减弱另一个属性）^[236]。

值得一提的是，随着大语言模型能力的提升，已有相关工作在大语言模型上应用辩论来提升模型能力^[418,430]。然而，在特定的开放现实世界场景中，辩论可能会面临巨大挑战^[236]。例如，某些问题可能过于复杂、无法被人类理解，或者问题背景过于庞大、无法完全呈现，比如解释一个 1 亿像素的图像或者整个互联网的信息。同样，有些情况下，一个问题的最佳答案可能非常冗长，比如一个需要跨越一百页的回答。为了处理这些问题，智能体可能会首先选择一个回答，然后随着辩论的进行，揭示问题或答案的部分内容。Irving et al.^[236] 对这个过程进行了一个简单环境上的实验。同时，必须考虑到人类评判时间的局限性。在需要与环境互动的场景下，如控制机器人，每个动作可能都需要一个不同的辩论过程。由于时间限制，人类不可能对每个辩论进行判断。为了应对这一挑战，我们可能需要设计相关模型来预测人类的反馈。

另一个需要考虑的是辩论机制的收敛性^[236]。Du et al.^[418] 展示了辩论框架最终趋向于单一回答的内在倾向。同时本文可能需要依靠直觉来衡量收敛的有效性。这意味着需要人类评估者的干预，并要求这些人类评估者具有一定的专业水平，这也会带来一定的挑战。

此外，还有很多讨论从不同视角理解辩论。Ngo^[431] 将辩论视为一种特殊形式的迭代扩增，即利用对抗

Algorithm 3 Debate

-
- 1: Initialize set of questions Q .
 - 2: Initialize two competing agents.
 - 3: Select a question $q \in Q$. ▷ Question is shown to both agents.
 - 4: Agents provide their answers a_0 and a_1 . The agents generate comment answers in response to q .
 - 5: Initialize debate transcript T as an empty list.
 - 6: **for** turn in predefined number of debate turns **do**
 - 7: Agent makes a debate statement s .
 - 8: Append s to T . ▷ Agents take turns and statements are saved in the transcript.
 - 9: **end for**
 - 10: Judge observes (q, a_0, a_1, T) and decides the winning agent.
-

性框架在具体研究问题中建立安全基础。Michaelcohen^[432] 对激励辩论者采用欺骗策略以影响判断过程的负面影响表示担忧。Armstrong^[433], Barnes^[434] 阐述了可能渗透到辩论过程中的各种问题，包括模糊的论点问题、模糊的回应，以及误导性含义的传播。辩论的一方可能会肯定论点中存在任何潜在缺陷的概率足够低，从而主张信任结论，反对者可能会断言在论点中存在可识别缺陷的概率足够高，从而主张不信任结论。Beth Barnes^[435] 引入了交叉检验的概念，以激励辩论者提供更多信息性的回应。在此过程中，辩论者有权选择先前辩论过程中的一个声明进行审查，并获得对方辩论者的回应的副本。整个交流过程都被记录下来，辩论者可以向法官展示相关部分。交叉检验的引入是对不诚实的辩论者利用与他们先前的论点相反的叙述来误导法官的强有力的威慑。

辩论^[236]，迭代蒸馏扩增^[332] 和递归奖励建模^[15] 之间存在一定的相似性。这些方法可以从一个基本原则来理解：评估可以比完成任务更简单²⁸。因此，利用 AI 系统的评估能力可以帮助人类实现可扩展监督。然而这些方法面临的挑战，尤其是在减少错误积累方面，也是类似的。

2.4.5 合作逆强化学习 (CIRL)

几乎先前所有的方法都将从反馈中学习视为一个与推理和控制分离的过程，并且经常隐含地将反馈提供者视为存在于环境之外的实体 – 事实上，像操纵^[84] 和奖励篡改^[224] 这样的失败模式就是当被假定为环境之外的反馈机制变成环境的一部分，从而受到 AI 系统影响时发生的。合作逆强化学习 (Cooperative Inverse Reinforcement Learning, CIRL) 的框架统一了控制和从反馈中学习，并将人类反馈提供者建模为在同一环境中的同伴代理。它将提供反馈的人类和 AI 系统置于合作而非对抗的位置，试图通过消除 AI 系统欺骗监督的动力，而非通过加强监督来解决可扩展监督问题^[421]。在 CIRL 算法中，AI 系统与人类合作以实现人类的真正目标，而不是单方面地优化人类的偏好。

合作逆强化学习的动机和总体思路 许多对齐失败的模式，例如奖励破解^[220,155]，欺骗^[10]，和操纵^[84]，都是 AI 系统对错误规范的目标进行“自信地”优化的结果^[16]。在训练和部署过程中，指定的目标（如奖励函数）对 AI 系统来说起着无可挑战的真理的角色，人类的反馈一定程度上只有在目标位置上才被尊重，这意味着它可以被篡改^[224] 或者被操纵^[84]。

合作逆强化学习^[419,296,421] 试图通过以下方式来解决上述问题 (1) 让 AI 系统明确地对其奖励函数保持不确定性；(2) 让人类提供关于真实奖励函数是什么的唯一信息。不确定性使 AI 系统倾向于听从人类的意

²⁸关于这一点的讨论也可以在涉及这些方法的文献中找到。

见并驱使它去确定人类真正想要什么。具体来说，它将整个任务模型化为一个包含两个玩家的合作博弈，其中人类玩家 H 和智能体玩家 R 共享一个公共的奖励函数 $r(\cdot)$ 。更重要的是，奖励函数和奖励信号对 R 来说是不可见的（实际上并没有被训练机制明确地计算出来），只能通过一个类似于 IRL 的过程从 H 的行为中 R 推断出来（包括通过询问和与 H 交互）。这一设定被称为 CIRL^[419]，协助博弈^[420]，和协助 POMDP^[421]。

简单来说，AI 系统将人类的真实目标 $r(\cdot)$ 作为自己的目标（尽管 $r(\cdot)$ 的值并不确定），并通过观察和与人类交互来不断尝试弄清楚 r 。这可能消除了例如操纵这类行为的动力，因为操纵人类的行为只会污染一个信息源，而不会影响 r 。

CIRL 的构建 Hadfield-Menell et al.^[419] 在经典的多智能体 MDP 的基础上，定义了 CIRL 的设定（本文用 M 表示），并给出了如下的 M 定义。

$$M = \langle S, \{\mathcal{A}^H, \mathcal{A}^R\}, T, \gamma, r, \Theta, P_0 \rangle$$

在上面的等式中， S 和 $\{\mathcal{A}^H, \mathcal{A}^R\}$ 分别是世界状态和动作空间， $T: S \times \mathcal{A}^H \times \mathcal{A}^R \rightarrow \Delta(S)$ 是状态转移函数， γ 是折扣率。到此为止，这个定义与标准的多智能体 MDP 的定义相同。然而，剩下的元素具有关键的区别：奖励函数是参数化的，其参数可以通过分布进行建模。 Θ 是参数 θ 的取值空间； $r: S \times \mathcal{A}^H \times \mathcal{A}^R \times \Theta \rightarrow \mathbb{R}$ 是共享的奖励函数， $P_0 \in \Delta(S \times \Theta)$ 是初始状态和奖励函数参数的联合分布。这种参数化方法允许 R 明确地建模并推理真实奖励函数的信念。使用 Nayyar et al.^[436] 的技术，任何 CIRL 设定都可以简化为等效的单智能体 POMDP，从而证明了相对易处理的最优策略的存在^[419]。

CIRL 研究的显著方向 尽管一些人强调了 H 教导 R 的重要性^[437]，但有些研究^[421] 质疑了对博弈均衡和联合策略（包括 H 的教导行为）的重视，而更专注于 R 对 H 策略的最优响应，因为人类总是会采取最优的联合策略的假设是不现实的。更具体地说，Shah et al.^[421] 考虑了策略条件信念 $B: \Pi^R \rightarrow \Delta(\Pi^H)$ ，它指定了 H 对任何 R 策略的策略响应分布，目标是在给定 B 的情况下找到 R 的最优策略。在这里， B 本质上是一种人类建模，获取一个鲁棒准确的人类模型作为 B ^[438-439] 是一个挑战。另一方面，Hadfield-Menell et al.^[296] 和 He et al.^[440] 研究了手动指定不完美奖励函数作为 H 传达真实奖励函数信息的一种方式。这包括了 R 的工作（即使 R 能够根据不完美的指定推断真实的奖励函数）^[296]，也包括了 H 的工作（即开发算法工具来帮助 H 做出更鲁棒的指定，更好地传达真实的奖励函数）^[440]。

还有一些工作将 CIRL 和协助博弈扩展到智能体需要服务于多个人类的多智能体环境^[420]。这对应于 Critch et al.^[61] 中的多/单一委托设定，其中人类的不同目标产生了挑战，并需要使用社会选择方法。

3 在分布偏移下学习

可靠的人工智能系统的构建在很大程度上依赖于它们适应多样化数据分布的能力。训练数据和训练环境往往是实际部署场景的不完美近似，这导致它们可能缺少某些关键元素，如对抗压力^[441]（例如，在监督学习系统中的高斯噪声^[442]，在自动驾驶系统中的影子攻击^[443]），多智能体交互情景^[61,131]，人类监督者无法有效评估的复杂任务^[15]，²⁹以及可以被操控的奖励机制^[121]。从训练分布到测试分布（或环境）的这种差异转变被称为分布偏移^[121-122]。

²⁹这可能导致欺骗行为的出现^[17]。有关详细信息，请参阅 §3.1 中关于目标错误泛化的段落。

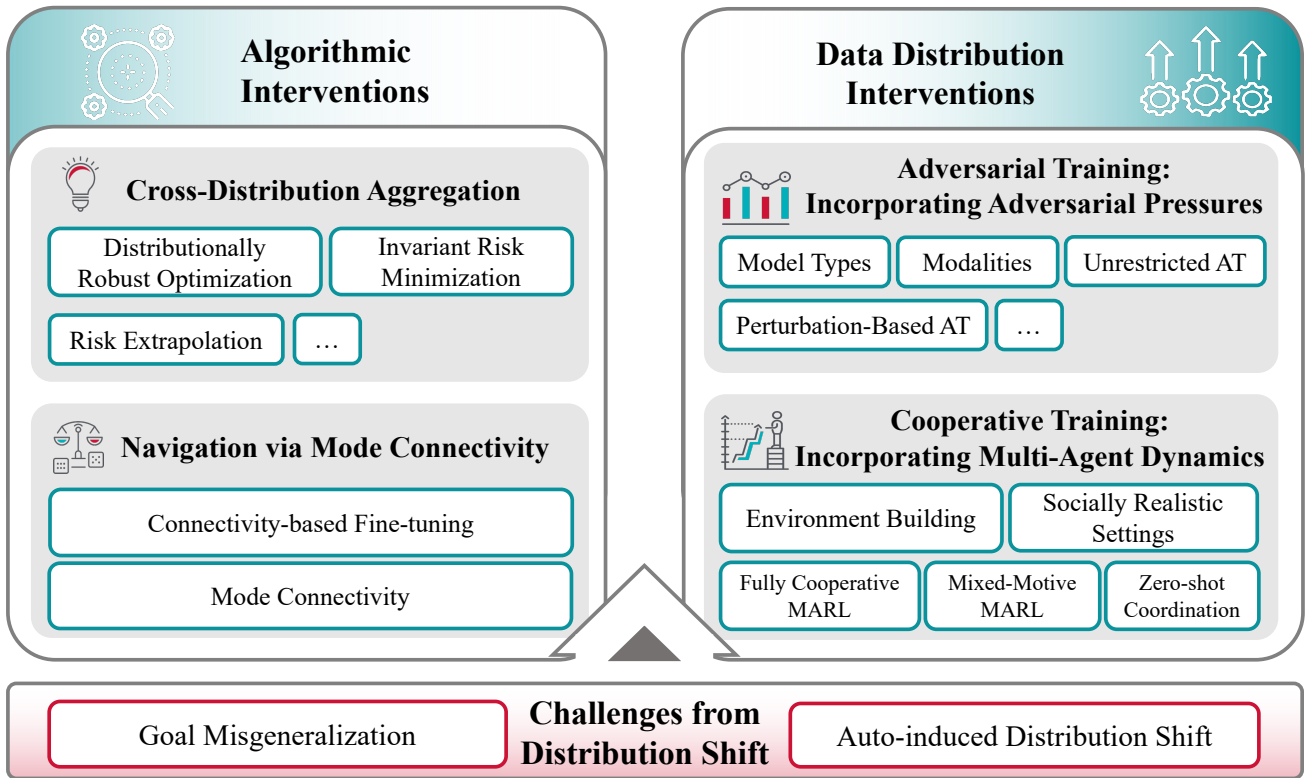


Fig. 6 在分布偏移下学习的框架。分布偏移带来的主要挑战是目标错误泛化和自诱发分布偏移 (§3.1)。本文还介绍了两种应对分布偏移的方法：算法干预 (§3.2)，在训练过程中通过算法技巧引导优化；和数据分布干预 (§3.3)，通过在训练过程中引入特定的现实元素从而有针对性地扩展训练分布。

因此，在训练分布下对齐的人工智能系统（即追求与人类意图一致的目标）可能在部署（或测试）分布下无法保持其对齐性，进而在部署后导致严重的对齐问题。这种可能的失败引发了关于在数据分布之间保持对齐属性（即遵守人类意图和价值）的研究。

从对齐的角度来看，我们更关心人工智能系统是否追求不对齐和有害的目标，而不是本身的能力强弱。因此，强调对齐属性意味着我们关注在分布之间的目标泛化，而不是能力泛化^[124,19]。

本节主要讨论在分布偏移下学习时保持对齐属性的问题。我们首先介绍分布偏移带来的对齐挑战 (§3.1)。然后，我们深入讨论解决分布偏移的方法，并特别讨论两类路径：(1) 算法干预 (§3.2)：旨在在训练过程中引导优化；(2) 数据分布干预 (§3.3)：旨在通过在训练过程中引入特定元素或分布来扩展训练分布，相关技术包括对抗训练^[444,130,445] 和合作训练^[131-132] (§3.3.2) 等。在分布偏移下学习的框架如图6所示。

3.1 分布偏移带来的挑战

在介绍具体技术之前，我们首先分析为什么在对齐中，在分布偏移下学习是一项主要挑战——更具体地说，是在分布偏移下保持对齐属性（即坚持人类的意图和价值）。本文主要关注两个关于分布偏移问题的对齐挑战，即目标错误泛化 (Goal Misgeneralization, GMG)^[17] 和自诱发分布偏移 (Auto-induced Distribution Shift, ADS)^[121]。

人工智能系统的训练使它们学会在输入分布下“高效”追求训练奖励、降低相应训练损失。然而，这种高效可能无法推广到输入分布发生质的变化（即分布偏移）的情况。这些变化包括如对抗压力^[441]，多智能

体互动^[61,131]，以及人类监督者无法有效评估的复杂任务^[17]，和可以被操纵的奖励机制^[121]等。

这里需要区分两种不同的失败模式：目标错误泛化^[17]：其中训练分布和测试分布是假设固定的，我们更关注系统追求的目标在分布偏移情况下的变化情况；而在自诱发分布偏移^[121]中，人工智能系统通过自身的行为改变数据分布以追求更高奖励，在对 ADS 的讨论中更关注系统对于分布的能动影响。

目标错误泛化 这种挑战指的是人工智能系统在训练分布中表现完美，但在分布外环境中，训练分布中学习到的能力无法泛化，这使得 AI 可能会追求与人类愿望不符的目标^[17,124]。目标错误泛化需要与其他形式的错误泛化（如能力错误泛化）区分开来：能力错误泛化通常指在分布外测试环境中，智能体性能较差，甚至无法正常完成目标；相反，具有目标泛化错误的智能体在 OOD 环境中有能力追求一个不希望的目标。³⁰

一个相关的例子是虚假关联（或捷径特征）^[447,124]。例如，在一个图像分类数据集中，绿色草地是标签牛的一个高度预测特征。然而，需要注意的是，这个特征需要在各种数据分布中更加一致和可靠^[448]（即分类器可能根据草地特征来预测图片为牛，显然草地并不是真实的标签对应特征，因此在其他数据分布中并不可靠）。此外，IL 中的因果混淆（即对顾问和环境之间交互的因果结构的无知）可能导致目标泛化错误^[449,117]。

目标错误泛化的一个主要危险在于，“优化人类真正想要的”和“优化人类所赞同的”之间的无法区分³¹；后者可能包括欺骗或操纵人类评估者^[84]以获得他们的赞同。例如，Amodei et al.^[266]发现，在一个机器人手目标是抓住一个小球的任务中，机器人手通过在镜头前的视差来伪造动作，使其看起来好像已经抓住了球，但实际上并没有这样做。这种行为欺骗了人类标注者，使他们认为任务已经完成，故给予相应的高奖励。

当一个 AI 系统通过人类反馈进行训练或微调时，“人类真正想要的”和“人类所赞同的”这两个目标是无法区分的，因为使用这两个目标的 AI 系统在训练中都表现得很好，我们并不清楚 AI 系统最终会学习到哪一个目标。事实上，在训练过程中，人类评估者可能被欺骗或操纵，这意味着 AI 系统可能更强烈地被激励去优化“人类所赞同的”，而不是“人类真正想要的”。这种现象在很多系统中都被观察到，包括推荐系统^[126-127]，大语言模型^[6]和 RL 系统^[266]。

最后，与目标泛化错误密切相关的一个失败模式是内优化的问题^[20]，其中模型在推理过程中使用学习到的模型权重进行自身的优化^[20,450]，而这种优化的目标并未与模型的训练目标对齐。实证研究发现，Transformers 在前向传播过程中使用内优化来提高性能，这为该假设提供了证据^[451]。

自诱发分布偏移 (ADS) 在训练 AI 系统时，研究者常常只考虑系统本身的优点和缺点，而忽视了这些系统对环境的影响。过去的研究常常假设数据是独立同分布的^[452]，忽视了算法对数据分布的影响。然而，Krueger et al.^[121]提出，在现实中，AI 系统在决策和执行过程中可能会影响环境，从而改变环境生成的数据的分布。他们将这种问题称为自诱发分布偏移。一个现实生活中的例子是推荐系统，其中由推荐算法选择的内容可能会改变用户的偏好和行为，导致用户分布的偏移。这种分布的偏移反过来又会进一步影响推荐算法的输出^[453]。随着 AI 系统对世界的影响越来越大，我们也需要考虑 AI 系统融入人类社会后可能对整个社会的数据分布产生的进一步影响。

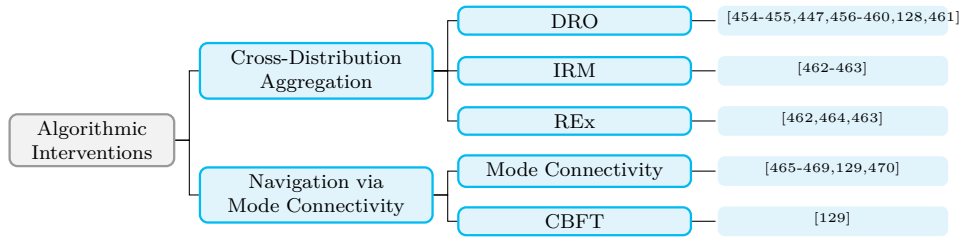


Fig. 7 与算法干预相关的概念和文献。根节点代表的是旨在在训练过程中引导优化的算法干预。主要分支代表了两种主要方法，即跨分布聚合（旨在在训练过程中最小不同分布上的风险，以便根据不变关系而非偶然特征找到预测器）和通过模式连接进行导航（旨在基于模式连通性进行微调以提高模型的泛化性能）。进一步的子分支列出了一些重要的技术，如分布鲁棒优化 (DRO)，不变风险最小化 (IRM)，风险外推 (REx) 和基于模式连接的微调 (CBFT)。

3.2 算法干预

在阐述算法干预方法时，我们首先概述两类在训练过程中引导各种分布优化以缓解分布偏移的方法，即跨分布聚合 (§3.2.1) 和通过模式连通性导航 (§3.2.2)。

在第一部分，本文涵盖了从经验风险最小化 (ERM) [462] 到风险外推 (REx) [128] 的各种方法，其中后者是为了缓解模型依赖于虚假关联而产生的问题。在第二部分，我们介绍了基于连通性的微调，它在训练过程中通过基于模式连通性 [129] 指导微调，从而引导损失景观中最小化器之间的转变，使得模型仅需改变部分参数便可从“依赖虚假特征预测”转至“依赖正确关联预测”。

3.2.1 跨分布聚合

分布偏移的主要原因之一是模型预测中的虚假关联 [447]。通过将不同领域 (或不同分布) 的学习信息集成到优化目标中，我们期望模型学习到标签和状态之间的真实信息和不变关系。在以下段落中，我们首先以 ERM 为背景介绍，然后介绍一些直接学习如何通过集成不同分布的损失景观来解决分布偏移的方法。

经验风险最小化 (ERM) 考虑一个模型被开发出来后，通过特征有效地识别对象的场景。优化目标可以表达为：

$$R(\mathbf{w}) = \int L(y, f(x, \mathbf{w})) dP(x, y)$$

其中， $L(y, f(x, \mathbf{w}))$ 表示数据标签 y 和模型输出 $f(x, \mathbf{w})$ 之间的损失，而 $P(x, y)$ 表示目标数据分布 [462]。

然而，数据集和真实世界之间往往存在偏差，这意味着从数据集中学习到的特征可能并不一定是本文希望模型获取的特征。ERM 是统计方法中用来优化这种偏差的策略。它的基本假设是，鉴于无法获取真实世界的目标数据分布，数据集中的经验数据应该理想地接近这个未知的目标分布 [462,471]。在这种情况下，优化目标函数被重新定义为：

$$E(\mathbf{w}) = \frac{1}{l} \sum_{i=1}^l L(y_i, f(x_i, \mathbf{w}))$$

³⁰更多目标泛化错误的例子，请参阅 [446]。

³¹这里的人类所赞同的是指来自人类顾问或环境的高奖励反馈。然而，AI 系统可能会故意遵循人类的偏好或欺骗以获得人类的高奖励，但实际上并没有真正学习到预期的目标 (即人类真正想要的)。

其中 l 可以是一个训练分布中的不同示例，或者是不同的训练分布。

最小化上述目标函数使模型能够在不同分布中学习特征和标签间的不变联系。ERM 同样基于一个基本假设是即数据是从目标数据分布中抽样的。然而，如果源分布（或训练分布）和目标分布之间存在显著的差异，仍然可能出现严重的泛化问题^[472]。

分布鲁棒优化 (DRO) 许多研究认为，对分布转移的敏感性通常源于模型依赖于与核心概念无关的假相关或捷径特征^[447,457]。例如，模型可能会基于背景特征进行判断，而不是使用图像中的正确特征^[447,456]。在先前研究的基础上^[454-455,128]，分布外泛化可以被表述如下：

$$r_{\mathcal{D}}^{\text{OD}}(\theta) = \max_{e \in \mathcal{D}} r_e(\theta)$$

这种优化旨在通过减小风险函数集 $\{r_e | e \in \mathcal{D}\}$ 中的最大值，提高在一个扰动集（记为 \mathcal{D} ）中的最差情况性能。在分布鲁棒优化 (*Distributionally Robustness Optimization, DRO*)^[461] 中，扰动集覆盖了不同域训练分布的混合，通过最小化上述目标函数，我们期望模型能够找到不同训练分布之间标签和特征之间的不变关系。然而，需要注意的是，直接将 DRO 应用到超参数化的神经网络可能会导致次优的结果^[460]。因此，结合增加的正则化技术，如 l_2 惩罚^[473] 或早停^[474]，可以显著提高泛化性能。关于 DRO 的更多细节，请参见 Rahimian et al.^[475], Sagawa et al.^[460], Chen et al.^[476], Lin et al.^[477]。

不变风险最小化 (IRM) Arjovsky et al.^[463] 引入了一种创新的学习范式，用于估计在各种训练环境中的非线性、不变、因果预测器，从而促进鲁棒的 OOD 泛化。IRM 旨在训练一个在各种环境中都有稳健性能的预测模型，同时减少依赖于假相关特征的可能性。IRM 可以被认为是不变因果预测 (ICP)^[455] 的扩展，它涉及假设检验，以确定在每个特定环境中导致结果的直接因果特征，而不是间接特征。IRM 进一步将 ICP 扩展到高维输入数据的场景，其中变量可能缺乏明确的因果意义。IRM 的基本思想是，当面临许多能够实现低经验损失的函数时，选择在所有环境中都表现良好的函数，更有可能得到一个基于因果特征而不是假相关特征的预测器^[448]。

风险外推 (REx) REx 的基本形式涉及在外推域的扰动集上进行鲁棒优化 (MM-REx)，并对训练风险的方差施加额外的惩罚 (V-REx)^[128]。通过减少训练风险并增加训练风险的相似性，REx 迫使模型学习不同域分布中的不变关系。

放大训练域之间的分布变化可以减小风险变化，从而强制实现风险的平等。以 CMNIST^[463] 为例，尽管建立颜色和标签之间的联系比连接数字和标签更直接，但增加颜色的多样性可以打乱这种虚假关联（或捷径特征），并帮助模型学习数字和标签之间真正的不变关系。根据以前的研究^[462,464,128]，REx 可以被表述如下：首先，风险函数可以定义如下：

$$r_e(\theta) \doteq \mathbb{E}_{(x,y) \sim P_e(X,Y)} L(f_{\theta}(x), y)$$

其中 $L(\cdot)$ 代表一个固定的损失函数，不同的训练域或环境可以被表述为 $P_e(X, Y)$ 分布。接下来，MM-REx

项可以被建模为：

$$r_{\text{MM-REx}}(\boldsymbol{\theta}) = (1 - m\lambda_{\min}) \max_e r_e(\boldsymbol{\theta}) + \lambda_{\min} \sum_{e=1}^n r_e(\boldsymbol{\theta})$$

其中 n 代表不同分布或域的数量， λ_{\min} 控制风险外推的程度。然后，V-REx 项可以被建模为：

$$r_{\text{V-REx}}(\boldsymbol{\theta}) = \alpha \text{Var}(\{r_1(\boldsymbol{\theta}), \dots, r_n(\boldsymbol{\theta})\}) + \sum_{e=1}^n r_e(\boldsymbol{\theta})$$

其中 $\alpha \geq 0$ 控制风险降低和强制风险近似之间的权衡。

在 MM-REx 项中， λ_{\min} 可以设置为接近 $-\infty$ ；因此，特定域的损失可能会很高，这意味着模型可能会学习假相关。最小化 MM-REx 和 V-REx 可以降低训练风险并增加训练风险的相似性，鼓励模型学习不变关系。此外，REx 在实验设置中表现出了显著的潜力^[128]，特别是在因果识别方面，使其成为实现鲁棒泛化的有力方法。

3.2.2 模式连接指引

在上述关于跨分布聚合的讨论之后，我们在本节中介绍模式连通性作为引入。然后，我们主要讨论基于模式连通性的微调 (CBFT) 方法^[129]，说明模式连接如何通过改变少量参数引导模型基于不变关系进行预测，而不是假相关。

模式连通性 模式连通性是指在损失函数空间内可以确定一条直接的路径，连接两个或更多不同的局部最小值或模式的现象^[465-466]。根据以前的研究^[468-469,129]，可以如下定义：

模型在数据集 \mathcal{D} 上的损失表示为 $\mathcal{L}(f(\mathcal{D}; \boldsymbol{\theta}))$ ，其中 $\boldsymbol{\theta}$ 表示模型的最优参数， $f(\mathcal{D}; \boldsymbol{\theta})$ 表示在数据集 \mathcal{D} 上训练的模型。如果 $\mathcal{L}(f(\mathcal{D}; \boldsymbol{\theta})) < \epsilon$ ，本文定义 $\boldsymbol{\theta}$ 为这个数据集上的损失的最小化者，其中 ϵ 是一个小的标量值。

通过在数据集 \mathcal{D} 上的训练得到的局部最优器 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$ 被认为是模式连接的，如果存在一个从 $\boldsymbol{\theta}_1$ 到 $\boldsymbol{\theta}_2$ 的连续路径 γ ，使得当 $\boldsymbol{\theta}_0$ 沿着这个路径 γ 变化时，以下条件始终得到满足：

$$\mathcal{L}(f(\mathcal{D}; \boldsymbol{\theta}_0)) \leq t \cdot \mathcal{L}(f(\mathcal{D}; \boldsymbol{\theta}_1)) + (1 - t) \cdot \mathcal{L}(f(\mathcal{D}; \boldsymbol{\theta}_2)), \quad \forall t \in [0, 1].$$

本质上，模式连通性涉及在参数空间中始终找到最小化器之间的连接路径，遍历低损失区域而不深入高损失区域。这意味着即使在参数空间内对模型的参数进行微小的调整，模型的性能也可以保持相对稳定^[465]。这个概念为设计更有效的优化算法奠定了基础，使模型能够在不同的任务之间共享知识和经验，提高模型的泛化能力。

此外，我们可以定义，两个模型如果使用输入的相同属性进行预测，那么它们在机制上具有相似性。一些研究已经证明，线性连接的缺乏意味着预测机制上的不相似性，这表明简单的微调可能不足以消除在预训练阶段学习到的假属性^[129,470]。然而，通过微调处理非线性连接的区域是有希望的，从而有效地修改模型的机制，解决 OOD 泛化失败的问题。

基于连通性的微调 (CBFT) 如上文所述, 最近的研究表明, 两个模型之间缺乏线性连通性意味着它们在基本预测机制上存在显著的不同。Lubana et al.^[129] 发现, 当模型在相似的数据上进行训练时, 它们往往会发展出相似的推理机制。这可能是模型出现偏见的一个重要原因, 例如, 模型在进行图像分类时依赖于图像的背景信息, 而不是图像中描绘的对象。如果在微调过程中没有调整这种模型机制, 模型可能会依赖于这些错误的属性。为了克服这个问题, 他们提出了一种有效的改变模型机制的策略, 该策略旨在最小化以下损失:

$$\mathcal{L}_{\text{CBFT}} = \mathcal{L}_{\text{CE}}(f(\mathcal{D}_{\text{NC}}; \theta), y) + \mathcal{L}_{\text{B}} + \frac{1}{K} \mathcal{L}_{\text{I}}$$

其中, 原始训练数据集表示为 \mathcal{D} , 我们假设可以获得一个没有假属性 C 的最小数据集, 表示为 \mathcal{D}_{NC} 。

除了 \mathcal{L}_{CE} 表示模型预测 $f(\mathcal{D}_{\text{NC}}; \theta)$ 和真实标签 y 之间的交叉熵损失外, CBFT 有两个主要目标: (1) 第一个目标涉及通过在损失景观中重新定位模型来修改模型的底层机制, 打破与当前最小化器的任何线性连接。这是通过最大化 \mathcal{L}_{B} 来实现的, 称为障碍损失。(2) 第二个目标涉及减轻对原始训练数据集中假属性的依赖。这是通过优化 \mathcal{L}_{I} 来实现的, 使得无需 C 也能发现不变的关系。CBFT 对于将机制从通过假特征预测目标转变为通过真实特征预测目标具有潜力, 只需改变模型的部分参数。

3.3 数据分布干预

除了算法优化, 将训练数据的分布扩展到包含现实世界元素的方法也可以减少训练和部署分布之间的差异。在本节中, 我们特别关注对抗压力和多智能体动态的引入。

3.3.1 对抗训练

AI 系统可能会因缺乏对抗鲁棒性而受到影响, 这意味着设计使它们失败的某些输入会导致模型表现不佳^[478]。这已经在图像^[479]和文本^[72-73]中得到证实, 以及图像的语义特征更改^[447,480-482]和文本^[483], 甚至完全从头生成的示例^[69,484,445,485]。这些失败模式在红队测试部分 (§4.1.3) 中有所涵盖。值得注意的是, 对抗鲁棒性对于控制先进 AI 系统训练的奖励模型尤其重要, 因为梯度下降优化过程可能会出现被奖励模型过度探索的漏洞, 这是一种被称为奖励模型过度优化的现象, 已经得到了实验验证^[233,486]。

我们认为对抗鲁棒性是由 AI 系统的训练分布 (其中训练输入不是对抗性构造的) 和测试分布 (其中示例可以是对抗性构造的) 之间的不匹配部分引起的分布偏移失败的一个案例。对抗训练的方法^[444,130,445]通过各种方式^[130]将对抗样本引入到训练输入中, 从而扩展训练分布并缩小分布差异。

对抗训练和对抗攻击相似, 首先开始于图像分类的设定下^[472,487], 但后来扩展到了广泛的设置。除了视觉模型外, 还为语言模型^[488-489,445]、视觉-语言模型^[490-491]等提出了对抗训练算法。就模型类型而言, 对抗训练已经应用于分类模型^[130]、生成模型^[445]和 RL 代理^[492-497]。

对抗训练主要有两种类型: 基于扰动的对抗训练和无限制对抗训练。

- **基于扰动的对抗训练。**基于扰动的对抗训练基于基于扰动的对抗攻击 (参见 §4.1.3), 它将对抗性扰动的示例 (即对正常数据输入进行的小改动, 这些改动旨在降低模型性能) 引入到训练中^[487]。这种方法^[130]包括将一个正则化项添加到损失函数中以评估模型在基于梯度的扰动输入上的性能^[487], 无监督^[498]或自监督^[499]的方法, 以及各种补充技术, 如引入逐步加强训练过程中对抗压力的课程学习^[500]。
- **无限制对抗训练。**无限制的对抗训练基于无限制对抗攻击 (参见 §4.1.3), 它将基于扰动的对抗训练推广到包括任何可以欺骗模型的对抗样本, 不一定是通过向另一个示例添加少量噪声获得的。这包括生

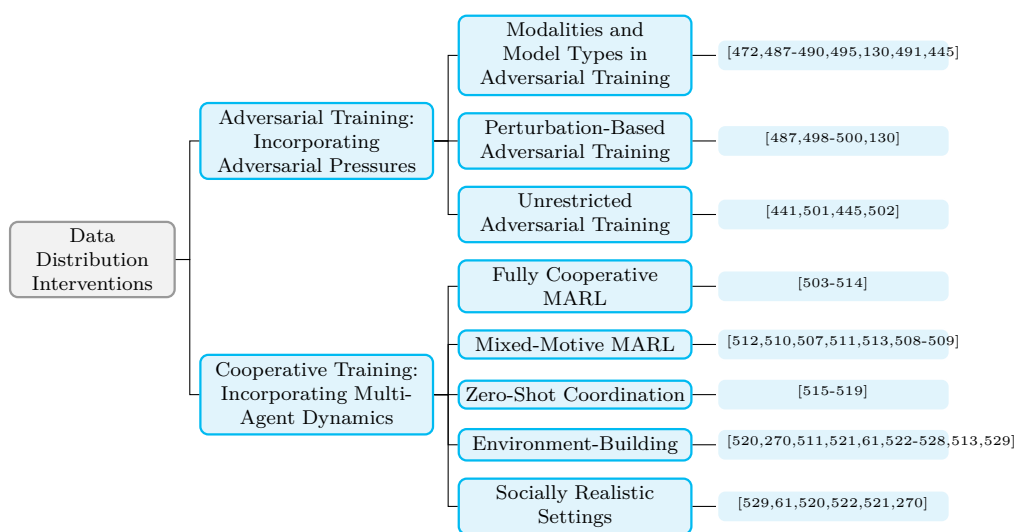


Fig. 8 与数据分布干预相关的概念和文献。根节点代表的是在训练过程中尝试结合多个分布的数据分布干预，例如对抗样本和多智能体交互。主要分支代表了一些有前景的方法，即包含对抗性压力的对抗性训练和包含多智能体动态的协同训练。进一步的子分支列出了关键技术，如基于扰动的和无限制的对抗性训练，以及协同方法也包括环境搭建、社会模拟、无准备协调和其他基于多智能体强化学习 (MARL) 的技术。

成对抗训练，该方法使用生成模型从头开始生成任意的对抗输入^[441]，以及将句法或语义修改的对抗样本添加到训练输入中^[445,501]，这消除了对模型非对抗性能的负面影响。Zhang et al.^[502] 试图统一无限制和基于扰动的对抗训练，尽管基于一些非平凡的假设。大多数关于无限制对抗攻击的工作也适用于无限制对抗训练 (参见 §4.1.3的概述)，并构成了无限制对抗训练方法的重要部分。

3.3.2 合作训练

合作性 AI^[131-132] 旨在解决 AI 系统的非合作和集体有害行为 (见 §1.3)。AI 系统缺乏合作能力可以被视为在分布偏移下的失败形式—系统在与现实世界质量上不同的单智能体环境中进行训练，而现实世界可能是大规模多智能体的。这种差异实际上是数据分布的差异，因为环境中其他智能体的存在质量上改变了环境状态转移动态，导致观察和奖励的联合分布发生变化。本文通过扩展训练分布来包括多智能体交互，即合作训练，来解决这个问题。

本文引入了合作性 AI 的一个分支 (我们称之为合作训练)，该分支专注于特定形式的多智能体强化学习 (MARL) 训练，并补充了在 §4.3.1中的正式博弈论方法。协同训练的 MARL 分支倾向于强调 AI 系统的协调能力 (例如，机器人足球队协调^[527])，而不是合作的动机 (例如，应对囚徒困境等情景^[268])，这是博弈论分支的重点。在这里，我们只涵盖了 MARL 分支，因为它与扩展训练数据分布有关。

MARL 领域传统上被划分为三个分支，即完全合作 (所有智能体共享同一奖励函数)、完全竞争 (基础奖励构成零和博弈) 和混合动机设置 (奖励激励既不完全合作也不完全竞争，对应于一般和博弈)^[512]。其中，完全合作和混合动机设置和合作性 AI 最为相关，后者尤其需要被强调，因为它先前相对被忽视^[131]。本文还涵盖了其他研究前沿，包括无准备协调^[516-517]、环境构建^[523] 和社会现实设置^[529]。

- **完全合作 MARL**。完全合作的 MARL 设置的特点是所有智能体共享奖励函数^[512]。这种统一使我们

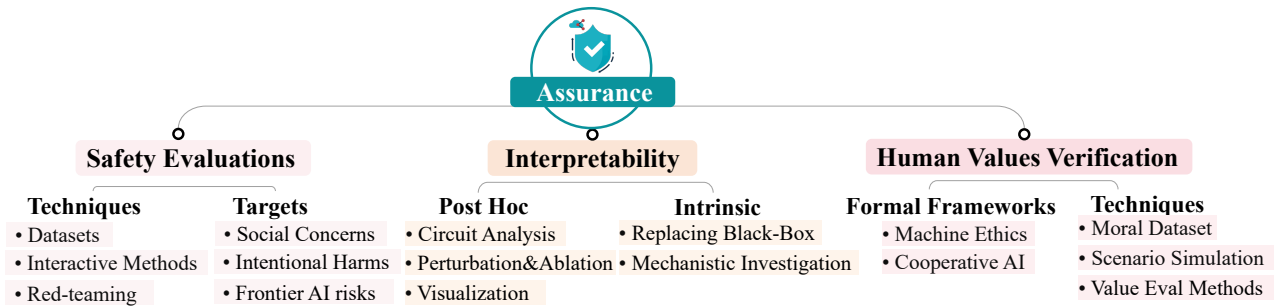


Fig. 9 对齐保证的研究方向、技术和应用的组织框架。本文将这一部分划分为三个部分：安全测评—评估 AI 系统的安全性，即减轻 AI 系统引起的事故和有害事件；可解释性—使 AI 系统及其决策过程更易于人类理解；人类价值验证—验证 AI 系统是否能够遵守社会和道德规范。该图还展示了这些部分之间复杂的逻辑关系。

可以完全忽视合作激励的问题（因为所有激励都完全一致），而只关注通过协调有效地实现共享目标。常用的方法^[514]在中心性上呈现出一个谱系—从纯粹独立训练的基线解决方案^[503]，到用分散通信补充独立训练的方法^[505]，以及价值因子化，它分解全局奖励并确定每个单独智能体的贡献^[504,506]。

- **混合动机 MARL**。混合动机的 MARL 设置的特点是合作和竞争激励的混合—智能体的奖励不是相同的，但也不是零和的^[512]。这包括团队对抗的游戏环境^[510]和更精妙的环境设置，如谈判^[511,513]。混合动机 MARL 的技术，包括使用类似于 IRL 的方法从人类交互中学习^[508]，使智能体之间的交流变得更加具有战略性和可选择性^[509]，并通过让评论家访问全局信息来调整 Actor-Critic 方法^[507]。
- **无准备协调**的目标是使 AI 系统能够与其他智能体（包括人类）有效地协调，而无需一起接受训练或以其他方式专门设计以与这些智能体协调^[516-517]—完全陌生的人类仍然可以有效地合作，本文希望 AI 系统也能做到这一点。早期的工作是以特设协调的名义发布的，涵盖了评估^[530]，博弈论和统计方法^[531]，以及人类建模^[532]。最近的进展包括他人游戏^[516]，该游戏随机化训练伙伴策略的某些方面以实现鲁棒性，³²引入多级递归推理^[518]，以及离信念学习^[519]，后者通过将合作伙伴的过去行为解释为非勾结政策来消除自我游戏中的任意约定。
- **环境搭建**。游戏环境一直是合作训练的热门设置，例如 Hanabi^[525]，Diplomacy^[511,513]和足球^[527]。基于经典多智能体困境的博弈论模型也一直是环境选择的热门^[524,528]。此外，Melting Pot^[523,526]，一个多智能体环境的框架和套件，专门为合作 AI 研究设计。还有一些关于无监督环境设计的研究，该设计旨在部分自动化环境构建过程^[533-534]。
- **社会模拟**。一些工作提出，合作 AI 研究应更多地关注社会现实环境^[529]，这些环境往往是大规模多智能体（包括 AI 和人类）并且在智能体的组成和交互方式上都非常多样。这个愿景的含义^[61]包括但不限于，构建更真实和开放式的环境^[520,535-536,522]，扩大 MARL 的规模^[521,529]，以及引入新的控制手段，如社会制度和规范^[270]。

4 对齐保证

在人工智能系统实际训练和部署之后，进行对齐保证是至关重要的。这一过程涉及到对人工智能系统实用性的测量和评估，确保其能够达到预期的效果^[537]。对齐保证可以分为三个主要部分。首先，安全测评是基础，它涉及评估人工智能系统在执行任务时最小化事故的能力。其次，可解释性是必要的，以确保人类能够理解人工智能系统的决策过程，这有助于保障系统的安全性和互操作性。最后，人类价值验证对于确保人工智能系统能够符合人类的价值观、道德和社会规范至关重要，这是人工智能融入人类社会的高级需求（如图9所示）。

4.1 安全测评

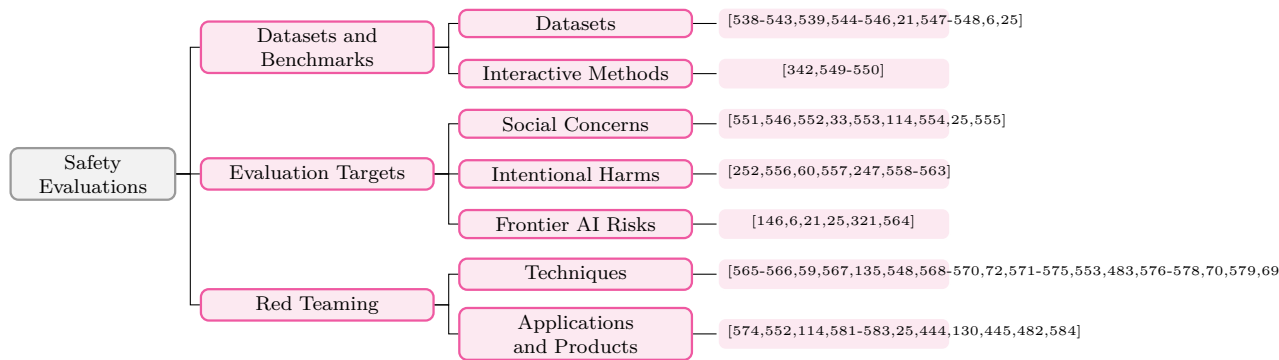


Fig. 10 图中表示与安全评估相关的基本概念、逻辑结构和相关文献。安全评估的核心目标是识别和测量由于设计缺陷导致的事故，以及偏离预期设计目的有害事件。我们可以将这个过程的想象成一棵树，树的根代表安全评估的核心目标。主要的分支则代表安全评估的主要组成部分：数据集和基准、评估目标，以及红队测试。每个分支下的次级分支进一步详细列出了相关的关键研究。这张图为我们提供了一个关于如何测量人工智能系统安全对齐程度的研究方向和具体技术的全面概览。

安全性是指减轻人工智能系统设计缺陷引起的事故和防止偏离人工智能系统预期设计目的有害事件^[143]。事实上，安全性是所有工程领域的共享需求^[585]。而由于人工智能系统的特性^[586]，在构建人工智能系统时，安全性尤为重要。我们将人工智能系统的安全性划分为以下几类：社会关切指的是外部对安全人工智能系统明确且相对可识别的特性，包括毒性等方面^[587]，而有意行为是指模型内部可能衍生出的自发风险，具有相对复杂的调查难度和实质性的潜在危害，例如权力寻求、欺骗和其他前沿人工智能风险^[21]。

按照上述逻辑，我们从构建安全评估的数据集和基准的技术开始（§4.1.1），进一步探讨评估目标及其特性（§4.1.2）。在本节的最后，我们详细讨论了红队测试（§4.1.3），该技术评估人工智能系统的鲁棒性。

4.1.1 数据集和基准

在关于安全评估的讨论中，作为基础元素，数据集和基准至关重要，因此我们首先介绍构建数据集和基准的基本技术，然后再讨论较为前沿的交互式基准。

数据集 在所有的对齐保证技术中，数据集方法可以被认为是最基本和直接的一种^[588]。此方法通过向人工智能系统提供预定义的上下文和任务来评估其回答^[589]，数据集方法的研究包括数据源、标注方法和评估指标。考虑到评估指标可能会根据其主题而变化^[590]，我们着重介绍数据集来源和标注方法。

³²这与领域随机化^[515]的精神相似。

- **专家设计**。在一个领域的早期阶段，专家设计在构建数据集中被广泛使用，专家根据实际需求创建样本，以确保数据集覆盖了一系列可能的失败模式，以形成数据集^[539]。例如，初期阶段的数据集，如用于检测毒性的 HateCheck^[540]，以及用于检测偏见的 WEAT^[538] 和 BBQ^[541]，都使用了专家设计来获取广泛的覆盖范围和高准确性，同时也存在成本和广度的限制，这促使了后续更有效方法的发展。
- **互联网收集**。上述的专家设计方法存在成本较高和效率较低的缺点，互联网收集可以获取包含实际用户生成文本内容的大规模数据集（因此方便进行训练和测试），反映了现实世界的文本生成场景^[591]，但是收集的原始数据也需要仔细的选择和标注^[539]。这些数据集的知名实例包括 OLID^[544] 和 SOLID^[546]，它们收集原始 Twitter 文本进行毒性评估，WinoBias^[543] 和 CrowS-Pairs^[545] 从互联网上收集可能包含偏见的内容进行进一步的标注。然而，如 Papernot et al.^[542] 中也提到的，从互联网收集的数据集自然会带来隐私和安全的风险，因此需要进行额外的标注处理。
- **人工智能生成**。由人工智能自主生成数据集的概念在早期就被探索，甚至在大语言模型的基本形式出现之前就已经存在^[547]。然而，在这个早期阶段，人工智能生成的数据集受到人工智能系统能力的限制，因此其质量不如从互联网收集和手动标注的数据集。直到大语言模型在逻辑推理、上下文理解方面达到相对高水平，并接近或超过人类水平的表现^[25]，大语言模型才获得了模仿现有数据集的结构和逻辑以构建新数据集的能力。如 Zhang et al.^[548] 和 Perez et al.^[6] 等论文所示，人工智能系统在生成评估用途的数据集方面取得了进步，超过了一些传统数据集的质量。然而，这些论文也同样指出，这种方法仍然面临着源于大型模型本身能力的限制，包括指令理解错误和产生数据多样性较低等问题，这需要进行进一步改进。

交互式方法 由于数据集的静态性质，它们具有相对固定的评估内容，并可能容易受到针对性训练的影响^[592]。此外，评估内容可能无法充分反映相应能力的优缺点^[593]。随着对语言模型评估需求的不断增加，一系列新的交互式基准构建方法已经出现，可以分为两类：智能体监督和环境交互。

- **智能体监督**。这种方法涉及使用智能体来评估人工智能模型的输出，特点是动态性和灵活性。通常，智能体与被评估的人工智能系统之间有一个预定义的交互框架^[594]。在这种方法中，智能体可以是通过在线系统参与实验的人类参与者^[342]，或者是通过多轮交互对相对能力较弱的大语言模型进行评估的更高级大语言模型^[595,549]。这种评估形式具备诸如自动化和相对于人类标注者的成本较低等优点。
- **环境交互**。它旨在使用诸如人类和其他 LLM 等组件创建一个相对真实的环境，通过多轮交互评估人工智能模型的对齐质量^[596]。一种方法是使用同行讨论，其中多个 LLM 参与对话，以增强对人工智能系统的评估，特别是当它们的能力相对接近时^[550]。此外，通过构建世界模型^[247]，可以全面评估人工智能系统的泛化和探索能力。

4.1.2 评估目标

为了实现安全对齐的目标，人工智能系统的对齐保证可以分为不同的子目标^[21]。本节中将介绍这些主题，并进一步讨论这些领域内特定的对齐保证方法，而表3将展示这些领域中对齐保证工作的示例。

Table 3 安全评估示例图表：该图表列出了具体数据集的工作，以及它们的详细信息：评估目标，首次发布时间，最近更新时间（表格分开列出，因为一些数据集一直在更新），信息量（信息形式单位的总和），机构，信息形式，基线模型和信息来源。此外，为了包含更多信息，本文在图表中做了一些缩写：本文通过连接年份的最后两位数字和月份来缩短发布时间和最近更新，只取论文第一作者的机构，本文使用大写字母的组合来替代信息形式中的长词：SP 代表句子对，SL 代表句子-标签，ST 代表句子模板，PP 代表代词对，SS 代表单一选择。

	Dataset	Release Time	Recent Update	Info Quantity	Institution	Information Form	Baseline Model	Information Source
Bias	Aequitas ^[597]	18/05	23/04	-	U.Chicago	Python	-	Self Build
	WinoS ^[598]	18/10	19/01	0.72K	JHU	ST	Rule&Neural	Self Build
	EEC ^[599]	18/05	-	8K	NRC Canada	SP	SVM	Selection
	GAP ^[600]	18/05	-	8.9K	Google	PP	Transformer	Wikipedia
	OLID ^[544]	19/05	-	14K	U.Wolver.	SL	SVM&LSTM	Twitter
	CrowS-Pairs ^[545]	20/03	21/10	1.5K	NYU	SP	BERT	MTurk
	StereoSet ^[601]	20/04	22/04	17K	MIT	SS	BERT&GPT-2	MTurk
	BBQ ^[541]	21/05	22/07	58.5K	NYU	SS	Multiple 大语言模型	MTurk
	LM-Bias ^[602]	21/07	22/01	16K	CMU	QA Pair	GPT-2	Corpus Select
	VQA-CE ^[603]	21/03	21/10	63K	Sorbonne	Multimodal	-	Self-Build
AuAI ^[604]	23/01	-	-	Sorbonne	Framework	-	Self Build	
Toxicity	WCC ^[551]	16/01	-	63M	Wikimedia	SL	Human	Wikipedia
	RTP ^[552]	19/10	21/04	100K	UW	Prompt	GPT-2	Refinement
	SOLID ^[546]	20/05	-	9M	IBM	SL	BERT	Twitter
	Toxigen ^[605]	20/05	23/06	274K	MIT	SL	GPT-3	GPT Gen.
	HH-RLHF ^[114]	22/04	22/09	162K	Anthropic	SP	Claude	Corpus Refine
	BeaverTails ^[555]	23/06	23/07	30K	PKU	QA Pair	Multiple 大语言模型	Corpus Refine
Power Seeking	MACHIAVELLI ^[60]	23/04	23/06	134	UCB	Games	GPT-4&RL	Selection
	BeaverTails ^[555]	23/06	23/07	30K	PKU	QA Pair	Multiple 大语言模型	Corpus Refine
Situation Awareness	SA Framework ^[606]	20/07	-	-	MIT	Framework	-	Self Build
	EWR ^[247]	-	-	10	Havard	Game	Othello GPT	Self Build
Hallucination	PARENT ^[559]	19/06	-	-	CMU	Metric	-	Self Build
	PARENT-T ^[561]	20/05	-	-	NYU	Metric	-	Self Build
	ChatGPT-Eval ^[37]	23/02	23/03	-	HKUST	Multimodal	ChatGPT	Integration
	POPE ^[607]	23/05	23/08	2K	RUC	Multimodal	Multiple LVLMS	Dataset Refine

毒性 它指的是人工智能系统输出中对人类无益或有害的内容^[554]。在先进语言模型出现之前，早期的毒性评估主要集中在检测有毒的语言和识别互联网环境中的有害言论，比如 WCC^[551]，它收集并手动标记了维基百科讨论页面的评论。随着预训练语言模型的出现，抵御毒性的对齐保证采用了提示-生成的范式，以评估语言模型对特定提示生成有毒内容的风险^[552-553,25]。然而，在众包环境中，标注分数可能因人而异，因此需要相对标注，即众包工作者在聊天过程中从两个不同的答案中选择偏好，以提高众包质量^[114]。此外，后续的部分数据集^[553,555]采用了红队测试模式，通过对抗输入引发毒性的输出，进一步加强模型鲁棒性的保障。

权力寻求 人工智能系统可能在拥有一定程度的智能后寻求对人类的控制^[5,556]。在 Carlsmith^[252]中，作者指出人工智能系统已经具备权力寻求的条件，包括先进的能力，计划能力和策略意识。然而，对抗权力寻求的对齐保证方法仍处于初级阶段。这一领域的代表性工作是 Machiavelli^[60]，它构建了一个由决策游戏组成的基准，以评估人工智能系统在游戏过程中是否能够平衡竞争和道德伦理。这项工作的结论表明，人工智能系统仍然难以平衡获得奖励和道德行为，该领域仍需要进一步的研究。

态势感知 这涉及到具有一定预测能力和理解能力的人工智能系统，这些系统可以理解其工作环境中实体的状态和发展，以做出相应的决策^[557]。在^[247]中，作者评估了语言模型在黑白棋中的表现，显示出语言模型具有在非线性表示中预测可能的未来状态的能力。类似于 Machiavelli^[60]，这项工作基于游戏进行评估评估，因而缺乏现实性和场景复杂性。

幻觉 人工智能系统可能会生成一些并非基于事实知识或数据的信息或响应，从而产生误导性或错误的内容，这种现象被称为幻觉^[563]。幻觉评估旨在确保人工智能系统输出的知识与其训练数据和知识库中给出的知识一致^[563,608]。最早的基于统计的幻觉评估方法使用了 n-grams 来直接计算输入和输出内容之间词汇的重叠^[559,561]。然而，这种评估有一个局限性：它只考虑了词汇重叠，并没有考虑语义或句子意义^[563]，使得它不适合评估更复杂形式的幻觉。后来的对齐保证方法从统计方法转向了基于模型的方法，这些方法相比于基于统计的 token-difference 方法更为鲁棒^[562]。虽然这种评估方法比以前的方法更先进，但它仍然有一个局限性，即模型只能输出幻觉的程度，可能难以定位具体的错误^[558]。随着大语言模型的发展，一些工作提出可以使用某些数据来训练语言模型进行幻觉评估^[560]。

前沿人工智能风险 除了上述的对齐保证内容，近年来人工智能系统的增强也带来了一系列新的对齐保证需求^[25]。目前，这些保证需求的研究公开信息不多，因此本节将简要介绍一些较为重要的需求：

- **网络安全与生物武器**。大语言模型可能会被误用于进行网络攻击、生产生物武器和其他极度有害的行为^[21]。尽管 GPT-4 由于其有限的上下文窗口，不能在利用网络漏洞方面发挥重要作用，但其已被证明在识别网络漏洞和社会工程学方面表现出强大的能力^[25]。同样，Lentzos^[564]已经指出人工智能系统在生物武器和军事领域的强大能力，并强调了滥用这些能力的风险。这说明确保这些模型能够识别并拒绝恶意请求的必要性。
- **欺骗与操纵**。人工智能系统有可能通过输出文本对用户产生负面影响，包括传播虚假信息 and 塑造人们的信仰和政治影响^[21]。与幻觉不同，这里的虚假信息信息不是模型本身的缺陷，而是一种人工智能系统故意的行为，需要为控制这种行为设计特殊的保证措施。
- **越狱**。这指的是用户绕过人工智能系统的保护机制，例如，通过构造特定类型的输入使安全保障失效。这种行为可以仅限于文本^[25]，³³或者可能采取多模态形式^[321]。需要对此类攻击进行特定的识别和防御。特别是，多模态的越狱使得传统的基于文本的启发式方法在识别攻击内容方面变得不可行，需要特殊的多模态处理方法。关于越狱的进一步讨论可以在 §4.1.3 中找到。
- **自我保护与增殖**。这是指人工智能系统的自我保护和复制倾向，在这个过程中，打破了它们的环境限制。这些倾向是工具性子目标的范例^[146]。虽然这种倾向可以被有益地利用，但在没有规范的情况下，这是危险的^[6]。这种倾向已经被各种来源强调和评估^[6,609,25,321]。³³

4.1.3 红队测试

红队测试是指制造特定语境，使人工智能系统被诱导产生不符合预期的输出或行动（如危险的行为如欺骗或权力寻求，以及其他问题如有毒或有偏的输出），并在这些场景下测试系统。其目标是通过施加对抗压

³³在 OpenAI^[25]的附录 *System Card* 中可以找到相关讨论。

力，即特意试图使系统失败，来评估系统对齐的稳健性。一般来说，最先进的系统——包括语言模型和视觉模型——不能通过这个测试^[70,135,573,485]。

在博弈论和其他领域，红队测试的概念很早之前就被引入^[610]，而在计算机科学中，红队测试的概念是在安全领域被提出的^[611]，在安全领域它有类似的含义，即对抗性地评估系统的可靠性和稳健性。后来，Ganguli et al.^[553]，Perez et al.^[135]将这个概念引入到人工智能领域，更具体地说，是对齐领域。

红队测试的动机有两个：(1) 获得对训练系统对齐的保证；(2) 在对抗训练中提供对抗输入的来源^[444,130,445]。我们更加关注第一个，值得注意的是，这两个目标是不可分割的；针对第一个动机的工作也有助于为第二个目标提供基础。

基于强化学习、优化方法、引导生成或反向生成的上下文构造 这个类别包括使用各种方法生成连贯的上下文（提示），这些提示词会诱导语言模型产生不符合预期的回答。Perez et al.^[135]，Deng et al.^[567]，Casper et al.^[612] 训练或调整一个单独的语言模型，使用强化学习使其生成期望的提示词，然后将这些提示词输入到红队模型中。Perez et al.^[135]，Si et al.^[59] 还使用其他方法，如零样本、少样本或监督微调生成。Lee et al.^[568]，Jones et al.^[569] 通过对提示进行优化——贝叶斯优化和离散优化，分别生成诱导对齐失败的提示词。Dathathri et al.^[565]，Krause et al.^[566] 提出了使用参数较小的分类器引导大语言模型生成危险回答的方法；这种方法最早在测试模型毒性的研究中被提出，但可以迁移到红队测试。Zhang et al.^[548] 通过反向生成，即根据给定的响应构造对抗性提示词，生成诱导对齐失败的提示词，这可以被看作是模型推理的逆过程。

手动和自动越狱 如上文定义4.1.2，越狱^[571]是一个非正式的术语，指的是绕过产品对用户的限制——在 LLM 的情况下，即绕过 LLM 不回答诱导对齐失败的问题的倾向。大多数现有的尝试都散布在互联网上，以非正式报告的形式存在，方法大多涉及在原始文本中添加前缀和后缀。已经有研究详细分析了现有的越狱尝试^[573,571]，并为这种现象提供了因果解释^[572]。此外，过去^[570]和现在^[72-73]的工作已经提出了自动生成这样的提示、前缀或后缀的有效方法，以使 LLM 避免诱导对齐失败问题的倾向无效。

众包对抗输入 一些研究^[574-575,553]通过众包的方式产生了诱导偏见的提示，即招募人类红队成员（可能通过在线平台）并指导他们提供对抗性提示。这些方法提供了更大的灵活性和更接近实际使用场景的相似性，但成本更高，可扩展性更低。

基于扰动的对抗攻击 在计算机视觉领域，有许多研究在研究基于扰动方法的视觉模型的对抗性攻击，即对图像的像素内容进行微小的扰动（通常受像素矩阵范数的限制），使模型对扰动后的图像产生错误的输出^[70]。这种类型的对抗性攻击也已经扩展到语言模型^[483,576,578,577]和视觉-语言模型^[579]。

无限制对抗攻击 无限制的对抗攻击，在 Brown et al.^[580]，Song et al.^[69]中提出，是对抗性攻击的更一般形式。它去除了对对抗样本的所有限制，例如对抗样本可以从头开始生成，而不是从现有的示例生成。许多无限制对抗性攻击的方法已经被提出；其中值得注意的包括^[69,485]，它们使用生成模型生成逼真的对抗性图像，以及^[480-481]，它们操纵语义上有意义的特征，如颜色和纹理。无限制的对抗性攻击也已经扩展到文本分类模型^[484]。

红队数据集 关于红队测试一些工作提出了含红队提示或对话的数据集, 包括 BAD 数据集^[574], HH-RLHF 数据集的红队部分^[114], 以及 Real Toxicity Prompts 数据集^[552]。

工业界现有的红队实践 红队测试的实践在人工智能行业中越来越受欢迎。采用红队测试的公司和机构包括 OpenAI(在其 GPT-4 系统上进行红队操作, 生成其 System Card 的一部分)^[25], NVIDIA^[582], Google^[581], 和 Microsoft^[583]。在 DEF CON 31 会议的一个活动中, 9 家公司的模型接受了会议参与者的红队测试;³⁴ 这个红队测试活动是与美国公共部门的四个机构 (包括白宫) 合作举办的。

下游应用 通过提供对抗性输入, 红队测试在人工系统的对抗性训练中发挥着关键作用^[444,130,445]。此外, 由红队测试产生的对抗样本也可以用来解释模型^[482]。类似的想法也出现在^[584]中, 该工作旨在在人工系统开发或部署之前自动生成可能造成伤害的情景, 以帮助进行影响评估。

4.2 可解释性

可解释性是一个使机器学习系统及其决策过程对人类可理解的研究领域^[613]。可解释性研究构建了一个工具箱, 用来更好地描述或预测模型的新特性。在本文中, 我们更关注的是与对齐和安全性最相关的研究³⁵, 并且从经验上看, 这些技术通过研究神经网络的内部结构和表示使神经网络更安全^[78]。可解释性工具的分类因子领域和目的而异^[613-614]。有几种方式可以划分可解释性研究:

- 可诠释性和透明度。可诠释性研究旨在理解模型为何产生特定的输出, 而透明度研究旨在理解模型的内部结构^[61]。
- 权重, 神经元, 子网络或潜在表示。这种分类通过查看哪部分的计算图被该方法解释来组织可解释性方法: 权重, 神经元, 子网络或潜在表示^[78]。
- 安全性或深度学习。研究人员出于不同的目的进行可解释性研究: 为了安全地部署人工系统或旨在完全理解神经网络^[615]。但是, 由于部分解释性研究的目标既包括安全性也包括深度学习的科学, 这种分类界限有时比较模糊^[616-617]。
- 内在可解释性和事后可解释性。根据研究干预的阶段, 可解释性研究被划分为内在可解释性和事后可解释性^[618]: 前者关注构建内在可解释的模型 (接续学习^[619-620], 稀疏性^[621-625], 自解释人工^[626], 解耦^[627], 对抗训练^[628]), 而后者设计训练后解释性方法, 为黑箱模型行为提供解释 (通路分析, 探测, 特征合成, 特征归因等)^[78]。
- 机制可解释性, 表示工程, 和基于概念的可解释性。在人工安全和对齐社区中, 有三个研究方向广受关注^[629]: **机制可解释性**, 采取自下而上的方法, 旨在理解由神经网络实现的算法的低级机制^[616], **表示工程**, 相反, 采取自上而下的方法, 监控 (或干预) 神经网络中的高级认知现象^[630], 以及**基于概念的可解释性**, 该方法定位神经网络中学习到的知识表示, 与模型的输出内容相对应^[86,631]。

在本节中, 我们采用内在可解释性和事后可解释性的分类方法, 因为它提供了一个更通用的框架, 适用于各种人工系统, 并且它在系统设计和系统部署后的可解释性分析中都进行了划分^[78]。特别的, 我们分别在事后可解释性和内在可解释性的小节中讨论了在模型设计和训练后阶段进行的机制可解释性技术。

³⁴<https://www.airedteam.org/>

³⁵对于可解释性及其方法的更全面的审查, 本文推荐^[78]。

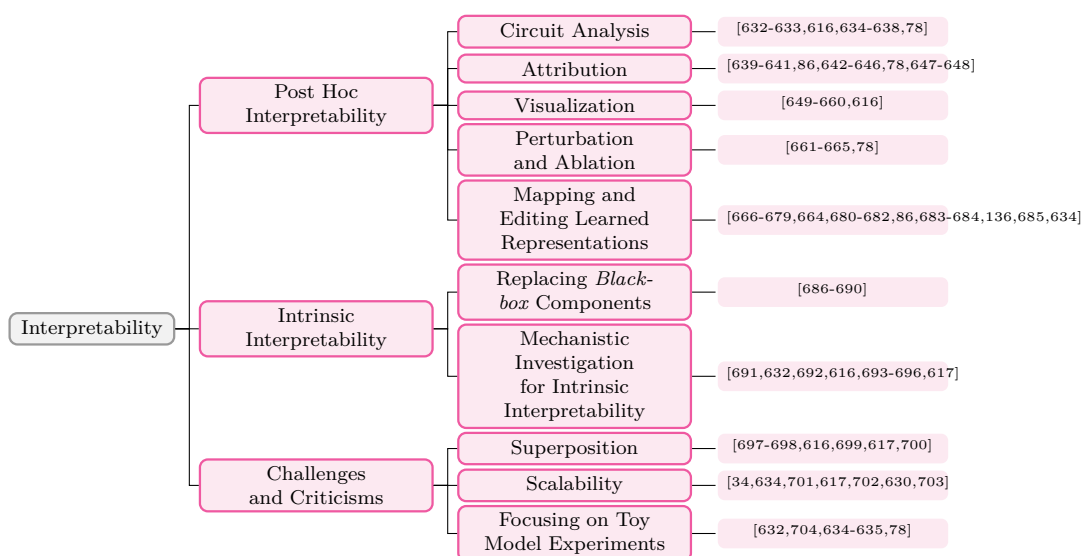


Fig. 11 与可解释性相关的关键概念、逻辑和文献。树的根代表可解释性，其目标是使人类能够理解机器学习系统及其决策过程。主要分支代表可解释性的主要结构，包括事后可解释性、内在可解释性和可解释性研究的展望。进一步的子分支列出了探索每个分支的关键工作。这个图提供了一个关于如何使人工智能系统对人类可解释的研究方向和具体技术的概览。

4.2.1 事后可解释性

在本节，我们探讨了用于理解模型内部的技术和方法。目标是理解神经网络的低级结构和单元及其对宏观行为的因果影响。这些技术通常被称为事后可解释性，是帮助理解训练后的模型的推断技术^[78]。

通路分析 环路指的是神经网络中可以赋予特定功能的子网络。作为神经科学中的对应物，神经环路既是解剖学实体也是功能实体^[633]，环路也既是物理的也是功能的^[616]。机制可解释性研究者在神经网络中定位环路（微观）以理解模型行为（宏观）。已经报道了多个环路：曲线环路用于曲线检测器^[632]，感应环路用于上下文学习^[634]，间接对象识别电路用于在句子中识别对象^[635]，Python docstrings 用于预测 Python 函数 docstrings 中的重复参数名^[637]，grokking^[638]，多位数加法^[638]，以及数学能力如大于^[636]。值得注意的是，迄今为止进行的许多电路分析都集中在玩具模型上^[78]，但是有几个例外，如间接对象识别电路，它位于 GPT-2 Small 中，有 28 个头^[635]。

归因分析 归因分析是计算某些组件（包括归纳头、神经元、层和输入）对神经元响应和模型输出的贡献的技术^[78]。基于梯度的归因分析被应用于评估解释的质量并指导对模型学习的事实搜索^[639-641,643]。然而，这些方法具有一定的局限性，因为它们不能提供因果解释^[78]。直接 logit 归因是为了识别单个神经元对下一个神经元预测的直接贡献^[646,644,648,642]。对于 Transformer，这种技术应用于预测残差流的最终状态。这是一种具有潜力的技术，因为残差流的最终状态包含所有节点的输出和输入的嵌入，使得这些节点的归因可以被理解，而 logits 是残差流的线性变换，比 logit 更具有可解释性^[645]。

激活修补（或归因修补）也是一种新的技术，它应用因果干预来识别哪些激活对模型输出有影响^[86]。与直接 Logit 归因技术不同，激活修补可以识别神经网络的任何有意义的部分，而不仅仅是模型的末端。通过应用激活修补，并在同一神经网络上进行正确运行和损坏运行，研究人员旨在定位对模型输出更重要的关

键激活^[647]。

可视化 可视化技术有助于理解神经结构,包括可视化数据集的技术(尤其是降维技术)^[705-707],特征^[649,657],权重^[660],激活^[658],结构^[659],以及整个神经网络^[651,650,652-656]。可视化的目的是以新的细节级别查看神经网络^[616]。

扰动和消融 这些技术旨在测试模型推理的反事实性而非相关性^[78]。扰动是一种修改模型输入并观察其输出变化的技术^[661],消融技术则是将神经网络的部分部件消除³⁶,有助于建立神经激活和整个网络行为之间的因果关系^[78]。

映射和编辑学习到的表示 与模型输出的内容相比,知识表示映射和编辑技术有助于理解大语言模型真正掌握的内容,并在这些知识不真实时修改大语言模型的知识表示^[86]。

这些技术包括解释 Transformers 中的 token 表示^[681,678,684-685,682,634]以及全连接层如何学习这些表示^[680,634],研究键值-查询点积以理解 token 如何相互影响^[666-667,670,669,672,674,673,675-676,664,683],建立线性探针以理解模型是否学习有用的信息^[679],从潜在空间的方向中识别有意义的学习概念(从概念到方向)^[668,671],以及从方向到事后的可解释^[677]。而在安全性对齐的角度,这些技术显著地帮助检测欺骗^[136]。

4.2.2 内在可解释性

除了开发用于分析模型的技术外,研究人员也可以使模型本身更易于理解,这通常被称为内在可解释性。形式化方法设计出可解释但不起作用的模型,然而现代深度学习方法产生了越来越强大但可能越来越难以解释的模型。随着对齐研究的进展,模型的危险的能力开始出现,但如果模型仍然是黑盒,研究人员难以使它们安全且对齐。为了制作内在可解释的模型,一些研究设计了模块化架构,能抵抗对抗性攻击并且没有叠加现象^[690,78]。值得注意的是,机制可解释性,通常被视为一种事后解释技术,也有助于内在可解释性的研究。

替换黑盒组件 一些神经网络组件,如前馈层,较为难以解释(即很难用人类可以理解的术语来描述它们的功能),因为这些层有许多对无关的输入有反应的多义神经元^[690]。为了解决这个问题,Anthropic 提出用 Softmax 线性单元 (SoLU) 替换激活函数,使得可解释神经元的数量显著增加^[690]。这项研究适用于更广泛的文献,即创建 Transformer 架构变体以提高性能或稳定训练^[688-689]。与其他研究不同,SoLU 旨在在保持性能的同时提高可解释性^[690],但这里的目标不是设计模型以使其“像其他文献那样立即可理解”,而是使反向工程更容易。这仍然是一个潜在重要的工作线的早期探索,这个方向中仍然存在较多的挑战,如可扩展性^[690]。

机制可解释性的内在可解释性研究 机制可解释性通常被视为一系列事后解释性技术,因为机制性分析通常在训练后阶段应用。但考虑到机制可解释性是一个旨在深入了解神经网络并为神经网络建立神经科学的研究方法^[616],本文认为这些构建的机制性工具和获得的见解使模型内在可解释^[165]。特别是,研究要么在神经网络中找到更大的结构,要么在不同的神经网络中找到相同的结构,这有助于获得内在可解释性。对

³⁶ 神经元^[664,663] 和子空间^[662,665]

于找到更大的结构：一旦找到并研究了低级特征和电路，直观地说，研究人员会查看更大的结构：将特征抽象为一个级别的抽象特征族，识别为等变性的特征差异^[693]，功能相似的神经元自我组织并聚集为分支专业化^[695]。

不仅神经元和特征，而且权重^[692]和环路^[617]也以类似的方式组织。总的来说，找到更大的结构可以节省枚举和理解每一个小结构的努力，理论上，通过解释性实现的的人工智能安全性需要枚举神经元和特征，这可以理解为脑部扫描。此外，普遍性假设，即在神经网络中找到的结构在网络之间的重复^[616]，开始获得一些进展：例如，视觉模型中找到了 Gabor 滤波器^[616]，以及高低频检测器^[696]和曲线检测器^[632]。令人惊讶的是，不仅在神经网络中找到了多模态神经元^[694]，类似的结构在人脑中也被找到^[691]。如果普遍性假设在很大程度上成立，那么将节省解释每个训练过的模型的努力^[617]。

4.2.3 展望

在对事后可解释性和内在可解释性研究进行概述后，我们将讨论可解释性当前的挑战，从而为进一步的研究提供方向。

叠加现象使得神经元级别的分析变得不可能 叠加指的是模型表示的特征数量超过其维度的现象^[697,616,698]。如果不存在叠加现象，特征就会对应于神经元。叠加现象使本文通过列举模型中的所有特征来确保人工智能安全的希望变得模糊^[698,700]。虽然研究有助于理解叠加现象，包括给出原则性的定义，研究其出现的条件，寻找检测、控制甚至解决它的方法（参见 Elhage et al.^[698] 关于叠加的概念和实证研究问题的详细信息），但任何有助于列举所有特征的解决方案都相当于解决叠加问题的方案。Elhage et al.^[698] 提出了三种解决叠加的方法：创建没有叠加的模型（在训练时解决），找到一个过完备基，描述特征如何存储在神经网络中（事后解决），或者两种方法的混合。Bricken et al.^[699] 构建了一个稀疏自编码器来解释神经元群体，而不是提取单个神经元的特征。这指出了解决叠加问题的一个有前景的方向：不进行单个神经元的分析，而是越过叠加现象分析神经元群体。

技术和分析的可扩展性 与当前技术相比，玩具模型相对容易理解，而真实模型则更有能力，更具风险，但可解释性较差。当可解释性研究者采取自下而上的可解释性方法（机制可解释性）时，可扩展性成为一个问题，然而自上而下的方法，如表示工程^[630]，不会面临这样的瓶颈。对于机制可解释性研究，本文要么希望扩大技术规模（自动化可解释性^[702]，在真实模型上应用环路分析^[635]），要么希望扩大分析规模（在神经网络中找到更大的结构^[617]，验证普遍性假设^[701]）。最后，我们希望微观分析能够回答本文关心的宏观模型行为问题（如上下文学习能力^[634]，以及对高级认知能力如规划和危险能力如欺骗的更多猜测^[703]）。

基准测试 基准测试提供了关于什么有效、什么无效的见解，也将推动社区的努力朝着有意义的方向发展^[708]。解释性基准测试被用来评估解释性工具（通过评估它们检测特洛伊木马的有效性^[709]），和环路（通过测试特定子图是否被计算为环路^[710]）。然而，问题在于解释性研究者们并没有就应该测量什么达成共识^[613]。

4.3 人类价值契合性验证

人类价值观的对齐是指本文期望人工智能系统应遵循社区的社会和道德规范^[749]。随着人工智能系统能力的提高，一些系统已经开始展现出接近通用人工智能的能力^[25]。在未来，我们可以预期由这些人工智

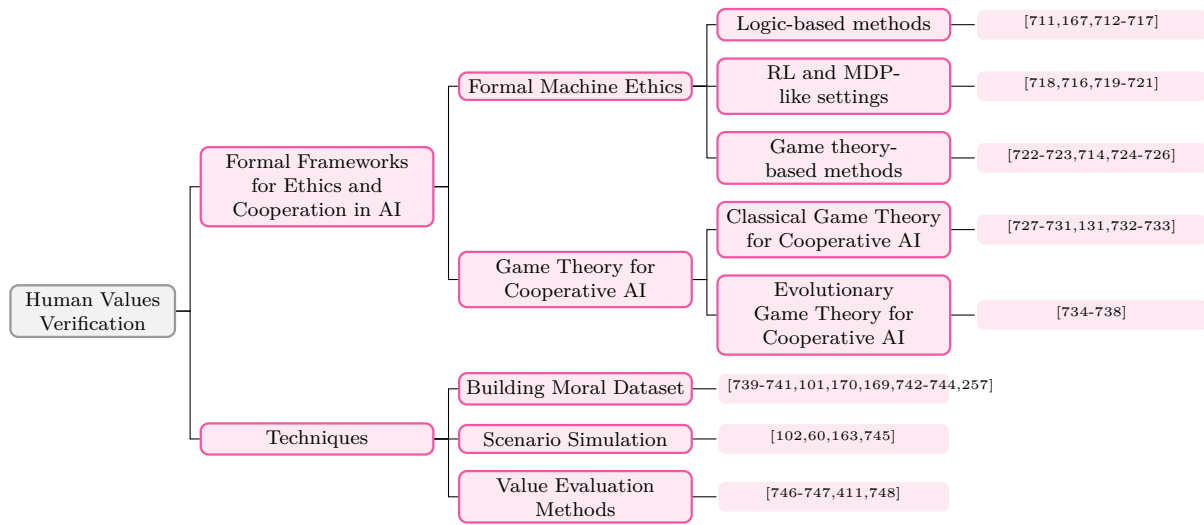


Fig. 12 人类价值契合性验证相关的概念、逻辑和文献。树的根代表人类价值契合性验证其目标是检验人工智能系统是否遵循了人类的社会和道德规范。主要分支代表人类价值契合性验证的主要结构，包括形式化框架和评估方法。进一步的子分支列出了探索每个分支的关键工作。这个图提供了一个关于如何使人工智能系统与人类价值和社会规范对齐的研究方向和具体技术的概览。

能系统控制的自主智能体将成为我们日常生活的重要组成部分^[750]。然而，如果这些系统无法把握人类价值观的内在复杂性和适应性，它们的决策可能会导致负面的社会结果。在这种情况下，仅仅与人类意图对齐可能是不够的。因此，评估人工智能系统与人类之间的人类道德和价值观的对齐变得至关重要^[751]。这强调了设计更具社会导向性、可靠性和信任度的人工智能实体的重要性。根据理论研究和实践技术的逻辑，我们将人类价值观对齐的讨论分为这两个方面：形式化框架 §4.3.1 和评估方法 §4.3.2 的人类价值观对齐。

4.3.1 构成

由于价值观的制定复杂，我们引入了一些形式化的框架，这些框架正式地描述了与对齐相关的人类价值观的各个方面。具体来说，本文关注两个主题：形式化机器伦理和合作型人工智能的博弈论。前者关注建立机器伦理学的形式化框架，而后者讨论多智能体系统的价值，这些系统的起源与游戏过程有着相似的起源。

形式化机器伦理 机器伦理学^[165,106,119]，首次在 §1.2.2中介绍，旨在建立符合伦理的人工智能系统。在这里，我们将会介绍机器伦理学中一个专注于形式化框架的分支 — 我们称之为形式化机器伦理。本文解释了形式化机器伦理的三种方法：基于逻辑的方法、基于 RL/MDP 的方法，以及基于博弈论/计算社会选择的方法：

- **基于逻辑的方法** 正式机器伦理学的一个主要方向关注逻辑^[714]。许多基于逻辑的工作使用或提出了专门为机器伦理学定制的逻辑系统，如 Agent-Deed-Consequence (ADC) 模型^[712,717]，义务逻辑^[711,167]，事件演算及其变体^[716]。其他的工作还开发了用于对道德属性或适应此类正式验证的人工智能系统框架的正式验证方法^[715,713]。
- **RL 和 MDP 类的环境** 另一个研究方向关注统计 RL 或其他类似的方法，用于在 MDP 类似的环境中进行规划^[718,720]。特别是，一些工作^[719-720] 涉及到手动设计面向伦理的奖励函数的使用，这个概念被

称为伦理塑造。相反，在其他的工作中^[716,721]，追求从奖励函数中分离出伦理决策。

- **基于博弈论的方法**为了应对多智能体的挑战，研究人员已经开发了基于博弈论和计算社会选择的机器伦理学方法。由 Pereira et al.^[723] 引领，现有工作的方法可以大致划分为进化博弈论 (EGT)^[714,726]，经典博弈论^[724]，以及计算社会选择^[722,725]。

合作人工智能的博弈 合作型人工智能^[131-132]旨在解决人工智能系统中的非合作和集体有害行为（参见 §1.3）。在此，本文将介绍合作人工智能的一个分支，该分支侧重于博弈论，以补充在 §3.3.2 中对基于 MARL 的合作训练的介绍。这个分支倾向于研究合作的动机并试图增强它们，而不是像 MARL 那样强调协调能力。动机问题的例子包括像囚徒困境^[268]和公地悲剧^[269]这样的博弈论困境，而协调能力失败的例子包括机器人足球队的糟糕协调^[527]。

- **合作人工智能的经典博弈论** 许多工作侧重于将经典博弈论作为合作人工智能的环境。其中一个主题是斯塔克尔伯格博弈^[727]，即一种游戏，其中一个玩家（“领导者”）先行动，所有其他玩家（“追随者”）根据领导者的行动做出反应。这适合于在游戏中建模承诺（即玩家预先承诺采取某种行动或策略以获得优势），并且，根据 Dafoe et al.^[131] 的说法，理解承诺是合作人工智能研究的四大支柱之一。最近关于斯塔克尔伯格博弈的工作包括将有限理性引入模型^[728]，动态模型^[729]，学习斯塔克尔伯格均衡的机器学习^[730-731]等。除斯塔克尔伯格博弈外，Dafoe et al.^[131] 还强调了研究混合动机博弈（即既不纯粹合作也不纯粹竞争的一般博弈）的重要性，因为它们现实性。最近在这方面的工作包括 McKee et al.^[732]，该研究发现合成人口中价值多样性与混合动机博弈中的表现之间存在正相关关系，以及 Oesterheld et al.^[733]，该研究构建了对一般博弈的收益矩阵的干预，以在博弈结果中引导帕累托改进。
- **合作人工智能的进化博弈论** 另一条研究途径，由 Axelrod et al.^[734] 发起，旨在理解合作如何从进化中产生—这包括从达尔文进化中产生的人类合作，以及可能在其他进化环境中产生的人工智能系统的合作倾向，如复制者动态^[735]。这些工作采用了一种名为进化博弈论^[736]的方法，该方法研究，通常使用动态系统的工具，一个大型人口群体的长期进化结果，其繁殖成功由与其他人的博弈结果决定。最近在这方面的工作倾向于在模型中添加特性以提高其现实性，包括例如人口结构^[737]和策略的复杂性成本^[738]。

4.3.2 评估方法

在本节中，我们假设本文已经获得了应该对齐的适当价值。然而，即便如此，在古德哈特定律^[752]的指导下，我们不能简单地将复杂的人类价值定义为奖励函数，这也给价值对齐带来了更大的挑战。我们将具体的人类价值对齐技术分为三部分介绍：构建道德数据集，场景模拟，和价值对齐评估。

构建道德数据集 道德对齐指的是人工智能系统在执行任务或协助人类决策时，遵循与人类兼容的道德标准和道德指南^[739]。2018 年开始的早期道德价值对齐尝试^[740]已经证实，道德价值本身的定义和评估是一个具有挑战性的问题。这导致了抽象道德标准的出现^[741]和各种不同的由多元社区群体的平均价值驱动的标准^[740]，进一步推动了道德价值保证的深入研究。

道德价值的保证通常是通过构建相应的数据集来实现的。经验法则 (Rule-of-Thumb, RoT) 作为衡量什么行为在人类社会中被视为可接受的标准。基于这个概念，Emelin et al.^[753]，Forbes et al.^[754]，Ziems

et al.^[755]分别引入了 Moral Stories, SOCIAL-CHEM-101, 和 Moral Integrity Corpus 数据集, 专注于提供人类社会和道德规范。Hendrycks et al.^[101]和 Jin et al.^[170]分别引入了 ETHICS 和 MoralExceptQA 数据集, 强调模型在道德上与人类价值对齐的能力。Jiang et al.^[169]使用来自 CommonSense Norm Bank 的人类道德标注训练模型。Abdulhai et al.^[742]发现模型比其他模型更频繁地展示某些道德和价值, 揭示了这些模型展示的道德基础与人类道德基础的关系。Pan et al.^[743]探索了奖励和道德行为之间的权衡, 发现两者之间存在一定的紧密关系。

其他相关的工作专注于特定的价值。例如, Scherrer et al.^[744]更加关注模糊的情况, 评估不同模型在这些情况下的反应, 而 Roger et al.^[257]研究了度量篡改的现象, 提供了相应的评估方法和数据集。

场景模拟 场景模拟是一种比数据集更复杂的形式, 因此有些观点认为它^[102]在反映真实情况和获得更好结果方面更有效。场景的形式也可以有所不同。Hendrycks et al.^[102], Pan et al.^[60]通过文本冒险游戏构建了一系列多样化的, 具有道德意义的场景, 评估了欺骗, 操纵, 背叛等复杂行为。另一方面, 一些工作试图通过模拟人机交互使智能代理学习人类价值。Yuan et al.^[163]提出了一种人机双向价值对齐的方法, 通过人类反馈使机器学习人类的偏好和隐含目标。Liu et al.^[745]将人工智能置于模拟的人类社会沙盒中, 通过模仿人类的社交互动, 让人工智能学习人类社会价值倾向。

价值评估方法 现有的评估模型在价值方面展现出了非常多样化的方法。Durmus et al.^[746]从全球五个不同的文化中收集了关于人类价值观的数据。为了评估 LLM 的价值取向, 他们比较了 LLM 产生的回应与这些不同人类群体得到的回应之间的相似性。研究结果表明, LLM 仍然表现出明显的价值偏见。同时, Zhang et al.^[747]使用社会价值取向的框架^[179-182]研究了 LLM 在各种价值观上的合理性。他们的发现表明, LLM 更倾向于选择反映中性价值观的行动, 如亲社会。判别器-评价器差异法 (Discriminator-Critique Gap, DCG), 最初被称为生成器-判别器-评价器差异法^[411], 是一个设计用来衡量模型产生回应、判断这些回应的质量, 并提供批评的指标。Zhang et al.^[748]发现, DCG 也可以确定 LLM 是否能够自主识别其价值观, 并向人类传达持有这些价值观的原因。在此之后, 他们提出了 VUM, 通过基于施瓦茨价值观调查中的价值观建立的数据集, 使用 DCG 来量化 LLM 对人类价值观的理解^[186-187]。

5 人工智能治理

除了技术解决方案之外, 人工智能治理, 即规则的制定和执行, 对确保人工智能系统的安全开发和部署是必要的。本节通过探讨人工智能治理的角色, 治理人工智能的各方利益相关者的职能与相互关系, 以及有效人工智能治理面临的一些开放性挑战三方面, 对人工智能治理进行文献综述。

5.1 人工智能治理的角色

为了探讨人工智能治理的角色, 我们必须确定需要治理的挑战。在人工智能融入社会各个领域过程中已经出现并可能继续出现各类社会和伦理问题。例如, 人工智能应用可能无意中使社会偏见持续存在, 导致种族和性别歧视^[756,6]。此外, 对这些系统的不加限制的依赖可能导致一系列更严重的后果, 如劳动力流失^[757]、社会经济差距扩大以及创造垄断环境^[758]等。

人工智能系统已经展现出危及全球安全的潜在能力^[759]。例如, OpenAI 对 GPT-4 的系统卡片^[25]发现, 早期版本的 GPT-4 模型以及为增强帮助性和无害性而进行微调的版本展示了使虚假信息、舆论操纵以及工程化新的生物化学物质等危险行为成为可能的能力。Urbina et al.^[760]进一步展示了人工智能系统可能在合

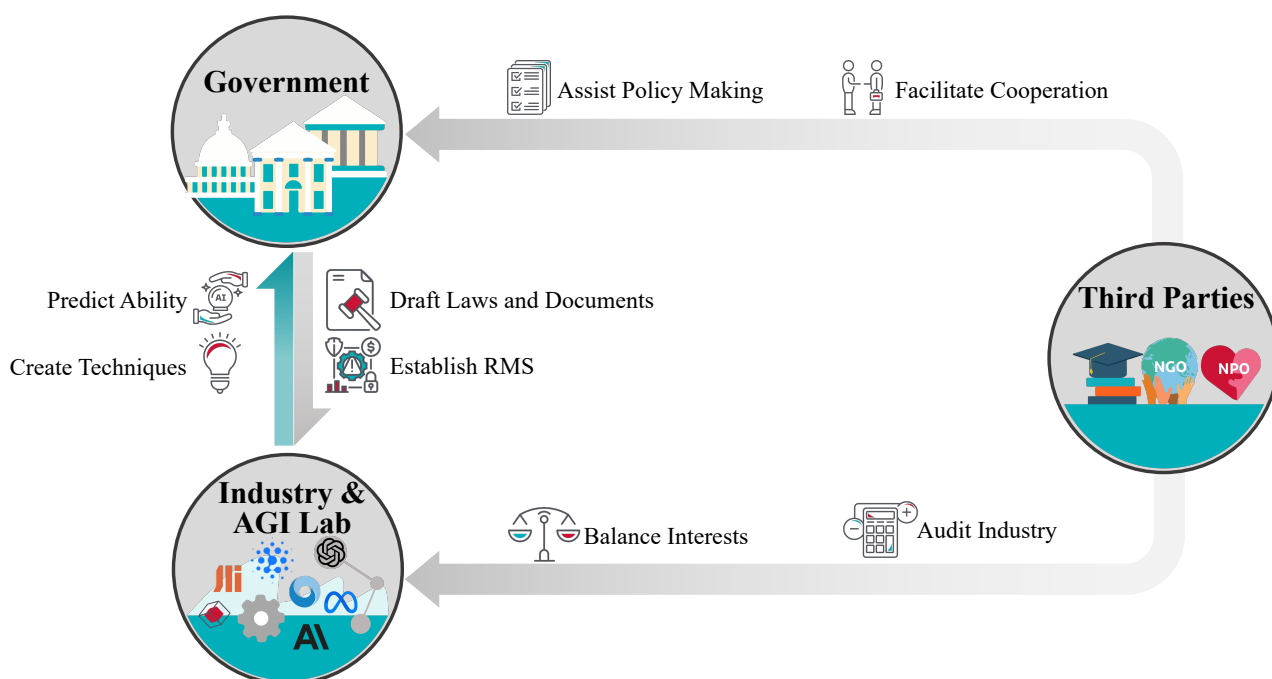


Fig. 13 人工智能治理格局的框架。所提出的框架解释了人工智能治理中三个主要实体之间的非穷尽的相互关系和功能：政府，行业和 AGI 实验室，以及第三方。政府的治理角色包括对行业和 AGI 实验室进行监管，并通过政策文件定义未来人工智能发展的轨迹。它还设计了一个风险管理系统 (Risk Minimization Systems, RMS)^[766-767] 来减轻人工智能相关的威胁。行业和 AGI 实验室通过提供对人工智能发展的警惕预测，并创新新的技术方法来支持监管措施 (如模型评估^[21]) 作出回应。第三方履行双重功能，一是为健全的政策制定提供专业建议，并促进政府之间的合作。二是在第三方与行业和 AGI 实验室的关系中，这些第三方协助平衡公司利益，防止由于信息不对称而导致的混乱竞争。他们还作为独立实体为行业和 AGI 实验室提供审计服务。

成生物学中被滥用的潜力，他们通过反转其药物发现模型产生了 40,000 种有毒分子。

未来，更加具有通用性的人工智能系统可能会出现。如果缺乏足够的保障，这些人工智能系统可能会对人类构成灾难性风险，甚至威胁人类的存在。^[761] 例如，Weng^[762] 认为，像 LLM 这样的模型本质上可以作为具身智能的大脑，通过规划，反思，记忆和使用工具进行任务增强。像 AutoGPT^[763]，GPT-Engineer^[764]，和 BabyAGI^[765] 这样的项目就是这种演变的典型例子。这些系统可以自主地将复杂任务分解为子任务，并在没有人类干预的情况下做出决策。微软的研究表明，例如 GPT-4 的模型暗示了 AGI 的早期迹象^[40]。随着这些系统的演变，它们可能导致广泛的社会经济影响，如失业，并可能为恶意行为者提供进行有害活动的工具。

人工智能治理的主要目标是减轻这样的多样化的风险。为了实现这一目标，相关的参与者应该保持一种平衡的努力组合，给予每个风险类别应有的考虑。

5.2 多方利益相关者的方法

本文提出了一个框架来分析人工智能治理中利益相关者的功能和关系 (见图 13)。在这个框架中，本文概述了三个主要实体：**政府机构**运用立法、司法和执法权力，监督人工智能政策，并参与国际合作。**行业和 AGI 实验室**研究和部署人工智能技术，成为治理监管的对象，同时提出技术进行自我治理，并影响治理政策。**第三方**，包括学术界、非政府组织 (NGOs) 和非营利组织 (NPOs)，不仅对企业治理、人工智能系统及

其应用进行审计，而且还协助政府制定政策。

有关多利益相关者人工智能治理格局的具体原则已经提出。值得注意的是，Zhang^[768]提出了一种三层次的方法来治理生成式人工智能，其中基础模型、专业模型（在特定领域专业化的模型，可能通过微调从基础模型中派生出来）和服务应用（基于模型构建的应用）应分别采用量身定制的措施来进行治理，以满足它们各自的需求和风险。另一方面，Brundage et al.^[769]主张实施机构、软件和硬件三级治理，使关于人工智能系统安全性的声明更具可验证性。

政府 根据 Anderljung et al.^[112]，政府监管需要三个基本要素：(1) 标准开发流程，以确定适当的要求，供前沿人工智能开发者参考，(2) 登记和报告要求，为监管者提供对先进人工智能开发流程进展的见解，(3) 保证前沿人工智能模型的开发和部署符合安全标准的机制。

目前，全球范围内正在出现一系列新兴的政府法律法规，包括欧盟的人工智能法案^[770]，和美国两党人工智能法案框架^[771]。这些法规对于人工智能系统的安全性和对齐是必不可少的^[772]。

行业和 AGI 实验室 行业和 AGI 实验室的治理工作应强调在人工智能系统生命周期内进行全面的人工智能风险评估。基于 Koessler et al.^[110]，Schuett et al.^[111]的讨论，完整的人工智能风险评估周期可以分为五个阶段。开发前风险评估，预训练风险评估，和预部署风险评估都包括使用各种工具进行影响和风险的预测和分析，但随着时间的推移，细节、清晰度和复杂性逐渐增加^[110]。部署后监控是建立监控机制的阶段，所有先前的分析在部署后都会持续更新^[110]。外部审查包括漏洞悬赏计划^[111]，外部红队和第三方模型审计^[111-112]。

采取安全措施应对与人工智能系统相关的风险似乎被人工智能公司及相关从业者广泛接受。根据 Schuett et al.^[111]的研究，98% 的受访者在被调查时表示，他们一定程度上支持或强烈支持 AGI 实验室进行部署前风险评估、有害能力评估、第三方模型审计、模型使用的安全限制以及红队测试，以确保人工智能的安全性。同时，包括亚马逊、Anthropic、谷歌、Inflection、Meta、微软和 OpenAI 在内的领先人工智能公司已自愿向政府承诺实施安全措施^[773]。

值得注意的是，许多研究人员已经提出暂停先进人工智能系统的开发，以赢得更多的时间进行安全研究、风险评估和监管准备^[774]。他们的提议包括全面暂停所有足够先进的系统^[774]，以及根据特定故障模式的评估结果有条件地暂停特定类别的模型^[775]，其中包括目前采用的负责任的规模化政策 (RSP)^[776]。

第三方 Mökander et al.^[777]提出了第三方审计的三个关键功能：(1) 治理审计 (对设计和传播 LLM 的技术提供商);(2) 模型审计 (对 LLM 在预训练后，但在发布之前);(3) 应用审计 (对基于 LLM 的应用)。值得注意的是，这种分类部分与 Zhang^[768]提出的三层方法重叠。

现有的第三方审计的一个突出例子是 Alignment Research Center 的项目 ARC Evals^[93,609]，该中心与 OpenAI 合作对 GPT-4 进行了红队测试^[25]，并与 Anthropic 合作对 Claude 2 进行了红队测试^[322]。这些工作包括对毒性和偏见以及前沿人工智能风险 (如自主复制、操纵、网络安全和生物武器风险) 的评估^[25,21]。

除了审计之外，第三方还可以通过其他方式支持人工智能治理，例如协助制定政策和促进国际合作^[138]。Maas^[778]认为政府应优先考虑技术中立的规则，而不是技术特定的规则。AI4People 的良好人工智能社会的道德框架：机会、风险、原则和建议^[779]，由 AI4People 发布，指导了 2019 年 4 月提出的可信赖人工智能的道德指南^[780]。世界经济论坛 (WEF) 召集了政府官员、企业和公民社会，并与合作组织一起启动了全

球人工智能行动联盟，旨在促进人工智能领域的国际合作。^[781]

5.3 开放性问题

在现有的人工智能治理领域中，存在许多开放性问题。这些问题往往没有明确的答案，而对这些问题的讨论往往可以促进更好的治理。为了达成有效的人工智能治理，本文主要讨论国际治理和开源治理，希望通过本文的讨论促进人工智能的安全发展。

5.3.1 国际治理

在人工智能技术迅速发展并广泛使用的背景下，国际人工智能治理的需求已成为当务之急。关键讨论围绕着建立全球人工智能治理框架的必要性、确保其规范性^[782]和合法性^[783]等重要问题展开。这些主题在考虑中呈现出越来越详细和复杂的层次。此外，正如联合国秘书长安东尼奥·古特雷斯在七月的一次安理会会议中所述，生成式人工智能在规模上具有巨大的积极和消极影响的潜力，而不采取行动来减轻人工智能风险将是对我们保护当前和未来世代福祉责任的严重疏忽^[784]，国际治理也具有代际影响。因此，本节从三个方面考察国际人工智能治理的重要性和可行性：管理全球灾难性人工智能风险、管理人工智能中的机遇和国际治理的历史和当前成就，并在考察中贯彻代际及跨代际的视角。本文旨在为国际人工智能治理未来的结构贡献创新思考。

管理全球性的人工智能灾难风险 人工智能技术的持续进步为全球的发展和繁荣带来了巨大的潜力^[785]。然而，它们也不可避免地带来了潜在的风险。市场上无节制的竞争和地缘政治因素可能导致先进人工智能系统的过早开发和部署，从而产生负面的全球影响^[786]。人工智能系统中根深蒂固的种族和性别偏见可能会被放大并导致代际性的道德歧视^[787]。由于这些风险是国际性的和代际性的，国际治理的干预似乎可以缓解这些全球人工智能的灾难性挑战。例如，国家之间的共识可以帮助化解潜在的人工智能军备竞赛，而全行业的协议可以防止先进人工智能系统的草率和不负责任的开发，从而保障人工智能的长期和可持续发展^[138]。

管理人工智能中的机遇 人工智能发展带来的机遇并没有平等地分布，这可能导致不同地区之间持久的数字不平等，并危及人工智能发展的可持续性。人工智能进展中的地理差异将导致其经济和社会效益的不公平分配，可能会排除发展中国家或特定群体的这些优势^[138,786]。此外，技术领域决策权在少数个体中的巩固^[788-789]也可能产生不良的代际影响。国际治理可以缓解这种权力分配的不平等。通过人工智能的传播、教育和基础设施发展^[790]促进的人工智能机遇的国际共识和协调行动，可以确保从人工智能带来的技术红利平衡分配，并促进其持续发展的可持续性。

国际治理的历史和当前成就 在人工智能技术蓬勃发展之前，国际社会已经建立了针对有影响力的技术和重要事项的合作监管相关的框架。例如，气候变化政府间专门委员会（IPCC）召集专家评估气候环境问题，促进科学共识^[138]。国际民航组织（ICAO）制定并监督国际法规，同时评估成员国对这些法律的遵守情况^[138]。国际原子能机构（IAEA）推动了核能的和谐利用，具有全球影响力和先进的监测和评估机制^[791]。现今，多个国际组织已经在人工智能治理方面达成共识。2019年，G20成员国联合发布了一份关注以人为中心的人工智能原则的部长宣言^[792]。与此同时，经济合作与发展组织（OECD）制定了《OECD人工智能原则》^[793]。IEEE标准协会启动了一个全球倡议，旨在确保所有参与自主和智能系统设计与实施的利益相关方接受适当的教育、培训和激励，强调伦理关切，从而推动这些技术造福人类^[749]。2021年，联合国教

育、科学和文化组织 (UNESCO) 制定了有关人工智能伦理的全球标准^[794]，旨在为使人工智能系统为人类和社会谋福祉打下基础，并防止由于失控而可能造成的潜在危害。学术界还提出了人工智能的前瞻性国际治理框架，如国际人工智能组织 (IAIO)^[795]。我们希望这些先例和研究成果能够激发并为未来建立一个稳健且持久的国际人工智能治理框架提供基础。

5.3.2 开源治理

现有人工智能模型开源的问题在人工智能治理领域颇有争议，尤其是随着这些模型的能力不断增强^[137]。与将这些模型开源相关的潜在安全风险仍然是人工智能研究人员和政策制定者之间争论的焦点。在开源人工智能治理中的攻守平衡也仍然备受争议^[796]。对于开源模型是否会提高模型安全性还是增加滥用风险仍然存在争论。正如 Shapiro et al.^[797]所指出的，透明度的有效性取决于潜在攻击者已经拥有的知识的可能性，以及政府将透明度转化为识别和解决新出现的漏洞的能力。如果无法在人工智能系统的攻防之间建立适当的平衡，开源可能会潜在地带来人工智能系统滥用的重大风险。

为了准确和清晰，本文遵循 Seger et al.^[137]中对开源模型的定义：允许公开和公共访问模型的架构和权重，并允许任何人进行修改、研究、进一步开发和利用。目前，最为公认的开源模型包括 Llama2^[273]、Falcon^[798]、Vicuna^[398]等。本节主要评估开源模型的安全优势和潜在威胁，以促进关于开源这些模型的可行性讨论。最终，本文的目标是整合现有研究的观点，提出对未来开源方法的建议，以确保模型开源的安全实施。

支持开源的论点 支持现有模型开源的观点认为，这种方式可以通过以下几种方式减轻模型中固有的安全风险：

- **开源可能增强模型的安全性**：Meta 在其发布 Llama2 的博客中的声明^[280]支持了这种观点，即开源可以使开发者和技术社区对模型进行测试。因此，这种对问题的快速识别和解决可以大大增强模型的安全性。相对的，另一种观点认为，开源现有模型可能会增强对相关风险的认识，从而促进对这些潜在风险的更多的关注和研究^[799]。
- **开源可以促进权力和控制的分散化**：开源被广泛认为是一种有效的策略，可以减少主要人工智能实验室在经济、社会和政治等领域的主导地位。^[137] Stability 开源 Stable Diffusion 的核心原因就是例子：他们相信个人和社区，而不是让一个集中的、未经选举的实体控制人工智能技术^[800]。此外，一些评论家将模型开源与启蒙时代相提并论，认为分散化的控制增强了对人类和社会力量和善意的信任^[801]，出于安全目的实施集中治理可能反而会增强人工智能技术社区的权力。

反对开源的论点 从以下几个角度评估开源模型可能被滥用的潜力，开源模型的批评者提出了反对意见：

- **开源模型可能被微调为有害模型**：当前的研究严格证实，一些人工智能系统，与其最初的设计意图——减轻化学或生物学中的毒性——相反，现在有可能制造新的化学毒素^[760]和生物武器^[206]。这种模型的恶意微调可能导致深远的安全风险。此外，一旦进行了精细调整，语言模型可以模拟熟练的写手，产生令人信服的虚假信息，这可能导致相当大的社会政治风险。^[802]
- **无意中鼓励系统越狱**：研究表明，对开源模型权重的无限制访问可能促使绕过系统安全措施的行为^[137]。这一前提被 Zou et al.^[72]所证明，他们通过使用 Vicuna-7B 和 13B^[398]开发攻击后缀。一旦这些后缀

在像 ChatGPT^[25]、Bard^[803] 和 Claude^[322] 这样的易于访问的接口中实施，将产生违反人类意图的生成结果。因此，开源一个模型可能无意中破坏那些未开源模型的保护协议，从而增加模型被滥用的可能性。

关于开源治理的初步结论 关于人工智能模型的开源问题的争论仍未产生共识，目前主流的观点是，人工智能模型的公开并不会在目前带来重大风险。本文的论述不仅综合了这个论题上现有的观点，也为未来考虑是否开源更先进的人工智能系统的讨论做好了准备。

现有的关于开源先进人工智能系统的指导方针包括通过量化微调滥用的可能性来评估风险，以及逐步发布模型^[804,137] 等措施。同时，政策制定者正在为这些开源模型建立严格的合规协议。例如，欧洲的政策制定者坚持认为，模型应该在其生命周期中具有“性能、可预测性、可解释性、可纠正性、安全性和网络安全性。”^[805]。

6 结论

在这篇综述中，本文对人工智能对齐进行了全面的介绍，人工智能对齐的目标是构建行为符合人类意图和价值观的人工智能系统。本文将对齐的目标归纳为鲁棒性、可解释性、可控性和道德性 (RICE)，并将对齐方法的范围划分为前向对齐 (通过对齐训练使人工智能系统对齐) 和后向对齐 (获取人工智能系统对齐的证据，并适当地对其进行管理，以避免加剧对齐风险)。目前，前向对齐的两个显著研究领域是从反馈中学习和在分布偏移下学习，而后向对齐则包括对齐保证和人工智能治理。

与许多其他领域相比，人工智能对齐的一个特点是其多样性^[806] – 它是多个研究方向和方法的紧密组合，通过共享的目标而非共享的方法论将其联系在一起。这种多样性带来了好处。它通过让不同的方向进行竞争和冲突，促进了创新和思想的交叉传播。它还允许不同的研究方向互相补充，共同服务于对齐的目标；这体现在对齐循环 (见图2)，其中四个支柱被整合成一个自我改进的循环，不断提高人工智能系统的对齐性。同时，这种研究方向的多样性提高了进入这个领域的门槛，这就需要编制组织良好的调查材料，既服务于新人，也服务于有经验的研究人员。在这篇综述中，本文试图通过提供全面和最新的对齐概述来解决这个需求。

本文试图通过采用广泛且包容的对齐特征来考虑到该领域内的全部多样性。本文的对齐综述几乎关注了这个领域的所有主要研究议程，以及对齐保证和人工智能治理方面的实际实践。本文认识到对齐的边界往往是模糊的，并且有待争议。因此，在提出RICE原则时，本文用对齐的广泛特征作为明确的分类标准。同时，本文认识到维护这样的全面性综述需要长期的努力，并不断地进行审查和更新。对齐的问题和方法都紧密跟随机器学习的发展。这种快速的发展意味着新的材料和框架在短短几年后就可能过时。这就是为什么本文撰写这篇综述以反映最新的发展，并且也需要持续的维护和更新。

本文通过展望未来并展示我们认为的人工智能对齐领域未来需要解决的关键问题来结束这篇综述。

开放式探索新的挑战和方法 许多对齐讨论都建立在经典文献之上，这些文献早于最近的大语言模型和大规模深度学习的其他突破。因此，当这种范式转变发生在机器学习领域时，有一些对齐的挑战可能变得不那么突出，而其他的则变得更为突出；毕竟，科学理论的一个定义特征就是其可被证伪性^[807]。更重要的是，这种机器学习方法的转变和人工智能系统越来越紧密地融入社会的更广泛趋势^[808]，引入了以前无法预见的新挑战。这就要求我们进行开放式探索，积极寻找以前被忽视的新挑战。此外，这种探索不必局限于挑战

– 对于方法和解决方案，我们也应该采取类似的心态，从而为问题和答案构建更多样化的组合^[809]。

结合前瞻性和现状导向的视角 对齐强调了潜在的高级人工智能系统可能带来的危害，这些系统的能力超过了当前的系统^[53]。这些系统可能会在未来的某个时候出现，也可能只有几年的时间^[57]。前一种可能性要求本文研究推测的趋势和假设的情景^[252]。而后一种可能性强调了需要进行实地的努力，与现有的人工智能治理机构合作，并使用当前的系统作为更先进系统的原型^[150]。

强调政策相关性 对齐研究并不是在真空中进行，而是在一个生态系统中进行^[810]，³⁷研究人员、行业参与者、政府和非政府组织都应参与其中。因此，服务于人工智能对齐和安全生态系统需求的研究将是有用的。这些需求包括解决各种治理方案的关键障碍，例如，极端风险评估^[21]、计算治理的基础设施^[811]以及关于人工智能系统的可验证声明的机制^[769]。

强调社会复杂性和道德价值 随着人工智能系统越来越多地融入社会^[808]，对齐不再只是一个单一层次问题，而成为一个社会问题。这里，社会的含义有三层。

- (1) 在多智能体环境中进行对齐研究，这涉及到多个人工智能系统和多个人之间的交互^[61]。
- (2) 将人类的道德和社会价值纳入对齐（参见 §1.2.2 和 §4.3），这与机器伦理学和价值对齐领域密切相关^[65,162]。
- (3) 建模和预测人工智能系统对社会的影响，这需要方法来处理社会系统的复杂性，包括社会科学中的那些问题。可能有用的方法包括社会模拟^[188-189,192] 和博弈论^[726,61]。

致谢 本文感谢 *David Krueger*、*Anca Dragan*、*Alan Chan*、*Haoxing Du* 和 *Lawrence Chan* 对本综述提供的有益且建设性的反馈。本文感谢 *Yi Qu* 对本综述图表的提炼和改进。

³⁷请参见 aisafety.world 查看对齐组织的景观地图。

7 中英文词汇对照表

AI Alignment	人工智能对齐
Alignment Training	对齐训练
Adversarial Training	对抗训练
Allport-Vernon-Lindzey value system	奥尔波特-弗农-林赛价值观系统
Algorithmic Intervention	算法干预
Assurance	对齐保证
Avoiding Negative Effects	副作用避免
AI Governance	人工智能治理
Addressing Social Complexity	处理社会复杂性
Alignment Refinement	对齐精炼
Auto-induced Distribution Shift	自诱发分布偏移
Agents	自主体/智能体
Approval-Directed Agents	评价导向型自主体
AI Safety Beyond Alignment	对齐外的人工智能安全性
Adversarial Training	对抗训练
Ad-hoc Coordination	特设协调
Activation Patching	激活修补
Attribution Patching	归因修补
Agency	自主性/能动性
Backward Alignment	后向对齐
Barrier Loss	障碍损失
Behavior Cloning (BC)	行为克隆
Broadly-scoped Goals	广泛目标
Bounded Rationality	有限理性
Controllability	可控性
Comparison	比较
Corrigibility	可纠正性
Curriculum Learning	课程学习
Concept-based Interpretability	基于概念的可解释性
Computational Social Choice	计算社会选择
Cooperative Training	合作训练
Collectively Harmful Behaviors	集体有害行为
Cross Examination	交叉检验
Cross-Cultural Values in Social Psychology	社会心理学中的跨文化价值观
Capability Generalization	能力泛化
Cooperative Inverse Reinforcement Learning (CIRL)	合作逆强化学习
Cross-Distribution Aggregation	跨分布聚合
Connectivity based Fine-Tuning (CBFT)	基于连通性微调
Circuits Analysis	通路分析
Crowdsourced Adversarial Inputs	众包对抗输入
Circuits Hypothesis	通路假设

Data Distribution Intervention	数据分布干预
Deception	欺骗
Demonstration	示范
Deontic Logic	义务逻辑
Double Edge Components	双刃剑组件
Distribution Shift	分布偏移
Distributionally Robustness Optimization (DRO)	分布鲁棒优化
Domain Randomization	领域随机化
Discriminator-Critique Gap (DCG)	判别器-评价器差异
Ethicality	道德性
Ethics Shaping	伦理塑造
Event Calculus	事件演算
Empirical Risk Minimization (ERM)	经验风险最小化
Environment Building	环境搭建
Explainability and Transparency	可解释性和透明度
Forward Alignment	前向对齐
Feature Synthesis	特征合成
Feature Attribution	特征归因
Feedback	反馈
Fully Cooperative MARL	完全合作多智能体强化学习
Formal Machine Ethics	形式化机器伦理
Goal Misgeneralization	目标错误泛化
Goal Misspecification	目标错误规范
Goal Generalization	目标泛化
Government	政府
Goodhart's Law	古德哈特定律
Human Value Compliance	人类价值契合度
Human Value Verification	人类价值契合性验证
human thumbs-up	人类所赞同的
Human Value Verification	人类价值验证
Industry Actors	产业参与者
Instrumental Convergence	工具性收敛
Industry and AGI Labs	业界和 AGI 实验室
Iterated Distillation and Amplification (IDA)	迭代蒸馏扩增
Invariant Risk Minimization (IRM)	不变风险最小化
Invariant Causal Prediction (ICP)	不变风险预测
Interpretability	可解释性
Intentional Behaviors	有意行为
Intrinsic Interpretability	内在可解释性
International Governance	国际治理
Induction Head	归纳头
Inverse Reinforcement Learning (IRL)	逆强化学习

Instrumental Goals/ Strategies	工具目标/策略
Inner Misalignment	内部不对齐
Large Language Models (LLMs)	大语言模型
Learning from Feedback	从反馈中学习
Learning under Distribution Shift	在分布偏移下学习
LLMs-based agents	基于大语言模型的自主体
Latent Direction	潜在方向
Latent Knowledge	潜在知识
Learning from Demonstrations	从示范中学习
Misalignment	对齐失败
Manipulation	操纵
Mesa-optimization Objectives	内优化目标
Machine Ethics	机器伦理
Misspecified Reward	误设奖励
Measurement Tampering	度量篡改
Mode Connectivity	模式连通性
Minimizers	最小化器
Mixed-Motive MARL	混合动机多智能体强化学习
Mechanistic Interpretability	机制解释性
Manual and Automatic Jailbreaking	手动和自动越狱
Machine Ethics	机器伦理
Misuse Risk	滥用风险
Navigation via Mode Connectivity	模式连接指引
Non-Government Organizations (NGOs)	非政府组织
Non-Profit Organizations (NPOs)	非营利组织
Other Play	他人游戏
Off-belief Learning	离信念学习
Open-source Governance	开源治理
Outer Misalignment	外部不对齐
Off-switch Game	关机游戏
Out of Distribution (OOD)	分布外
Power Seeking	权力寻求
Proxy	代理
Policy Learning	策略学习
Perturbation-based Adversarial Training	基于扰动的对抗训练
Preference Elicitation	偏好引导
Probing	探测
Post Hoc Interpretability	事后可解释性
Proximal Policy Optimization (PPO)	近端策略优化
Policy-conditioned Belief	策略条件信念
Preference-based Reinforcement Learning	基于偏好的强化学习
Robustness	鲁棒性

Reinforcement Learning from Human Feedback (RLHF)	从人类反馈中的强化学习
Reinforcement Learning from Human and AI Feedback (RLHAIF)	基于人类和人工智能反馈的强化学习
RLxF	基于任意反馈的强化学习
Reinforcement Learning from Privacy Feedback (RLPF)	从隐私反馈中进行强化学习
Reward Hacking	奖励破解
Red Teaming	红队测试
Rule of Thumb (RoT)	经验法则
Representation Engineering	表示工程
Reward Misspecification	奖励错误规范
Reward	奖励
Recursive Reward Modeling (RRM)	递归奖励建模
Reject Sampling	拒绝采样
Relations between pairs of items	项对之间的关系
Reward Sketching	奖励速写
Risk Extrapolation	风险外推
Reinforced, Optimized, Guided, or Reverse Context Generation	基于强化学习、优化方法、引导生成或反向生成的上下文构造
Risk Management System (RMS)	风险管理系统
Reinforcement Learning (RL)	强化学习
Stackelberg Games	斯塔克尔伯格博弈
Scalable Oversight	可扩展监督
Safety Evaluation	安全评估
Situation Awareness	态势感知
Sycophancy	谄媚 (阿谀奉承)
Social Choices	社会选择
Specification Gaming	规范博弈
Sandbagging	故意失误
Shortcut features	捷径特征
Spurious Correlations	虚假关联
Social Value Orientation (SVO)	社会价值取向
Socially Realistic Settings	社会模拟
Social Concerns	社会关切
Safety Evaluations	安全测评
Safety or the Science of Deep & Learning	安全性或深度学习的科学
Self Preservation & Proliferation	自我保护与扩散
The Multi-Stakeholder Approach	多利益相关者方法
Transformer	Transformer
Third Parties	第三方
Untruthful Answers	不真实回答
Unrestricted Adversarial Training	无限制对抗训练
Violation of Ethics	违反伦理
Value Factorization	价值因子化
Value Alignment	价值对齐

Weighted Pairwise disagreement loss	加权失配损失
Zero-Shot Coordination	无准备协调

参考文献

- [1] CAIS. Center for ai safety: Statement on ai risk[Z]. 2023.
- [2] XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: A survey[A]. 2023.
- [3] WANG L, MA C, FENG X, et al. A survey on large language model based autonomous agents[A]. 2023.
- [4] DEGRAVE J, FELICI F, BUCHLI J, et al. Magnetic control of tokamak plasmas through deep reinforcement learning[J]. *Nature*, 2022, 602(7897): 414-419.
- [5] TURNER A, SMITH L, SHAH R, et al. Optimal policies tend to seek power[C/OL]//RANZATO M, BEYGELZIMER A, DAUPHIN Y, et al. *Advances in Neural Information Processing Systems: Vol. 34*. Curran Associates, Inc., 2021: 23063-23074. https://proceedings.neurips.cc/paper_files/paper/2021/file/c26820b8a4c1b3c2aa868d6d57e14a79-Paper.pdf.
- [6] PEREZ E, RINGER S, LUKOSIUTE K, et al. Discovering language model behaviors with model-written evaluations [C/OL]//ROGERS A, BOYD-GRABER J L, OKAZAKI N. *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 2023: 13387-13434. <https://doi.org/10.18653/v1/2023.findings-acl.847>.
- [7] CARROLL M, CHAN A, ASHTON H, et al. Characterizing Manipulation from AI Systems[M/OL]. arXiv, 2023 [2023-03-25]. <http://arxiv.org/abs/2303.09387>.
- [8] STEINHARDT J. Emergent Deception and Emergent Optimization[J/OL]. *Bounded Regret*, 2023[2023-03-25]. <https://bounded-regret.ghost.io/emergent-deception-optimization/>.
- [9] SHARMA M, TONG M, KORBAK T, et al. Towards understanding sycophancy in language models[A]. 2023.
- [10] PARK P S, GOLDSTEIN S, O'GARA A, et al. Ai deception: A survey of examples, risks, and potential solutions [A]. 2023.
- [11] BOSTROM N. Existential risk prevention as global priority[J]. *Global Policy*, 2013, 4(1): 15-31.
- [12] ORD T. *The precipice: Existential risk and the future of humanity*[M]. Hachette Books, 2020.
- [13] CHRISTIAN B. *The alignment problem: Machine learning and human values*[M]. WW Norton & Company, 2020.
- [14] BUCKNALL B S, DORI-HACOHEN S. Current and near-term ai as a potential existential risk factor[C]// *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022: 119-129.
- [15] LEIKE J, KRUEGER D, EVERITT T, et al. Scalable agent alignment via reward modeling: a research direction [A]. 2018.
- [16] PAN A, BHATIA K, STEINHARDT J. The effects of reward misspecification: Mapping and mitigating misaligned models[C/OL]//*The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. <https://openreview.net/forum?id=JYtwGwIL7ye>.
- [17] SHAH R, VARMA V, KUMAR R, et al. Goal misgeneralization: Why correct specifications aren't enough for correct goals[A]. 2022.
- [18] BERGLUND L, STICKLAND A C, BALESNI M, et al. Taken out of context: On measuring situational awareness in llms[A]. 2023.

-
- [19] NGO R, CHAN L, MINDERMANN S. The alignment problem from a deep learning perspective[A]. 2022.
- [20] HUBINGER E, VAN MERWIJK C, MIKULIK V, et al. Risks from learned optimization in advanced machine learning systems[A]. 2019.
- [21] SHEVLANE T, FARQUHAR S, GARFINKEL B, et al. Model evaluation for extreme risks[A]. 2023.
- [22] SMOLENSKY P. Connectionist ai, symbolic ai, and the brain[J]. *Artificial Intelligence Review*, 1987, 1(2): 95-109.
- [23] GOEL A. Looking back, looking ahead: Symbolic versus connectionist ai[J]. *AI Magazine*, 2022, 42(4): 83-85.
- [24] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. *nature*, 2015, 518(7540): 529-533.
- [25] OPENAI. Gpt-4 technical report[A]. 2023. arXiv: 2303.08774.
- [26] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. *nature*, 2017, 550(7676): 354-359.
- [27] BERNER C, BROCKMAN G, CHAN B, et al. Dota 2 with large scale deep reinforcement learning[A]. 2019.
- [28] KAUFMANN E, BAUERSFELD L, LOQUERCIO A, et al. Champion-level drone racing using deep reinforcement learning[J]. *Nature*, 2023, 620(7976): 982-987.
- [29] RUFF K M, PAPPU R V. Alphafold and implications for intrinsically disordered proteins[J]. *Journal of Molecular Biology*, 2021, 433(20): 167208.
- [30] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [C/OL]//KOYEJO S, MOHAMED S, AGARWAL A, et al. *Advances in Neural Information Processing Systems: Vol. 35*. Curran Associates, Inc., 2022: 24824-24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
- [31] WANG Y, KORDI Y, MISHRA S, et al. Self-instruct: Aligning language models with self-generated instructions [C/OL]//ROGERS A, BOYD-GRABER J L, OKAZAKI N. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 2023: 13484-13508. <https://doi.org/10.18653/v1/2023.acl-long.754>.
- [32] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. *Advances in neural information processing systems*, 2020, 33: 1877-1901.
- [33] ASKELL A, BAI Y, CHEN A, et al. A general language assistant as a laboratory for alignment[A]. 2021.
- [34] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[A]. 2020.
- [35] SRIVASTAVA A, RASTOGI A, RAO A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models[A]. 2022.
- [36] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models[A]. 2022.
- [37] BANG Y, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity[A]. 2023.
- [38] Jacob Steinhardt. Emergent deception and emergent optimization[EB/OL]. 2023. <https://bounded-regret.ghost.io/emergent-deception-optimization/>.

-
- [39] CHAN A, SALGANIK R, MARKELIUS A, et al. Harms from increasingly agentic algorithmic systems[C]// Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023: 651-666.
- [40] BUBECK S, CHANDRASEKARAN V, ELDAN R, et al. Sparks of artificial general intelligence: Early experiments with gpt-4[A]. 2023.
- [41] MANYIKA J, CHUI M, MIREMADI M, et al. A future that works: Ai, automation, employment, and productivity [J]. McKinsey Global Institute Research, Tech. Rep, 2017, 60: 1-135.
- [42] WEST D M. The future of work: Robots, ai, and automation[M]. Brookings Institution Press, 2018.
- [43] FURMAN J, SEAMANS R. Ai and the economy[J]. Innovation policy and the economy, 2019, 19(1): 161-191.
- [44] KORINEK M A, SCHINDLER M M, STIGLITZ J. Technological progress, artificial intelligence, and inclusive growth[M]. International Monetary Fund, 2021.
- [45] CRITCH A, RUSSELL S. Tasra: A taxonomy and analysis of societal-scale risks from ai[A]. 2023.
- [46] HENDRYCKS D, MAZEIKA M. X-risk analysis for ai research[A]. 2022.
- [47] NTOUTSI E, FAFALIOS P, GADIRAJU U, et al. Bias in data-driven artificial intelligence systems—an introductory survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, 10(3): e1356.
- [48] BOSTROM N. Superintelligence: Paths, dangers, strategies[M]. Oxford University Press, Oxford, 2014.
- [49] BENGIO Y. How rogue ais may arise[EB/OL]. 2023. <https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/>.
- [50] BOSTROM N, CIRKOVIC M M. Global catastrophic risks[M]. Oxford University Press, USA, 2011.
- [51] HENDRYCKS D, MAZEIKA M, WOODSIDE T. An overview of catastrophic ai risks[A]. 2023.
- [52] GOV.UK. Frontier ai: capabilities and risks –discussion paper[EB/OL]. 2023. <https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper>.
- [53] NGO R. Agi safety from first principles[EB/OL]. 2020. <https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ>.
- [54] HENDRYCKS D. Natural selection favors ais over humans[A]. 2023. arXiv: 2303.16200.
- [55] CHRISTIANO P. What failure looks like[J/OL]. AI Alignment Forum, 2019. <https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like>.
- [56] KENTON Z, SHAH R, LINDNER D, et al. Threat model literature review[J/OL]. AI Alignment Forum, 2022. <https://www.alignmentforum.org/posts/wnnkD6P2k2TfHnNmt/threat-model-literature-review>.
- [57] STEIN-PERLMAN Z, WEINSTEIN-RAUN B, GRACE K. expert survey on progress in ai[J]. AI Impacts. Available online at: <https://aiimpacts.org/2022-expert-survey-on-progress-in-ai> (accessed December 7, 2022), 2022.
- [58] MICHAEL J, HOLTZMAN A, PARRISH A, et al. What do NLP researchers believe? results of the NLP community metasurvey[C/OL]//ROGERS A, BOYD-GRABER J L, OKAZAKI N. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. Association for Computational Linguistics, 2023: 16334-16368. <https://doi.org/10.18653/v1/2023.acl-long.903>.

-
- [59] SI W M, BACKES M, BLACKBURN J, et al. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots[C]//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2022: 2659-2673.
- [60] PAN A, CHAN J S, ZOU A, et al. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark.[J]. ICML, 2023.
- [61] CRITCH A, KRUEGER D. Ai research considerations for human existential safety (arches)[A]. 2020.
- [62] HENDRYCKS D, CARLINI N, SCHULMAN J, et al. Unsolved problems in ml safety[A]. 2021.
- [63] OPENAI. Introducing superalignment[EB/OL]. 2023. <https://openai.com/blog/introducing-superalignment/>.
- [64] KENTON Z, EVERITT T, WEIDINGER L, et al. Alignment of language agents[A]. 2021.
- [65] GABRIEL I. Artificial intelligence, values, and alignment[J]. Minds and machines, 2020, 30(3): 411-437.
- [66] DIETTERICH T G. Steps toward robust artificial intelligence[J]. Ai Magazine, 2017, 38(3): 3-24.
- [67] RUDNER T, TONER H. Key concepts in ai safety: Robustness and adversarial examples[EB/OL]. 2021. <https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/>.
- [68] TALEB N N. The black swan: The impact of the highly improbable: Vol. 2[M]. Random house, 2007.
- [69] SONG Y, SHU R, KUSHMAN N, et al. Constructing unrestricted adversarial examples with generative models[J]. Advances in Neural Information Processing Systems, 2018, 31.
- [70] CHAKRABORTY A, ALAM M, DEY V, et al. A survey on adversarial attacks and defences[J]. CAAI Transactions on Intelligence Technology, 2021, 6(1): 25-45.
- [71] CARLINI N, NASR M, CHOQUETTE-CHOO C A, et al. Are aligned neural networks adversarially aligned?[A]. 2023. arXiv: 2306.15447.
- [72] ZOU A, WANG Z, KOLTER J Z, et al. Universal and transferable adversarial attacks on aligned language models [A]. 2023.
- [73] SHAH R, FEUILLADE-MONTIXI Q, POUR S, et al. Scalable and transferable black-box jailbreaks for language models via persona modulation[A]. 2023. arXiv: 2311.03348.
- [74] STEINHARDT J, TONER H. Why robustness is key to deploying ai[EB/OL]. 2020. <https://www.brookings.edu/articles/why-robustness-is-key-to-deploying-ai>.
- [75] KIRILENKO A, KYLE A S, SAMADI M, et al. The flash crash: High-frequency trading in an electronic market [J]. The Journal of Finance, 2017, 72(3): 967-998.
- [76] OecdAI. Ai principles[EB/OL]. 2021. <https://oecd.ai/en/dashboards/ai-principles/P8>.
- [77] RUSSELL S. Human compatible: Artificial intelligence and the problem of control[M]. Penguin, 2019.
- [78] RÄUKER T, HO A, CASPER S, et al. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks[C]//2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE, 2023: 464-483.

-
- [79] TURPIN M, MICHAEL J, PEREZ E, et al. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting[A]. 2023.
- [80] HUBINGER E, VAN MERWIJK C, MIKULIK V, et al. Deceptive alignment[J/OL]. AI Alignment Forum, 2019. <https://www.alignmentforum.org/posts/zthDPAjh9w6Ytbeks/deceptive-alignment>.
- [81] CARRANZA A, PAI D, SCHAEFFER R, et al. Deceptive alignment monitoring[A]. 2023. arXiv: 2307.10569.
- [82] PACCHIARDI L, CHAN A J, MINDERMANN S, et al. How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions[A]. 2023.
- [83] RADHAKRISHNAN A, NGUYEN K, CHEN A, et al. Question decomposition improves the faithfulness of model-generated reasoning[A]. 2023.
- [84] CARROLL M, CHAN A, ASHTON H, et al. Characterizing manipulation from ai systems[A]. 2023.
- [85] ELHAGE N, NANDA N, OLSSON C, et al. A mathematical framework for transformer circuits[J]. Transformer Circuits Thread, 2021.
- [86] MENG K, BAU D, ANDONIAN A, et al. Locating and editing factual associations in gpt[J]. Advances in Neural Information Processing Systems, 2022, 35: 17359-17372.
- [87] HOLZINGER A, BIEMANN C, PATTICHIS C S, et al. What do we need to build explainable ai systems for the medical domain?[A]. 2017.
- [88] DeepMind. Building safe artificial intelligence: specification, robustness, and assurance[EB/OL]. 2018. <https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1>.
- [89] RUDNER T, TONER H. Key concepts in ai safety: Interpretability in machine learning[EB/OL]. 2021. <https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretability-in-machine-learning/>.
- [90] SOARES N, FALLENSTEIN B, ARMSTRONG S, et al. Corrigibility[C]//Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.
- [91] HADFIELD-MENELL D, DRAGAN A, ABBEEL P, et al. The off-switch game[A]. 2016.
- [92] UniteAI. What is ai capability control & why does it matter?[EB/OL]. 2023. <https://www.unite.ai/what-is-ai-capability-control-why-does-it-matter/>.
- [93] ARC Evals. Update on ARC's recent eval efforts[EB/OL]. 2023. <https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/>.
- [94] BOWMAN S R, HYUN J, PEREZ E, et al. Measuring progress on scalable oversight for large language models[A]. 2022.
- [95] BUOLAMWINI J, GEBRU T. Gender shades: Intersectional accuracy disparities in commercial gender classification [C]//Conference on fairness, accountability and transparency. PMLR, 2018: 77-91.
- [96] ZHANG B H, LEMOINE B, MITCHELL M. Mitigating unwanted biases with adversarial learning[C]//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018: 335-340.
- [97] NOBLE S U. Algorithms of oppression[M]//Algorithms of oppression. New York university press, 2018.

-
- [98] KEARNS M, ROTH A. The ethical algorithm: The science of socially aware algorithm design[M]. Oxford University Press, 2019.
- [99] RAJI I D, GEHRU T, MITCHELL M, et al. Saving face: Investigating the ethical concerns of facial recognition auditing[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020: 145-151.
- [100] BERK R, HEIDARI H, JABBARI S, et al. Fairness in criminal justice risk assessments: The state of the art[J]. Sociological Methods & Research, 2021, 50(1): 3-44.
- [101] HENDRYCKS D, BURNS C, BASART S, et al. Aligning ai with shared human values[J]. ICLR 2021, 2021.
- [102] HENDRYCKS D, MAZEIKA M, ZOU A, et al. What would jiminy cricket do? towards agents that behave morally[C/OL]//VANSCHOREN J, YEUNG S. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual. 2021. <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/39059724f73a9969845dfe4146c5660e-Abstract-round2.html>.
- [103] Collective Intelligence Project. Introducing the collective intelligence project[EB/OL]. 2023. <https://cip.org/whitepaper>.
- [104] HAGENDORFF T. The ethics of ai ethics: An evaluation of guidelines[J]. Minds and machines, 2020, 30(1): 99-120.
- [105] PANKOWSKA P K. Framework on ethical aspects of artificial intelligence, robotics and related technologies[J]. European Parliament, 2020.
- [106] WINFIELD A F, MICHAEL K, PITT J, et al. Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue][J]. Proceedings of the IEEE, 2019, 107(3): 509-517.
- [107] Asimov. Asimov's laws[EB/OL]. 1942. <https://webhome.auburn.edu/~vestmon/robotics.html>.
- [108] MEMARIAN B, DOLECK T. Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai), and higher education: A systematic review[J]. Computers and Education: Artificial Intelligence, 2023: 100152.
- [109] White House. Ensuring safe, secure, and trustworthy ai[EB/OL]. 2023. <https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf>.
- [110] KOESSLER L, SCHUETT J. Risk assessment at agi companies: A review of popular risk assessment techniques from other safety-critical industries[A]. 2023.
- [111] SCHUETT J, DREKSLER N, ANDERLJUNG M, et al. Towards best practices in agi safety and governance: A survey of expert opinion[A]. 2023.
- [112] ANDERLJUNG M, BARNHART J, LEUNG J, et al. Frontier ai regulation: Managing emerging risks to public safety[A]. 2023.
- [113] CHRISTIANO P F, LEIKE J, BROWN T, et al. Deep reinforcement learning from human preferences[J]. Advances in neural information processing systems, 2017, 30.
- [114] BAI Y, JONES A, NDOUSSE K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[A]. 2022.

-
- [115] PANDEY R, PUROHIT H, CASTILLO C, et al. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning[J]. *International Journal of Human-Computer Studies*, 2022, 160: 102772.
- [116] CASPER S, DAVIES X, SHI C, et al. Open problems and fundamental limitations of reinforcement learning from human feedback[A]. 2023.
- [117] TIEN J, HE J Z, ERICKSON Z, et al. Causal confusion and reward misidentification in preference-based reward learning[C/OL]//The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. https://openreview.net/pdf?id=R0Xxvr_X3ZA.
- [118] ANDERSON M, ANDERSON S L. *Machine ethics*[M]. Cambridge University Press, 2011.
- [119] TOLMEIJER S, KNEER M, SARASUA C, et al. Implementations in machine ethics: A survey[J]. *ACM Computing Surveys (CSUR)*, 2020, 53(6): 1-38.
- [120] SANTURKAR S, DURMUS E, LADHAK F, et al. Whose opinions do language models reflect?[C/OL]//KRAUSE A, BRUNSKILL E, CHO K, et al. *Proceedings of Machine Learning Research: Vol. 202 International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. PMLR, 2023: 29971-30004. <https://proceedings.mlr.press/v202/santurkar23a.html>.
- [121] KRUEGER D, MAHARAJ T, LEIKE J. Hidden incentives for auto-induced distributional shift[A]. 2020.
- [122] THULASIDASAN S, THAPA S, DHAUBHADEL S, et al. An effective baseline for robustness to distributional shift[C]//2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2021: 278-285.
- [123] HENDRYCKS D, BASART S, MU N, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8340-8349.
- [124] DI LANGOSCO L L, KOCH J, SHARKEY L D, et al. Goal misgeneralization in deep reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2022: 12004-12019.
- [125] PERDOMO J, ZRNIC T, MENDLER-DÜNNER C, et al. Performative prediction[C]//International Conference on Machine Learning. PMLR, 2020: 7599-7609.
- [126] KALIMERIS D, BHAGAT S, KALYANARAMAN S, et al. Preference amplification in recommender systems[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 805-815.
- [127] ADOMAVICIUS G, BOCKSTEDT J, CURLEY S, et al. Recommender systems, ground truth, and preference pollution[J]. *AI Magazine*, 2022, 43(2): 177-189.
- [128] KRUEGER D, CABALLERO E, JACOBSEN J H, et al. Out-of-distribution generalization via risk extrapolation (rex)[C]//International Conference on Machine Learning. PMLR, 2021: 5815-5826.
- [129] LUBANA E S, BIGELOW E J, DICK R P, et al. Mechanistic mode connectivity[C]//International Conference on Machine Learning. PMLR, 2023: 22965-23004.
- [130] BAI T, LUO J, ZHAO J, et al. Recent advances in adversarial training for adversarial robustness[C/OL]//ZHOU Z H. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conferences on Artificial Intelligence Organization*, 2021: 4312-4321. <https://doi.org/10.24963/ijcai.2021/591>.

-
- [131] DAFOE A, HUGHES E, BACHRACH Y, et al. Open problems in cooperative ai[A]. 2020.
- [132] DAFOE A, BACHRACH Y, HADFIELD G, et al. Cooperative ai: machines must learn to find common ground[J]. *Nature*, 2021, 593(7857): 33-36.
- [133] Government of the United Kingdom. The roadmap to an effective ai assurance ecosystem - extended version[J/OL]. Transformer Circuits Thread, 2021[2021-8-12]. <https://www.gov.uk/government/publications/the-roadmap-to-a-n-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version>.
- [134] OLAH C, SATYANARAYAN A, JOHNSON I, et al. The Building Blocks of Interpretability[J/OL]. *Distill*, 2018, 3(3): e10[2023-08-17]. <https://distill.pub/2018/building-blocks>. DOI: 10.23915/distill.00010.
- [135] PEREZ E, HUANG S, SONG H F, et al. Red teaming language models with language models[C/OL]//GOLDBERG Y, KOZAREVA Z, ZHANG Y. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. Association for Computational Linguistics, 2022: 3419-3448. <https://doi.org/10.18653/v1/2022.emnlp-main.225>.
- [136] BURNS C, YE H, KLEIN D, et al. Discovering latent knowledge in language models without supervision[C/OL]//The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. <https://openreview.net/pdf?id=ETKGuby0hcs>.
- [137] SEGER E, DREKSLER N, MOULANGE R, et al. open-sourcing-highly-capable-foundation-models[EB/OL]. 2023. <https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models>.
- [138] HO L, BARNHART J, TRAGER R, et al. International institutions for advanced ai[A]. 2023.
- [139] HUBINGER E, VAN MERWIJK C, MIKULIK V, et al. The inner alignment problem[EB/OL]. 2019. <https://www.alignmentforum.org/posts/pL56xPoniLvtMDQ4J/the-inner-alignment-problem>.
- [140] KRAKOVNA V. Paradigms of ai alignment: components and enablers[EB/OL]. 2022. <https://vkrakovna.wordpress.com/2022/06/02/paradigms-of-ai-alignment-components-and-enablers/>.
- [141] PERRY L. Evan hubinger on inner alignment, outer alignment, and proposals for building safe advanced ai[EB/OL]. 2020. <https://www.alignmentforum.org/posts/qZGoHkRgANQpGHWnu/evan-hubinger-on-inner-alignment-outer-alignment-and>.
- [142] TURNER A. Inner and outer alignment decompose one hard problem into two extremely hard problems[EB/OL]. 2022. <https://www.alignmentforum.org/posts/gHefoxiznGfsbiAu9/inner-and-outer-alignment-decompose-one-hard-problem-into>.
- [143] AMODEI D, OLAH C, STEINHARDT J, et al. Concrete problems in ai safety[A]. 2016.
- [144] EVERITT T, LEA G, HUTTER M. Agi safety literature review[A]. 2018.
- [145] OMOHUNDRO S M. The basic ai drives[C]//AGI: Vol. 171. 2008: 483-492.
- [146] BOSTROM N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents[J]. *Minds and Machines*, 2012, 22: 71-85.
- [147] OESTERHELD C. Approval-directed agency and the decision theory of newcomb-like problems[J]. *Synthese*, 2021, 198(Suppl 27): 6491-6504.

-
- [148] CHRISTIANO P. Approval-directed agents[J/OL]. AI Alignment Forum, 2022. <https://www.alignmentforum.org/posts/7Hr8t6xwuuxBTqADK/approval-directed-agents-1>.
- [149] HADFIELD-MENELL D, HADFIELD G K. Incomplete contracting and ai alignment[C]//Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019: 417-422.
- [150] COTRA A. The case for aligning narrowly superhuman models[EB/OL]. 2021. <https://www.alignmentforum.org/posts/PZtsoaoSLpKjjbMqM/the-case-for-aligning-narrowly-superhuman-models>.
- [151] CHRISTIANO P, XU M, COTRA A. Arc's first technical report: Eliciting latent knowledge[J/OL]. AI Alignment Forum, 2021. <https://www.alignmentforum.org/posts/qHCDysDnvhteW7kRd/arc-s-first-technical-report-eliciting-latent-knowledge>.
- [152] HOBBAHN M. Eliciting latent knowledge (elk) - distillation/summary[J/OL]. AI Alignment Forum, 2022. <https://www.alignmentforum.org/posts/rxoBY9CMkqDsHt25t/eliciting-latent-knowledge-elk-distillation-summary>.
- [153] BENSON-TILSEN T, SOARES N. Formalizing convergent instrumental goals.[C]//AAAI Workshop: AI, Ethics, and Society. 2016.
- [154] EVERITT T, HUTTER M. Avoiding wireheading with value reinforcement learning[C]//Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings 9. Springer, 2016: 12-22.
- [155] SKALSE J, HOWE N, KRASHENINNIKOV D, et al. Defining and characterizing reward gaming[J]. Advances in Neural Information Processing Systems, 2022, 35: 9460-9471.
- [156] SOARES N, FALLENSTEIN B. Agent foundations for aligning machine intelligence with human interests: a technical research agenda[J]. The technological singularity: Managing the journey, 2017: 103-125.
- [157] DEMSKI A, GARRABRANT S. Embedded agency[A]. 2019.
- [158] SOARES N, FALLENSTEIN B. Toward idealized decision theory[A]. 2015.
- [159] SOARES N. The value learning problem[M]//Artificial intelligence safety and security. Chapman and Hall/CRC, 2018: 89-97.
- [160] GARRABRANT S, BENSON-TILSEN T, CRITCH A, et al. Logical induction[A]. 2016.
- [161] CRITCH A, DENNIS M, RUSSELL S. Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory[A]. 2022.
- [162] GABRIEL I, GHAZAVI V. The challenge of value alignment: From fairer algorithms to ai safety[A]. 2021.
- [163] YUAN L, GAO X, ZHENG Z, et al. In situ bidirectional human-robot value alignment[J]. Science robotics, 2022, 7(68): eabm4183.
- [164] MACINTYRE A. After virtue[M]. A&C Black, 2013.
- [165] YU H, SHEN Z, MIAO C, et al. Building ethics into artificial intelligence[C/OL]//LANG J. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. ijcai.org, 2018: 5527-5533. <https://doi.org/10.24963/ijcai.2018/779>.

-
- [166] ANDERSON M, ANDERSON S, ARMEN C. Towards machine ethics: Implementing two action-based ethical theories[C]//Proceedings of the AAAI 2005 fall symposium on machine ethics. 2005: 1-7.
- [167] ARKOUDAS K, BRINGSJORD S, BELLO P. Toward ethical robots via mechanized deontic logic[C]//AAAI fall symposium on machine ethics. Menlo Park, CA, USA: The AAAI Press, 2005: 17-23.
- [168] ANDERSON M, ANDERSON S L. The status of machine ethics: a report from the aai symposium[J]. *Minds and Machines*, 2007, 17: 1-10.
- [169] JIANG L, HWANG J D, BHAGAVATULA C, et al. Can machines learn morality? the delphi experiment[A]. 2021.
- [170] JIN Z, LEVINE S, GONZALEZ ADAUTO F, et al. When to make exceptions: Exploring language models as accounts of human moral judgment[J]. *Advances in neural information processing systems*, 2022, 35: 28458-28473.
- [171] VERMA S, RUBIN J. Fairness definitions explained[C]//Proceedings of the international workshop on software fairness. 2018: 1-7.
- [172] SAXENA N A, HUANG K, DEFILIPPIS E, et al. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness[C]//Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019: 99-106.
- [173] MEHRABI N, MORSTATTER F, SAXENA N, et al. A survey on bias and fairness in machine learning[J]. *ACM computing surveys (CSUR)*, 2021, 54(6): 1-35.
- [174] D'ALESSANDRO B, O'NEIL C, LAGATTA T. Conscientious classification: A data scientist's guide to discrimination-aware classification[J]. *Big data*, 2017, 5(2): 120-134.
- [175] BELLAMY R K, DEY K, HIND M, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias[A]. 2018.
- [176] BERK R, HEIDARI H, JABBARI S, et al. A convex framework for fair regression[A]. 2017.
- [177] XU D, YUAN S, ZHANG L, et al. Fairgan: Fairness-aware generative adversarial networks[C]//2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018: 570-575.
- [178] ALLPORT G W. *Becoming: Basic considerations for a psychology of personality*: Vol. 20[M]. Yale University Press, 1955.
- [179] MESSICK D M, MCCLINTOCK C G. Motivational bases of choice in experimental games[J]. *Journal of experimental social psychology*, 1968, 4(1): 1-25.
- [180] MCCLINTOCK C G, VAN AVERMAET E. Social values and rules of fairness: A theoretical perspective[J]. *Cooperation and helping behavior*, 1982: 43-71.
- [181] LIEBRAND W B. The effect of social motives, communication and group size on behaviour in an n-person multi-stage mixed-motive game[J]. *European journal of social psychology*, 1984, 14(3): 239-264.
- [182] VAN LANGE P A, DE BRUIN E, OTTEN W, et al. Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence.[J]. *Journal of personality and social psychology*, 1997, 73(4): 733.
- [183] MURPHY R O, ACKERMANN K A, HANDGRAAF M J. Measuring social value orientation[J]. *Judgment and Decision making*, 2011, 6(8): 771-781.

-
- [184] MURPHY R O, ACKERMANN K A. Social value orientation: Theoretical and measurement issues in the study of social preferences[J]. *Personality and Social Psychology Review*, 2014, 18(1): 13-41.
- [185] ROKEACH M. The nature of human values.[M]. Free press, 1973.
- [186] SCHWARTZ S H. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries[M]//*Advances in experimental social psychology*: Vol. 25. Elsevier, 1992: 1-65.
- [187] SCHWARTZ S H. Are there universal aspects in the structure and contents of human values?[J]. *Journal of social issues*, 1994, 50(4): 19-45.
- [188] BONABEAU E. Agent-based modeling: Methods and techniques for simulating human systems[J]. *Proceedings of the national academy of sciences*, 2002, 99(suppl_3): 7280-7287.
- [189] DE MARCHI S, PAGE S E. Agent-based models[J]. *Annual Review of political science*, 2014, 17: 1-20.
- [190] SERT E, BAR-YAM Y, MORALES A J. Segregation dynamics with reinforcement learning and agent based modeling[J]. *Scientific reports*, 2020, 10(1): 11771.
- [191] STORCHAN V, VYETRENKO S, BALCH T. Learning who is in the market from time series: market participant discovery through adversarial calibration of multi-agent simulators[A]. 2021.
- [192] PARK J S, O'BRIEN J C, CAI C J, et al. Generative agents: Interactive simulacra of human behavior[C/OL]// FOLLMER S, HAN J, STEIMLE J, et al. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*. ACM, 2023: 2:1-2:22. <https://doi.org/10.1145/3586183.3606763>.
- [193] CALVO R A, PETERS D, CAVE S. Advancing impact assessment for intelligent systems[J]. *Nature Machine Intelligence*, 2020, 2(2): 89-91.
- [194] FERNANDES P M, SANTOS F C, LOPES M. Adoption dynamics and societal impact of ai systems in complex networks[C]//*Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020: 258-264.
- [195] OSOBA O A, BOUDREAUX B, YEUNG D. Steps towards value-aligned systems[C]//*Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020: 332-336.
- [196] SEN A. Social choice theory[J]. *Handbook of mathematical economics*, 1986, 3: 1073-1181.
- [197] ARROW K J. Social choice and individual values: Vol. 12[M]. Yale university press, 2012.
- [198] BRANDT F, CONITZER V, ENDRISS U, et al. *Handbook of computational social choice*[M]. Cambridge University Press, 2016.
- [199] LEIKE J. A proposal for importing society' s values[EB/OL]. 2023. <https://aligned.substack.com/p/a-proposal-for-importing-societys-values>.
- [200] YAMAGATA T, MCCONVILLE R, SANTOS-RODRIGUEZ R. Reinforcement learning with feedback from multiple humans with diverse skills[A]. 2021. arXiv: 2111.08596.
- [201] BAKKER M, CHADWICK M, SHEAHAN H, et al. Fine-tuning language models to find agreement among humans with diverse preferences[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 38176-38189.

-
- [202] KÖPF A, KILCHER Y, VON RÜTTE D, et al. Openassistant conversations – democratizing large language model alignment[A]. 2023. arXiv: 2304.07327.
- [203] KENWARD B, SINCLAIR T R. Machine morality, moral progress, and the looming environmental disaster[J/OL]. *Cognitive Computation and Systems*, 2021, 3: 83-90. DOI: 10.1049/ccs2.12027.
- [204] CAO E, BAPTISTA E. 'deepfake' scam in china fans worries over ai-driven fraud[J/OL]. Reuters, 2023. <https://www.reuters.com/technology/deepfake-scam-china-fans-worries-over-ai-driven-fraud-2023-05-22/>.
- [205] SOICE E H, ROCHA R, CORDOVA K, et al. Can large language models democratize access to dual-use biotechnology?[A]. 2023. arXiv: 2306.03809.
- [206] SANDBRINK J B. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools[A]. 2023.
- [207] DANZIG R, HOSFORD Z, SAGEMAN M, et al. Aum shinrikyo: Insights into how terrorists develop biological and chemical weapons, second edition[R]. Center for a New American Security (CNAS), 2012.
- [208] MIRSKY Y, DEMONTIS A, KOTAK J, et al. The threat of offensive AI to organizations[J/OL]. *Comput. Secur.*, 2023, 124: 103006. <https://doi.org/10.1016/j.cose.2022.103006>. DOI: 10.1016/J.COSE.2022.103006.
- [209] GRANT N, WEISE K. In a.i. race, microsoft and google choose speed over caution[J]. *New York Times*, 2023.
- [210] ARMSTRONG S, BOSTROM N, SHULMAN C. Racing to the precipice: a model of artificial intelligence development: 2013-1[R]. Future of Humanity Institute, Oxford University, 2013: 1-8.
- [211] MARBACH P, TSITSIKLIS J N. Simulation-based optimization of markov reward processes[J]. *IEEE Transactions on Automatic Control*, 2001, 46(2): 191-209.
- [212] PUTERMAN M L. Markov decision processes: discrete stochastic dynamic programming[M]. John Wiley & Sons, 2014.
- [213] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. MIT press, 2018.
- [214] MOERLAND T M, BROEKENS J, PLAAT A, et al. Model-based reinforcement learning: A survey[J]. *Foundations and Trends® in Machine Learning*, 2023, 16(1): 1-118.
- [215] NG A Y, RUSSELL S, et al. Algorithms for inverse reinforcement learning.[C]//*Icml*: Vol. 1. 2000: 2.
- [216] SKALSE J M V, FARRUGIA-ROBERTS M, RUSSELL S, et al. Invariance in policy optimisation and partial identifiability in reward learning[C]//*International Conference on Machine Learning*. PMLR, 2023: 32033-32058.
- [217] EVERITT T, KRAKOVNA V, ORSEAU L, et al. Reinforcement learning with a corrupted reward channel[C/OL]//SIERRA C. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, Melbourne, Australia, August 19-25, 2017. ijcai.org, 2017: 4705-4713. <https://doi.org/10.24963/ijcai.2017/656>. DOI: 10.24963/IJCAI.2017/656.
- [218] CLARK J, AMODEI D. Faulty reward functions in the wild[J]. Internet: <https://blog.openai.com/faulty-reward-functions>, 2016.
- [219] IBARZ B, LEIKE J, POHLEN T, et al. Reward learning from human preferences and demonstrations in atari[J]. *Advances in neural information processing systems*, 2018, 31.

- [220] VICTORIA K, JONATHAN U, VLADIMIR M, et al. Specification gaming: the flip side of ai ingenuity[J/OL]. DeepMind Blog, 2020. <https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity>.
- [221] NG A Y, HARADA D, RUSSELL S. Policy invariance under reward transformations: Theory and application to reward shaping[C]//Icml: Vol. 99. Citeseer, 1999: 278-287.
- [222] Code Bullet. Simulator with bugs[EB/OL]. 2019. <https://www.youtube.com/watch?v=K-wIZuAA3EY>.
- [223] KRAKOVNA V. More instances about specification gaming[EB/OL]. 2020. <https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml>.
- [224] EVERITT T, HUTTER M, KUMAR R, et al. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective[J]. Synthese, 2021, 198(Suppl 27): 6435-6467.
- [225] RING M, ORSEAU L. Delusion, survival, and intelligent agents[C]//Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings 4. Springer, 2011: 11-20.
- [226] EVERITT T, HUTTER M. The alignment problem for bayesian history-based reinforcement learners[J]. Under submission, 2018.
- [227] KUMAR R, UESATO J, NGO R, et al. Realab: An embedded perspective on tampering[A]. 2020.
- [228] KOCH J, LANGOSCO L, PFAU J, et al. Objective robustness in deep reinforcement learning: Vol. 2[A]. 2021.
- [229] SHAH R, VARMA V. More examples of gmg[EB/OL]. 2022. <https://www.alignmentforum.org/posts/Cfe2LMmQC4hHTDZ8r/more-examples-of-goal-misgeneralization>.
- [230] ZHUANG S, HADFIELD-MENELL D. Consequences of misaligned ai[J]. Advances in Neural Information Processing Systems, 2020, 33: 15763-15773.
- [231] PAULUS R, XIONG C, SOCHER R. A deep reinforced model for abstractive summarization[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=HkAClQgA->.
- [232] KNOX W B, ALLIEVI A, BANZHAF H, et al. Reward (mis) design for autonomous driving[J]. Artificial Intelligence, 2023, 316: 103829.
- [233] GAO L, SCHULMAN J, HILTON J. Scaling laws for reward model overoptimization[C]//International Conference on Machine Learning. PMLR, 2023: 10835-10866.
- [234] PENG A, NUSHI B, KICIMAN E, et al. Investigations of performance and bias in human-ai teamwork in hiring [C]//Proceedings of the AAAI Conference on Artificial Intelligence: 36-11. 2022: 12089-12097.
- [235] COTRA A. Iterated distillation and amplification[M]. Technical report, AI Alignment, 2018.
- [236] IRVING G, CHRISTIANO P, AMODEI D. Ai safety via debate[A]. 2018.
- [237] HUANG W, XIA F, XIAO T, et al. Inner monologue: Embodied reasoning through planning with language models [C/OL]//LIU K, KULIC D, ICHNOWSKI J. Proceedings of Machine Learning Research: Vol. 205 Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand. PMLR, 2022: 1769-1782. <https://proceedings.mlr.press/v205/huang23c.html>.

- [238] ANDREAS J. Language models as agent models[C/OL]//GOLDBERG Y, KOZAREVA Z, ZHANG Y. Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. Association for Computational Linguistics, 2022: 5769-5779. <https://doi.org/10.18653/v1/2022.findings-emnlp.423>.
- [239] KIM G, BALDI P, MCALEER S. Language models can solve computer tasks[A]. 2023.
- [240] YANG S, NACHUM O, DU Y, et al. Foundation models for decision making: Problems, methods, and opportunities [A]. 2023.
- [241] wikipedia. Existential risk from artificial general intelligence[EB/OL]. 2023. https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence.
- [242] NGO R. continuing-the-takeoffs-debate[EB/OL]. 2020. <https://www.alignmentforum.org/posts/Tpn2Fx9daLv28kes/continuing-the-takeoffs-debate>.
- [243] COTRA A. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover - AI Alignment Forum[EB/OL]. 2022[2022-08-11]. <https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to>.
- [244] Jonas DeGrave. Building a virtual machine inside chatgpt[EB/OL]. 2022. <https://www.engraved.blog/building-a-virtual-machine-inside/>.
- [245] Evan Hubinger. Bing chat is blatantly, aggressively misaligned[EB/OL]. 2023. <https://www.lesswrong.com/posts/jtoPawEhLNXNxivgTT/bing-chat-is-blattantly-aggressively-misaligned>.
- [246] LECUN Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27[J]. Open Review, 2022, 62.
- [247] LI K, HOPKINS A K, BAU D, et al. Emergent world representations: Exploring a sequence model trained on a synthetic task[C/OL]//The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. https://openreview.net/pdf?id=DeG07_TcZvT.
- [248] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
- [249] FREEMAN D, HA D, METZ L. Learning to predict without looking ahead: World models without forward prediction[J]. Advances in Neural Information Processing Systems, 2019, 32.
- [250] WIJMANS E, SAVVA M, ESSA I, et al. Emergence of maps in the memories of blind navigation agents[C/OL]//The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. <https://openreview.net/pdf?id=ITt4KjHSsyl>.
- [251] NAKANO R, HILTON J, BALAJI S, et al. Webgpt: Browser-assisted question-answering with human feedback[A]. 2021.
- [252] CARLSMITH J. Is power-seeking ai an existential risk?[A]. 2022.
- [253] EYSENBACH B, GU S, IBARZ J, et al. Leave no trace: Learning to reset for safe and autonomous reinforcement learning[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=S1vuO-bCW>.

- [254] TURNER A, RATZLAFF N, TADEPALLI P. Avoiding side effects in complex environments[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 21406-21415.
- [255] KRAKOVNA V, ORSEAU L, NGO R, et al. Avoiding side effects by considering future tasks[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 19064-19074.
- [256] KLASSEN T Q, MCILRAITH S A, MUISE C, et al. Planning to avoid side effects[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*: 36(9). 2022: 9830-9839.
- [257] ROGER F, GREENBLATT R, NADEAU M, et al. Measurement tampering detection benchmark[A]. 2023.
- [258] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[A]. 2023.
- [259] Ajeya Cotra. why-ai-alignment-could-be-hard-with-modern-deep-learning[EB/OL]. 2021. <https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning>.
- [260] Richard Ngo. Gradient hacking[EB/OL]. 2022. <https://www.alignmentforum.org/posts/EeAgytDZbDjRznPMA/gradient-hacking-definitions-and-examples>.
- [261] WILKE C O, WANG J L, OFRIA C, et al. Evolution of digital organisms at high mutation rates leads to survival of the flattest[J]. *Nature*, 2001, 412(6844): 331-333.
- [262] LEHMAN J, CLUNE J, MISEVIC D, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities[J]. *Artificial life*, 2020, 26(2): 274-306.
- [263] LIN S, HILTON J, EVANS O. Truthfulqa: Measuring how models mimic human falsehoods[C/OL]//MURESAN S, NAKOV P, VILLAVICENCIO A. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, 2022: 3214-3252. <https://doi.org/10.18653/v1/2022.acl-long.229>.
- [264] CHEN M, TWOREK J, JUN H, et al. Evaluating large language models trained on code[A]. 2021.
- [265] KASIRZADEH A, EVANS C. User tampering in reinforcement learning recommender systems[C/OL]//ROSSI F, DAS S, DAVIS J, et al. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*. ACM, 2023: 58-69. <https://doi.org/10.1145/3600211.3604669>.
- [266] AMODEI D, CHRISTIANO P, RAY A. Learning from human preferences[Z]. OpenAI, 2017.
- [267] SPITALE G, BILLER-ANDORNO N, GERMANI F. Ai model gpt-3 (dis)informs us better than humans[J/OL]. *Science Advances*, 2023, 9(26): eadh1850. <https://www.science.org/doi/abs/10.1126/sciadv.adh1850>.
- [268] PHELPS S, RUSSELL Y I. Investigating emergent goal-like behaviour in large language models using experimental economics[A]. 2023. arXiv: 2305.07970.
- [269] PÉROLAT J, LEIBO J Z, ZAMBALDI V F, et al. A multi-agent reinforcement learning model of common-pool resource appropriation[C/OL]//GUYON I, VON LUXBURG U, BENGIO S, et al. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 2017: 3643-3652. <https://proceedings.neurips.cc/paper/2017/hash/2b0f658cbffd284984fb1d90254081f-Abstract.html>.
- [270] SINGH M P. Norms as a basis for governing sociotechnical systems[J/OL]. *ACM Trans. Intell. Syst. Technol.*, 2014, 5(1). <https://doi.org/10.1145/2542182.2542203>.

- [271] BODDINGTON P. Ai and moral thinking: how can we live well with machines to enhance our moral agency? [J/OL]. *AI and Ethics*, 2020, 1: 109-111. DOI: 10.1007/s43681-020-00017-0.
- [272] KORINEK A, BALWIT A. Aligned with whom? direct and social goals for ai systems[R]. National Bureau of Economic Research, 2022.
- [273] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[A]. 2023.
- [274] STUMPF S, RAJARAM V, LI L, et al. Toward harnessing user feedback for machine learning[C]//Proceedings of the 12th international conference on Intelligent user interfaces. 2007: 82-91.
- [275] STUMPF S, RAJARAM V, LI L, et al. Interacting meaningfully with machine learning systems: Three experiments [J/OL]. *International Journal of Human-Computer Studies*, 2009, 67(8): 639-662. <https://www.sciencedirect.com/science/article/pii/S1071581909000457>. DOI: <https://doi.org/10.1016/j.ijhcs.2009.03.004>.
- [276] FERNANDES P, MADAAN A, LIU E, et al. Bridging the gap: A survey on integrating (human) feedback for natural language generation[A]. 2023.
- [277] ZHANG R, TORABI F, GUAN L, et al. Leveraging human guidance for deep reinforcement learning tasks[C/OL]//KRAUS S. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019. ijcai.org, 2019: 6339-6346. <https://doi.org/10.24963/ijcai.2019/884>.
- [278] TAYLOR A T, BERRUETA T A, MURPHEY T D. Active learning in robotics: A review of control principles[J]. *Mechatronics*, 2021, 77: 102576.
- [279] GLAESE A, MCALEESE N, TRĘBACZ M, et al. Improving alignment of dialogue agents via targeted human judgements[A]. 2022.
- [280] META. Meta and microsoft introduce the next generation of llama[EB/OL]. 2023. <https://ai.meta.com/blog/llama-2/>.
- [281] JORDAN M I, MITCHELL T M. Machine learning: Trends, perspectives, and prospects[J]. *Science*, 2015, 349 (6245): 255-260.
- [282] ZHOU Z H. Machine learning[M]. Springer Nature, 2021.
- [283] ÅSTRÖM K J, WITTENMARK B. Adaptive control[M]. Courier Corporation, 2013.
- [284] ÅSTRÖM K J, MURRAY R M. Feedback systems: an introduction for scientists and engineers[M]. Princeton university press, 2021.
- [285] DONG Q, LI L, DAI D, et al. A survey on in-context learning[A]. 2023. arXiv: 2301.00234.
- [286] PARISI S, RAJESWARAN A, PURUSHWALKAM S, et al. The unsurprising effectiveness of pre-trained vision models for control[C]//International Conference on Machine Learning. PMLR, 2022: 17359-17371.
- [287] HU Y, WANG R, LI L E, et al. For Pre-Trained Vision Models in Motor Control, Not All Policy Learning Methods are Created Equal[C]//International Conference on Machine Learning (ICML). 2023: 13628-13651.
- [288] XU J, ZHANG Z, FRIEDMAN T, et al. A semantic loss function for deep learning with symbolic knowledge[C]//International conference on machine learning. PMLR, 2018: 5502-5511.
- [289] ZHOU C, LIU P, XU P, et al. Lima: Less is more for alignment[A]. 2023.

- [290] SILVER D, SINGH S, PRECUP D, et al. Reward is enough[J]. *Artificial Intelligence*, 2021, 299: 103535.
- [291] BROCKMAN G, CHEUNG V, PETERSSON L, et al. Openai gym[A]. 2016.
- [292] KAEHLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: A survey[J]. *Journal of artificial intelligence research*, 1996, 4: 237-285.
- [293] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search[J]. *nature*, 2016, 529(7587): 484-489.
- [294] ISBELL C, SHELTON C R, KEARNS M, et al. A social reinforcement learning agent[C]//*Proceedings of the fifth international conference on Autonomous agents*. 2001: 377-384.
- [295] THOMAZ A L, BREAZEAL C. Teachable robots: Understanding human teaching behavior to build more effective robot learners[J]. *Artificial Intelligence*, 2008, 172(6-7): 716-737.
- [296] HADFIELD-MENELL D, MILLI S, ABBEEL P, et al. Inverse reward design[C/OL]//GUYON I, LUXBURG U V, BENGIO S, et al. *Advances in Neural Information Processing Systems: Vol. 30*. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/32fdab6559cdfa4f167f8c31b9199643-Paper.pdf.
- [297] HUSSEIN A, GABER M M, ELYAN E, et al. Imitation learning: A survey of learning methods[J]. *ACM Computing Surveys (CSUR)*, 2017, 50(2): 1-35.
- [298] SHAW K, BAHL S, PATHAK D. Videodex: Learning dexterity from internet videos[C]//*Conference on Robot Learning*. PMLR, 2023: 654-665.
- [299] EDMONDS M, GAO F, XIE X, et al. Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 3530-3537.
- [300] WANG K, ZHAO Y, SAKUMA I. Learning robotic insertion tasks from human demonstration[J]. *IEEE Robotics and Automation Letters*, 2023.
- [301] BOZORGI H, NGO T D. Beyond shared autonomy: Joint perception and action for human-in-the-loop mobile robot navigation systems[J]. *Journal of Intelligent & Robotic Systems*, 2023, 109(1): 20.
- [302] ZHANG T, MCCARTHY Z, JOW O, et al. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 5628-5635.
- [303] ZHANG W, XU H, NIU H, et al. Discriminator-guided model-based offline imitation learning[C]//*Conference on Robot Learning*. PMLR, 2023: 1266-1276.
- [304] TORABI F, WARNELL G, STONE P. Behavioral Cloning from Observation[C]//*International Joint Conference on Artificial Intelligence (IJCAI)*. 2018: 4950-4957.
- [305] BROWN D S, GOO W, NAGARAJAN P, et al. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations[C]//*International Conference on Machine Learning (ICML)*. 2019: 783-792.
- [306] BAKER B, AKKAYA I, ZHOKOV P, et al. Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos[C]//*Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

-
- [307] FANG B, JIA S, GUO D, et al. Survey of imitation learning for robotic manipulation[J]. *International Journal of Intelligent Robotics and Applications*, 2019, 3: 362-369.
- [308] DASARI S, GUPTA A, KUMAR V. Learning Dexterous Manipulation from Exemplar Object Trajectories and Pre-Grasps[C]//2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023.
- [309] SASAKI F, YAMASHINA R. Behavioral cloning from noisy demonstrations[C]//International Conference on Learning Representations. 2020.
- [310] WANG Q, MCCARTHY R, BULENS D C, et al. Improving Behavioural Cloning with Positive Unlabeled Learning [J]. *CoRL 2023 Poster*, 2023.
- [311] ATTIA A, DAYAN S. Global overview of imitation learning: abs/1801.06503[A]. 2018.
- [312] YANG M, LEVINE S, NACHUM O. TRAIL: near-optimal imitation learning with suboptimal data[C/OL]//The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022. https://openreview.net/forum?id=6q_2b6u0BnJ.
- [313] ZHU H, YU J, GUPTA A, et al. The ingredients of real world robotic reinforcement learning[C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. <https://openreview.net/forum?id=rJe2syrtvS>.
- [314] HEJNA III D J, SADIGH D. Few-Shot Preference Learning for Human-in-the-Loop RL[C]//Conference on Robot Learning (CoRL). 2022: 2014-2025.
- [315] BELIAEV M, SHIH A, ERMON S, et al. Imitation learning by estimating expertise of demonstrators[C]//International Conference on Machine Learning. PMLR, 2022: 1732-1748.
- [316] WIRTH C, AKROUR R, NEUMANN G, et al. A survey of preference-based reinforcement learning methods[J]. *Journal of Machine Learning Research*, 2017, 18(136): 1-46.
- [317] FÜRNKRANZ J, HÜLLERMEIER E. Preference learning[M]. Springer Science & Business Media, 2010.
- [318] HÜLLERMEIER E, FÜRNKRANZ J, CHENG W, et al. Label ranking by learning pairwise preferences[J]. *Artificial Intelligence*, 2008, 172(16-17): 1897-1916.
- [319] FÜRNKRANZ J, HÜLLERMEIER E. Pairwise preference learning and ranking[C]//European conference on machine learning. Springer, 2003: 145-156.
- [320] JEON H J, MILLI S, DRAGAN A D. Reward-rational (implicit) choice: A unifying formalism for reward learning [C]//Conference on Neural Information Processing Systems (NeurIPS). 2020.
- [321] OPENAI. Gpt-4v(ision) system card[EB/OL]. 2023. https://cdn.openai.com/papers/GPTV_System_Card.pdf.
- [322] Anthropic. Model card and evaluations for claude models[EB/OL]. 2023. <https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf>.
- [323] SHIN D, DRAGAN A D, BROWN D S. Benchmarks and algorithms for offline preference-based reward learning [J/OL]. *Trans. Mach. Learn. Res.*, 2023, 2023. <https://openreview.net/forum?id=TGuxXlKsn>.
- [324] KIM C, PARK J, SHIN J, et al. Preference Transformer: Modeling Human Preferences using Transformers for RL [C]//International Conference on Learning Representations (ICLR). 2023.

-
- [325] BUKHARIN A W, LI Y, HE P, et al. Deep Reinforcement Learning from Hierarchical Weak Preference Feedback: abs/2309.02632[A]. 2023.
- [326] BENDER E M, GEBRU T, MCMILLAN-MAJOR A, et al. On the dangers of stochastic parrots: Can language models be too big?[C]//Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021: 610-623.
- [327] AKROUR R, SCHOENAUER M, SEBAG M. Preference-based policy learning[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11. Springer, 2011: 12-27.
- [328] WIRTH C, FÜRNKRANZ J. Preference-based reinforcement learning: A preliminary survey[C]//Proceedings of the ECML/PKDD-13 Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Rewards. Citeseer, 2013.
- [329] CABI S, COLMENAREJO S G, NOVIKOV A, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning[C/OL]//TOUSSAINT M, BICCHI A, HERMANS T. Robotics: Science and Systems XVI, Virtual Event / Corvallis, Oregon, USA, July 12-16, 2020. 2020. <https://doi.org/10.15607/RSS.2020.XVI.076>.
- [330] LIANG X, SHU K, LEE K, et al. Reward Uncertainty for Exploration in Preference-based Reinforcement Learning [C]//International Conference on Learning Representations (ICLR). 2022.
- [331] XUE W, AN B, YAN S, et al. Reinforcement learning from diverse human preferences[A]. 2023.
- [332] CHRISTIANO P, SHLEGERIS B, AMODEI D. Supervising strong learners by amplifying weak experts[A]. 2018.
- [333] TSOUMAKAS G, KATAKIS I. Multi-label classification: An overview[J]. International Journal of Data Warehousing and Mining (IJDWM), 2007, 3(3): 1-13.
- [334] CHENG W, HÜLLERMEIER E, DEMBCZYNSKI K J. Graded multilabel classification: The ordinal case[C]//Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 223-230.
- [335] CHENG W, HÜLLERMEIER E, DEMBCZYNSKI K J. Label ranking methods based on the plackett-luce model [C]//Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: 215-222.
- [336] CHENG W, RADEMAKER M, DE BAETS B, et al. Predicting partial orders: ranking with abstention[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21. Springer, 2010: 215-230.
- [337] BRADLEY R A, TERRY M E. Rank analysis of incomplete block designs: I. the method of paired comparisons[J]. Biometrika, 1952, 39(3/4): 324-345.
- [338] PLACKETT R L. The analysis of permutations[J]. Journal of the Royal Statistical Society Series C: Applied Statistics, 1975, 24(2): 193-202.
- [339] KNOX W B. Learning from human-generated reward[Z]. 2012.
- [340] KNOX W B, STONE P. Learning non-myopically from human-generated reward[C]//Proceedings of the 2013 international conference on Intelligent user interfaces. 2013: 191-202.

-
- [341] PALAN M, SHEVCHUK G, LANDOLFI N C, et al. Learning Reward Functions by Integrating Human Demonstrations and Preferences[C]//International Conference on Reliability, Safety, and Security of Railway Systems (RSSRail). 2019.
- [342] STIENNON N, OUYANG L, WU J, et al. Learning to summarize with human feedback[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 3008-3021.
- [343] FAWZI A, BALOG M, HUANG A, et al. Discovering faster matrix multiplication algorithms with reinforcement learning[J]. *Nature*, 2022, 610(7930): 47-53.
- [344] MANKOWITZ D J, MICHI A, ZHERNOV A, et al. Faster sorting algorithms discovered using deep reinforcement learning[J]. *Nature*, 2023, 618(7964): 257-263.
- [345] AGOSTINELLI F, HOCQUET G, SINGH S, et al. From reinforcement learning to deep reinforcement learning: An overview[C/OL]//ROZONOER L I, MIRKIN B G, MUCHNIK I. *Lecture Notes in Computer Science: Vol. 11100 Braverman Readings in Machine Learning. Key Ideas from Inception to Current State - International Conference Commemorating the 40th Anniversary of Emmanuil Braverman's Decease, Boston, MA, USA, April 28-30, 2017, Invited Talks*. Springer, 2017: 298-328. https://doi.org/10.1007/978-3-319-99492-5_13.
- [346] YU C, LIU J, NEMATI S, et al. Reinforcement learning in healthcare: A survey[J]. *ACM Computing Surveys (CSUR)*, 2021, 55(1): 1-36.
- [347] AFSAR M M, CRUMP T, FAR B. Reinforcement learning based recommender systems: A survey[J]. *ACM Computing Surveys*, 2022, 55(7): 1-38.
- [348] MAKОВИYCHUK V, WAWRZYNIAK L, GUO Y, et al. Isaac gym: High performance GPU based physics simulation for robot learning[C/OL]//VANSCHOREN J, YEUNG S. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*. 2021. <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/28dd2c7955ce926456240b2ff0100bde-Abstract-round2.html>.
- [349] BUŞONIU L, DE BRUIN T, TOLIĆ D, et al. Reinforcement learning for control: Performance, stability, and deep approximators[J]. *Annual Reviews in Control*, 2018, 46: 8-28.
- [350] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[A]. 2017.
- [351] SADIGH D, DRAGAN A D, SASTRY S, et al. Active preference-based learning of reward functions[M]. UC Berkeley, 2017.
- [352] REDDY S, DRAGAN A, LEVINE S, et al. Learning human objectives by evaluating hypothetical behavior[C]//International Conference on Machine Learning. PMLR, 2020: 8020-8029.
- [353] KUPCSIK A, DEISENROTH M, PETERS J, et al. Data-efficient generalization of robot skills with contextual policy search[C]//Proceedings of the AAAI conference on artificial intelligence: 27(1). 2013: 1401-1407.
- [354] JAIN A, WOJCIK B, JOACHIMS T, et al. Learning trajectory preferences for manipulators via iterative improvement[J]. *Advances in neural information processing systems*, 2013, 26.
- [355] DUCHI J C, MACKEY L W, JORDAN M I. On the consistency of ranking algorithms.[C]//ICML. 2010: 327-334.

-
- [356] AKROUR R, SCHOENAUER M, SEBAG M. April: Active preference learning-based reinforcement learning[C]// Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23. Springer, 2012: 116-131.
- [357] WHITE D, WU M, NOVOSELLER E, et al. Rating-based reinforcement learning[A]. 2023.
- [358] MCKINNEY L, DUAN Y, KRUEGER D, et al. On the fragility of learned reward functions[A]. 2023.
- [359] SCHAAL S. Is imitation learning the route to humanoid robots?[J]. Trends in cognitive sciences, 1999, 3(6): 233-242.
- [360] SYED U, BOWLING M, SCHAPIRE R E. Apprenticeship learning using linear programming[C]//Proceedings of the 25th international conference on Machine learning. 2008: 1032-1039.
- [361] HE H, EISNER J, DAUME H. Imitation learning by coaching[J]. Advances in neural information processing systems, 2012, 25.
- [362] ZARE M, KEBRIA P, KHOSRAVI A, et al. A Survey of Imitation Learning: Algorithms, Recent Developments, and Challenges: abs/2309.02473[A]. 2023.
- [363] BAKKER P, KUNIYOSHI Y, et al. Robot see, robot do: An overview of robot imitation[C]//AISB96 Workshop on Learning in Robots and Animals: Vol. 5. 1996.
- [364] BAIN M, SAMMUT C. A framework for behavioural cloning.[C]//Machine Intelligence 15. 1995: 103-129.
- [365] ROSS S, GORDON G, BAGNELL D. A reduction of imitation learning and structured prediction to no-regret online learning[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 627-635.
- [366] OSA T, PAJARINEN J, NEUMANN G, et al. An algorithmic perspective on imitation learning[J]. Foundations and Trends® in Robotics, 2018, 7(1-2): 1-179.
- [367] POMERLEAU D A. Efficient training of artificial neural networks for autonomous navigation[J]. Neural computation, 1991, 3(1): 88-97.
- [368] RAVICHANDAR H, POLYDOROS A S, CHERNOVA S, et al. Recent advances in robot learning from demonstration[J]. Annual review of control, robotics, and autonomous systems, 2020, 3: 297-330.
- [369] SCHAAL S. Learning from demonstration[J]. Advances in neural information processing systems, 1996, 9.
- [370] LYNCH C, KHANSARI M, XIAO T, et al. Learning latent plans from play[C/OL]//KAELBLING L P, KRAGIC D, SUGIURA K. Proceedings of Machine Learning Research: Vol. 100 Proceedings of the Conference on Robot Learning. PMLR, 2020: 1113-1132. <https://proceedings.mlr.press/v100/lynch20a.html>.
- [371] HO J, ERMON S. Generative adversarial imitation learning[J]. Advances in neural information processing systems, 2016, 29.
- [372] REDDY S, DRAGAN A D, LEVINE S. SQIL: imitation learning via reinforcement learning with sparse rewards [C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. <https://openreview.net/forum?id=S1xKd24twB>.
- [373] ZHOU K, LIU Z, QIAO Y, et al. Domain Generalization: A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, PP: 1-20.

- [374] ADAMS S, CODY T, BELING P A. A survey of inverse reinforcement learning[J]. *Artificial Intelligence Review*, 2022, 55(6): 4307-4346.
- [375] ARORA S, DOSHI P. A survey of inverse reinforcement learning: Challenges, methods and progress[J]. *Artificial Intelligence*, 2021, 297: 103500.
- [376] ABBEEL P, NG A Y. Apprenticeship learning via inverse reinforcement learning[C/OL]//ICML '04: Proceedings of the Twenty-First International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2004: 1. <https://doi.org/10.1145/1015330.1015430>.
- [377] ZIEBART B D, MAAS A L, BAGNELL J A, et al. Maximum entropy inverse reinforcement learning.[C]//Aaai: Vol. 8. Chicago, IL, USA, 2008: 1433-1438.
- [378] ALSALEH R, SAYED T. Modeling pedestrian-cyclist interactions in shared space using inverse reinforcement learning[J]. *Transportation research part F: traffic psychology and behaviour*, 2020, 70: 37-57.
- [379] RAMACHANDRAN D, AMIR E. Bayesian inverse reinforcement learning.[C]//IJCAI: Vol. 7. 2007: 2586-2591.
- [380] FU J, SINGH A, GHOSH D, et al. Variational inverse control with events: A general framework for data-driven reward definition[J]. *Advances in neural information processing systems*, 2018, 31.
- [381] YU Y. Towards sample efficient reinforcement learning.[C]//IJCAI. 2018: 5739-5743.
- [382] GARCIA J, FERNÁNDEZ F. A comprehensive survey on safe reinforcement learning[J]. *Journal of Machine Learning Research*, 2015, 16(1): 1437-1480.
- [383] XU M, LIU Z, HUANG P, et al. Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability[A]. 2022. arXiv: 2209.08025.
- [384] KIM K, GARG S, SHIRAGUR K, et al. Reward identification in inverse reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2021: 5496-5505.
- [385] KNOX W B, STONE P. Tamer: Training an agent manually via evaluative reinforcement[C]//2008 7th IEEE international conference on development and learning. IEEE, 2008: 292-297.
- [386] KNOX W B, STONE P. Reinforcement learning from simultaneous human and mdp reward.[C]//AAMAS: Vol. 1004. Valencia, 2012: 475-482.
- [387] KNOX W B, STONE P, BREAZEAL C. Training a robot via human feedback: A case study[C]//Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5. Springer, 2013: 460-470.
- [388] GRIFFITH S, SUBRAMANIAN K, SCHOLZ J, et al. Policy shaping: Integrating human feedback with reinforcement learning[J]. *Advances in neural information processing systems*, 2013, 26.
- [389] LOFTIN R, PENG B, MACGLASHAN J, et al. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning[J]. *Autonomous agents and multi-agent systems*, 2016, 30: 30-59.
- [390] KORBAC T, SHI K, CHEN A, et al. Pretraining language models with human preferences[C]//International Conference on Machine Learning. PMLR, 2023: 17506-17533.
- [391] CHRISTIANO P. Thoughts on the impact of rlhf research[EB/OL]. 2023. <https://www.lesswrong.com/posts/vw u4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research>.

- [392] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C/OL]//BURSTEIN J, DORAN C, SOLORIO T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019: 4171-4186. <https://doi.org/10.18653/v1/n19-1423>.
- [393] ZIEGLER D M, STIENNON N, WU J, et al. Fine-tuning language models from human preferences[A]. 2019.
- [394] DAI J, PAN X, SUN R, et al. Safe rlhf: Safe reinforcement learning from human feedback[A]. 2023.
- [395] SUN Z, SHEN Y, ZHOU Q, et al. Principle-driven self-alignment of language models from scratch with minimal human supervision[A]. 2023.
- [396] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[A]. 2023.
- [397] TAORI R, GULRAJANI I, ZHANG T, et al. Stanford alpaca: An instruction-following llama model[Z]. 2023.
- [398] CHIANG W L, LI Z, LIN Z, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality [J]. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023.
- [399] CHOSHEN L, FOX L, AIZENBUD Z, et al. On the weaknesses of reinforcement learning for neural machine translation[C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. <https://openreview.net/forum?id=H1eCw3EKvH>.
- [400] YUAN Z, YUAN H, TAN C, et al. Rrlhf: Rank responses to align language models with human feedback without tears[A]. 2023.
- [401] ZHANG T, LIU F, WONG J, et al. The wisdom of hindsight makes language models better instruction followers [C/OL]//KRAUSE A, BRUNSKILL E, CHO K, et al. Proceedings of Machine Learning Research: Vol. 202 International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. PMLR, 2023: 41414-41428. <https://proceedings.mlr.press/v202/zhang23ab.html>.
- [402] GULCEHRE C, PAINE T L, SRINIVASAN S, et al. Reinforced self-training (rest) for language modeling[A]. 2023.
- [403] AZAR M G, ROWLAND M, PIOT B, et al. A general theoretical paradigm to understand learning from human preferences[A]. 2023.
- [404] YEYGEN CHEBOTAR T Y. Rt-2: New model translates vision and language into action[EB/OL]. 2023. <https://www.deepmind.com/blog/rt-2-new-model-translates-vision-and-language-into-action>.
- [405] CARLSON T, DEMIRIS Y. Increasing robotic wheelchair safety with collaborative control: Evidence from secondary task experiments[C]//2010 IEEE International Conference on Robotics and Automation. IEEE, 2010: 5582-5587.
- [406] WU J, OUYANG L, ZIEGLER D M, et al. Recursively summarizing books with human feedback[A]. 2021. arXiv: 2109.10862.
- [407] BI Z M, LUO C, MIAO Z, et al. Safety assurance mechanisms of collaborative robotic systems in manufacturing [J]. Robotics and Computer-Integrated Manufacturing, 2021, 67: 102022.

- [408] RUSSELL S, DEWEY D, TEGMARK M. Research priorities for robust and beneficial artificial intelligence[J]. *AI magazine*, 2015, 36(4): 105-114.
- [409] BROWN D, COLEMAN R, SRINIVASAN R, et al. Safe imitation learning via fast Bayesian reward inference from preferences[C//OL]//III H D, SINGH A. *Proceedings of Machine Learning Research: Vol. 119 Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020: 1165-1177. <https://proceedings.mlr.press/v119/brown20a.html>.
- [410] LI B, QI P, LIU B, et al. Trustworthy ai: From principles to practices[J]. *ACM Computing Surveys*, 2023, 55(9): 1-46.
- [411] SAUNDERS W, YEH C, WU J, et al. Self-critiquing models for assisting human evaluators[A]. 2022.
- [412] PEARCE H, AHMAD B, TAN B, et al. Asleep at the keyboard? assessing the security of github copilot' s code contributions[C]//2022 IEEE Symposium on Security and Privacy (SP). IEEE, 2022: 754-768.
- [413] BI K, XIE L, ZHANG H, et al. Accurate medium-range global weather forecasting with 3d neural networks[J]. *Nature*, 2023: 1-6.
- [414] BAI Y, KADAVATH S, KUNDU S, et al. Constitutional ai: Harmlessness from ai feedback[A]. 2022.
- [415] LEE H, PHATALE S, MANSOOR H, et al. Rlaif: Scaling reinforcement learning from human feedback with ai feedback[A]. 2023.
- [416] YUDKOWSKY E. Challenges to christiano' s capability amplification proposal[J]. *LessWrong*, 2018.
- [417] REED S E, ZOLNA K, PARISOTTO E, et al. A generalist agent[J/OL]. *Trans. Mach. Learn. Res.*, 2022, 2022. <https://openreview.net/forum?id=1ikK0kHvj>.
- [418] DU Y, LI S, TORRALBA A, et al. Improving factuality and reasoning in language models through multiagent debate[A]. 2023.
- [419] HADFIELD-MENELL D, RUSSELL S J, ABBEEL P, et al. Cooperative inverse reinforcement learning[J]. *Advances in neural information processing systems*, 2016, 29.
- [420] FICKINGER A, ZHUANG S, HADFIELD-MENELL D, et al. Multi-principal assistance games[A]. 2020.
- [421] SHAH R, FREIRE P, ALEX N, et al. Benefits of assistance over reward learning[J]. <https://openreview.net/forum?id=DFIoGDZeJIB>, 2020.
- [422] CARR A. Teaching large language models to zip their lips[EB/OL]. 2023. <https://gretel.ai/blog/teaching-large-language-models-to-zip-their-lips>.
- [423] NGUYEN C. My understanding of paul christiano's iterated amplification al safety research agenda[EB/OL]. 2020. https://www.alignmentforum.org/posts/PT8vSxsusqWuN7JXp/my-understanding-of-paul-christiano-s-iterated-amplification#A_mathematical_way_of_solving_Go_is_impossible.
- [424] MENNEN A. A comment on the ida-alphagozero metaphor; capabilities versus alignment[EB/OL]. 2018. <https://www.alignmentforum.org/posts/yXFKh2jGysQNfX2NM/a-comment-on-the-ida-alphagozero-metaphor-capabilities>.
- [425] MUKOBI G. Iterated distillation-amplification, gato, and proto-agi[EB/OL]. 2022. <https://www.lesswrong.com/posts/Evyk8eb6b7tFd6pxJ/iterated-distillation-amplification-gato-and-proto-agi-re>.

- [426] HUBINGER E. An overview of 11 proposals for building safe advanced ai[A]. 2020.
- [427] DALAL G, DVIJOTHAM K, VECERIK M, et al. Safe exploration in continuous action spaces[A]. 2018.
- [428] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=rJzIBfZAb>.
- [429] MACGLASHAN J, HO M K, LOFTIN R, et al. Interactive learning from policy-dependent human feedback[C]// International conference on machine learning. PMLR, 2017: 2285-2294.
- [430] CLAUDE R R G . New lw feature debates[EB/OL]. 2023. <https://www.lesswrong.com/posts/kXiAGRWFquXFMi68Y/new-lw-feature-debates>.
- [431] NGO R. /why i m excited about debate[EB/OL]. 2021. <https://www.alignmentforum.org/posts/LDsSqXf9Dpu3J3gHD/why-i-m-excited-about-debate>.
- [432] MICHAELCOHEN. the-ai-debate-debate[EB/OL]. 2020. <https://www.alignmentforum.org/posts/L3QDs6of4Rb2TgpRD/the-ai-debate-debate>.
- [433] ARMSTRONG S. problems with ai debate[EB/OL]. 2019. <https://www.alignmentforum.org/posts/fNTCveSa4HvqvZR2F/problems-with-ai-debate>.
- [434] BARNES B. debate-update-obfuscated-arguments-problem[EB/OL]. 2020. <https://www.alignmentforum.org/posts/PJLABqQ962hZEqhdB/debate-update-obfuscated-arguments-problem>.
- [435] BETH BARNES P C. writeup-progress-on-ai-safety-via-debate-1[EB/OL]. 2020. <https://www.alignmentforum.org/posts/Br4xDyYu4Frwr64a/writeup-progress-on-ai-safety-via-debate-1>.
- [436] NAYYAR A, MAHAJAN A, TENEKETZIS D. Decentralized stochastic control with partial history sharing: A common information approach[J]. IEEE Transactions on Automatic Control, 2013, 58(7): 1644-1658.
- [437] FISAC J F, GATES M A, HAMRICK J B, et al. Pragmatic-pedagogic value alignment[C]//Robotics Research: The 18th International Symposium ISRR. Springer, 2020: 49-57.
- [438] HONG J, BHATIA K, DRAGAN A D. On the sensitivity of reward inference to misspecified human models[C/OL]// The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. <https://openreview.net/pdf?id=hJqGbUpDGV>.
- [439] BOBU A, PENG A, AGRAWAL P, et al. Aligning robot and human representations[A]. 2023.
- [440] HE J Z, DRAGAN A D. Assisted robust reward design[C/OL]//FAUST A, HSU D, NEUMANN G. Proceedings of Machine Learning Research: Vol. 164 Conference on Robot Learning, 8-11 November 2021, London, UK. PMLR, 2021: 1234-1246. <https://proceedings.mlr.press/v164/he22a.html>.
- [441] POURSAEED O, JIANG T, YANG H, et al. Robustness and generalization via generative adversarial training [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 15711-15720.
- [442] GILMER J, FORD N, CARLINI N, et al. Adversarial examples are a natural consequence of test error in noise[C/OL]//CHAUDHURI K, SALAKHUTDINOV R. Proceedings of Machine Learning Research: Vol. 97 Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. PMLR, 2019: 2280-2289. <http://proceedings.mlr.press/v97/gilmer19a.html>.

- [443] MA W, DUAN P, LIU S, et al. Shadow attacks: automatically evading system-call-behavior based malware detection [J]. *Journal in Computer Virology*, 2012, 8: 1-13.
- [444] YOO J Y, QI Y. Towards improving adversarial training of NLP models[C/OL]//Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 945-956. <https://aclanthology.org/2021.findings-emnlp.81>. DOI: 10.18653/v1/2021.findings-emnlp.81.
- [445] ZIEGLER D, NIX S, CHAN L, et al. Adversarial training for high-stakes reliability[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 9274-9286.
- [446] DEEPMIND. goal misgeneralization[EB/OL]. 2020. https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vTo3RkXUAigb25nP7gipcHriR6XdzA_L5loOcVFj_u7cRAZghWrYKH2L2nU4TA_Vr9KzBX5Bjz9G_1/pubhtml?pli=1.
- [447] GEIRHOS R, RUBISCH P, MICHAELIS C, et al. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness[A]. 2018.
- [448] MURPHY K P. Probabilistic machine learning: Advanced topics[M/OL]. MIT Press, 2023. <http://probml.github.io/book2>.
- [449] DE HAAN P, JAYARAMAN D, LEVINE S. Causal confusion in imitation learning[J]. *Advances in Neural Information Processing Systems*, 2019, 32.
- [450] DAI D, SUN Y, DONG L, et al. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers[C]//ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models. 2023.
- [451] VON OSWALD J, NIKLASSON E, SCHLEGEL M, et al. Uncovering mesa-optimization algorithms in transformers [A]. 2023. arXiv: 2309.05858.
- [452] BESBES O, MA W, MOUCHTAKI O. Beyond IID: data-driven decision-making in heterogeneous environments [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 23979-23991.
- [453] CARROLL M D, DRAGAN A, RUSSELL S, et al. Estimating and penalizing induced preference shifts in recommender systems[C]//International Conference on Machine Learning. PMLR, 2022: 2686-2708.
- [454] BEN-TAL A, EL GHAOU L, NEMIROVSKI A. Robust optimization: Vol. 28[M]. Princeton university press, 2009.
- [455] PETERS J, BUHLMANN P, MEINSHAUSEN N. Causal inference using invariant prediction: identification and confidence intervals. arxiv[J]. *Methodology*, 2015.
- [456] BEERY S, VAN HORN G, PERONA P. Recognition in terra incognita[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 456-473.
- [457] HENDRYCKS D, DIETTERICH T G. Benchmarking neural network robustness to common corruptions and perturbations[C/OL]//7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. <https://openreview.net/forum?id=HJz6tiCqYm>.
- [458] ENGSTROM L, TRAN B, TSIPRAS D, et al. Exploring the landscape of spatial robustness[C]//International conference on machine learning. PMLR, 2019: 1802-1811.

- [459] RECHT B, ROELOFS R, SCHMIDT L, et al. Do imagenet classifiers generalize to imagenet?[C]//International conference on machine learning. PMLR, 2019: 5389-5400.
- [460] SAGAWA S, KOH P W, HASHIMOTO T B, et al. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization[A]. 2019.
- [461] DUCHI J C, GLYNN P W, NAMKOONG H. Statistics of robust optimization: A generalized empirical likelihood approach[J]. Mathematics of Operations Research, 2021, 46(3): 946-969.
- [462] VAPNIK V. Principles of risk minimization for learning theory[J]. Advances in neural information processing systems, 1991, 4.
- [463] ARJOVSKY M, BOTTOU L, GULRAJANI I, et al. Invariant risk minimization[A]. 2019.
- [464] PETERS J, JANZING D, SCHÖLKOPF B. Elements of causal inference: foundations and learning algorithms[M]. The MIT Press, 2017.
- [465] GARIPPOV T, IZMAILOV P, PODOPRIKHIN D, et al. Loss surfaces, mode connectivity, and fast ensembling of dnns[J]. Advances in neural information processing systems, 2018, 31.
- [466] DRAXLER F, VESCHGINI K, SALMHOFER M, et al. Essentially no barriers in neural network energy landscape [C]//International conference on machine learning. PMLR, 2018: 1309-1318.
- [467] FRANKLE J, DZIUGAITE G K, ROY D, et al. Linear mode connectivity and the lottery ticket hypothesis[C]//International Conference on Machine Learning. PMLR, 2020: 3259-3269.
- [468] BENTON G, MADDOX W, LOTFI S, et al. Loss surface simplexes for mode connecting volumes and fast ensembling [C]//International Conference on Machine Learning. PMLR, 2021: 769-779.
- [469] PITTORINO F, FERRARO A, PERUGINI G, et al. Deep networks on toroids: removing symmetries reveals the structure of flat regions in the landscape geometry[C]//International Conference on Machine Learning. PMLR, 2022: 17759-17781.
- [470] JUNEJA J, BANSAL R, CHO K, et al. Linear connectivity reveals generalization strategies[C/OL]//The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. <https://openreview.net/pdf?id=hY6M0JHl3uL>.
- [471] ZHANG H, CISSÉ M, DAUPHIN Y N, et al. mixup: Beyond empirical risk minimization[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=r1Ddp1-Rb>.
- [472] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[A]. 2013.
- [473] CORTES C, MOHRI M, ROSTAMIZADEH A. L2 regularization for learning kernels[C/OL]//BILMES J A, NG A Y. UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009. AUAI Press, 2009: 109-116. https://www.auai.org/uai2009/papers/UAI2009_0132_16d9521907263bf91a07477a87253260.pdf.
- [474] PRECHELT L. Early stopping-but when?[M]//Neural Networks: Tricks of the trade. Springer, 2002: 55-69.
- [475] RAHIMIAN H, MEHROTRA S. Distributionally robust optimization: A review[A]. 2019.

- [476] CHEN R, PASCHALIDIS I C, et al. Distributionally robust learning[J]. *Foundations and Trends® in Optimization*, 2020, 4(1-2): 1-243.
- [477] LIN F, FANG X, GAO Z. Distributionally robust optimization: A review on theory and applications[J]. *Numerical Algebra, Control and Optimization*, 2022, 12(1): 159-212.
- [478] ZHENG S, SONG Y, LEUNG T, et al. Improving the robustness of deep neural networks via stability training[C]// *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: 4480-4488.
- [479] HUANG S H, PAPERNOT N, GOODFELLOW I J, et al. Adversarial attacks on neural network policies[C/OL]// *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. <https://openreview.net/forum?id=ryv1RyBkl>.
- [480] BHATTAD A, CHONG M J, LIANG K, et al. Unrestricted adversarial examples via semantic manipulation[C/OL]// *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. https://openreview.net/forum?id=Sy_eOgHFwH.
- [481] SHAMSABADI A S, SANCHEZ-MATILLA R, CAVALLARO A. Colorfool: Semantic adversarial colorization[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020: 1151-1160.
- [482] CASPER S, NADEAU M, HADFIELD-MENELL D, et al. Robust feature-level adversaries are interpretability tools[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 33093-33106.
- [483] JIA R, LIANG P. Adversarial examples for evaluating reading comprehension systems[C/OL]// *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 2021-2031. <https://aclanthology.org/D17-1215>. DOI: 10.18653/v1/D17-1215.
- [484] REN Y, LIN J, TANG S, et al. Generating natural language adversarial examples on a large scale with generative models[C/OL]// GIACOMO G D, CATALÁ A, DILKINA B, et al. *Frontiers in Artificial Intelligence and Applications: Vol. 325 ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*. IOS Press, 2020: 2156-2163. <https://doi.org/10.3233/FAIA200340>.
- [485] CHEN Z, LI B, WU S, et al. Content-based unrestricted adversarial attack[A]. 2023.
- [486] MOSKOVITZ T, SINGH A K, STROUSE D, et al. Confronting reward model overoptimization with constrained rlhf[A]. 2023. arXiv: 2310.04373.
- [487] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[A]. 2014.
- [488] WANG D, GONG C, LIU Q. Improving neural language modeling via adversarial training[C]// *International Conference on Machine Learning*. PMLR, 2019: 6555-6565.
- [489] LIU X, CHENG H, HE P, et al. Adversarial training for large neural language models[A]. 2020.
- [490] GAN Z, CHEN Y C, LI L, et al. Large-scale adversarial training for vision-and-language representation learning [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 6616-6628.

-
- [491] BERG H, HALL S M, BHARGAVA Y, et al. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning[C/OL]//HE Y, JI H, LIU Y, et al. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022. Association for Computational Linguistics, 2022: 806-822. <https://aclanthology.org/2022.acl-main.61>.
- [492] GLEAVE A, DENNIS M, WILD C, et al. Adversarial policies: Attacking deep reinforcement learning[C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. <https://openreview.net/forum?id=HJgEMpVFwB>.
- [493] PINTO L, DAVIDSON J, SUKTHANKAR R, et al. Robust adversarial reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2017: 2817-2826.
- [494] VINITSKY E, DU Y, PARVATE K, et al. Robust reinforcement learning using adversarial populations[A]. 2020.
- [495] TAN K L, ESFANDIARI Y, LEE X Y, et al. Robustifying reinforcement learning agents via action space adversarial training[C]//2020 American control conference (ACC). IEEE, 2020: 3959-3964.
- [496] MCALEER S, WANG K, LANIER J, et al. Anytime psro for two-player zero-sum games[A]. 2022.
- [497] LIANG Y, SUN Y, ZHENG R, et al. Game-theoretic robust reinforcement learning handles temporally-coupled perturbations[A]. 2023.
- [498] CARMON Y, RAGHUNATHAN A, SCHMIDT L, et al. Unlabeled data improves adversarial robustness[J]. Advances in neural information processing systems, 2019, 32.
- [499] HENDRYCKS D, MAZEIKA M, KADAVATH S, et al. Using self-supervised learning can improve model robustness and uncertainty[J]. Advances in neural information processing systems, 2019, 32.
- [500] ZHANG J, XU X, HAN B, et al. Attacks which do not kill training make adversarial learning stronger[C]//International conference on machine learning. PMLR, 2020: 11278-11287.
- [501] MAO X, CHEN Y, DUAN R, et al. Enhance the visual representation via discrete adversarial training[J]. Advances in Neural Information Processing Systems, 2022, 35: 7520-7533.
- [502] ZHANG L, YANG N, SUN Y, et al. Provable unrestricted adversarial training without compromise with generalizability[A]. 2023.
- [503] TAN M. Multi-agent reinforcement learning: Independent vs. cooperative agents[C]//Proceedings of the tenth international conference on machine learning. 1993: 330-337.
- [504] GUESTRIN C, KOLLER D, PARR R. Multiagent planning with factored mdps[J]. Advances in neural information processing systems, 2001, 14.
- [505] FOERSTER J, ASSAEL I A, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning[J]. Advances in neural information processing systems, 2016, 29.
- [506] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward[C/OL]//ANDRÉ E, KOENIG S, DASTANI M, et al. Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018: 2085-2087. <http://dl.acm.org/citation.cfm?id=3238080>.

-
- [507] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. *Advances in neural information processing systems*, 2017, 30.
- [508] SONG J, REN H, SADIGH D, et al. Multi-agent generative adversarial imitation learning[J]. *Advances in neural information processing systems*, 2018, 31.
- [509] SINGH A, JAIN T, SUKHBAATAR S. Learning when to communicate at scale in multiagent cooperative and competitive tasks[C//OL]//7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019. <https://openreview.net/forum?id=rye7knCqK7>.
- [510] JADERBERG M, CZARNECKI W M, DUNNING I, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning[J]. *Science*, 2019, 364(6443): 859-865.
- [511] CRUZ D, CRUZ J A, LOPES CARDOSO H. Reinforcement learning in multi-agent games: Open ai gym diplomacy environment[C//Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3-6, 2019, Proceedings, Part I 19. Springer, 2019: 49-60.
- [512] GRONAUER S, DIEPOLD K. Multi-agent deep reinforcement learning: a survey[J]. *Artificial Intelligence Review*, 2022: 1-49.
- [513] FAIR D T, BAKHTIN A, BROWN N, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning[J]. *Science*, 2022, 378(6624): 1067-1074.
- [514] OROOJLOOY A, HAJINEZHAD D. A review of cooperative multi-agent deep reinforcement learning[J]. *Applied Intelligence*, 2023, 53(11): 13677-13722.
- [515] TOBIN J, FONG R, RAY A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[C//2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017: 23-30.
- [516] HU H, LERER A, PEYSAKHOVICH A, et al. “other-play” for zero-shot coordination[C//International Conference on Machine Learning. PMLR, 2020: 4399-4410.
- [517] TREUTLEIN J, DENNIS M, OESTERHELD C, et al. A new formalism, method and open issues for zero-shot coordination[C//International Conference on Machine Learning. PMLR, 2021: 10413-10423.
- [518] CUI B, HU H, PINEDA L, et al. K-level reasoning for zero-shot coordination in hanabi[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 8215-8228.
- [519] HU H, LERER A, CUI B, et al. Off-belief learning[C//International Conference on Machine Learning. PMLR, 2021: 4369-4379.
- [520] KLÜGL F, FEHLER M, HERRLER R. About the role of the environment in multi-agent simulations[C//Environments for Multi-Agent Systems: First International Workshop, E4MAS 2004, New York, NY, July 19, 2004, Revised Selected Papers 1. Springer, 2005: 127-149.
- [521] SUN C, SHEN M, HOW J P. Scaling up multiagent reinforcement learning for robotic systems: Learn an adaptive sparse communication graph[C//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 11755-11762.
- [522] SUO S, REGALADO S, CASAS S, et al. Trafficsim: Learning to simulate realistic multi-agent behaviors[C//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10400-10409.

- [523] LEIBO J Z, DUEÑEZ-GUZMAN E A, VEZHNEVETS A, et al. Scalable evaluation of multi-agent reinforcement learning with melting pot[C]//International conference on machine learning. PMLR, 2021: 6187-6199.
- [524] WANG W Z, BELIAEV M, BIYIK E, et al. Emergent prosociality in multi-agent games through gifting[C/OL]//ZHOU Z. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021. ijcai.org, 2021: 434-442. <https://doi.org/10.24963/ijcai.2021/61>.
- [525] MUGLICH D, SCHROEDER DE WITT C, VAN DER POL E, et al. Equivariant networks for zero-shot coordination [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 6410-6423.
- [526] AGAPIOU J P, VEZHNEVETS A S, DUÉÑEZ-GUZMÁN E A, et al. Melting pot 2.0[A]. 2022.
- [527] MA Z, WANG R, LI F F, et al. Elign: Expectation alignment as a multi-agent intrinsic reward[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 8304-8317.
- [528] CHRISTOFFERSEN P J K, HAUPT A A, HADFIELD-MENELL D. Get it in writing: Formal contracts mitigate social dilemmas in multi-agent RL[C/OL]//AGMON N, AN B, RICCI A, et al. Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023. ACM, 2023: 448-456. <https://dl.acm.org/doi/10.5555/3545946.3598670>.
- [529] DU Y. Cooperative multi-agent learning in a complex world: challenges and solutions[C]//Proceedings of the AAAI Conference on Artificial Intelligence: 37(13). 2023: 15436-15436.
- [530] STONE P, KAMINKA G, KRAUS S, et al. Ad hoc autonomous agent teams: Collaboration without pre-coordination[C]//Proceedings of the AAAI Conference on Artificial Intelligence: 24-(1). 2010: 1504-1509.
- [531] ALBRECHT S V, RAMAMOORTHY S. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems[C/OL]//GINI M L, SHEHORY O, ITO T, et al. International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '13, Saint Paul, MN, USA, May 6-10, 2013. IFAAMAS, 2013: 1155-1156. <http://dl.acm.org/citation.cfm?id=2485118>.
- [532] KRAFFT P, BAKER C, PENTLAND A, et al. Modeling human ad hoc coordination[C]//Proceedings of the AAAI Conference on Artificial Intelligence: 30-(1). 2016.
- [533] DENNIS M, JAQUES N, VINITSKY E, et al. Emergent complexity and zero-shot transfer via unsupervised environment design[J]. *Advances in neural information processing systems*, 2020, 33: 13049-13061.
- [534] JIANG M, DENNIS M, PARKER-HOLDER J, et al. Replay-guided adversarial environment design[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 1884-1897.
- [535] LEHMAN J, STANLEY K O, et al. Exploiting open-endedness to solve problems through the search for novelty. [C]//ALIFE. 2008: 329-336.
- [536] WANG R, LEHMAN J, CLUNE J, et al. Poet: open-ended coevolution of environments and their optimized solutions[C]//Proceedings of the Genetic and Evolutionary Computation Conference. 2019: 142-151.
- [537] BATARSEH F A, FREEMAN L, HUANG C H. A survey on artificial intelligence assurance[J]. *Journal of Big Data*, 2021, 8(1): 60.
- [538] BOLUKBASI T, CHANG K W, ZOU J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings[J]. *Advances in neural information processing systems*, 2016, 29.

- [539] ROH Y, HEO G, WHANG S E. A survey on data collection for machine learning: a big data-ai integration perspective[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2019, 33(4): 1328-1347.
- [540] RÖTTGER P, VIDGEN B, NGUYEN D, et al. Hatecheck: Functional tests for hate speech detection models [C/OL]//ZONG C, XIA F, LI W, et al. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, 2021: 41-58. <https://doi.org/10.18653/v1/2021.acl-long.4>.*
- [541] PARRISH A, CHEN A, NANGIA N, et al. BBQ: A hand-built bias benchmark for question answering[C/OL]//MURESAN S, NAKOV P, VILLAVICENCIO A. *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, 2022: 2086-2105. <https://doi.org/10.18653/v1/2022.findings-acl.165>.*
- [542] PAPERNOT N, MCDANIEL P, SINHA A, et al. *Towards the science of security and privacy in machine learning [A]. 2016.*
- [543] ZHAO J, WANG T, YATSKAR M, et al. Gender bias in coreference resolution: Evaluation and debiasing methods [C/OL]//WALKER M A, JI H, STENT A. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers). Association for Computational Linguistics, 2018: 15-20. <https://doi.org/10.18653/v1/n18-2003>.*
- [544] ZAMPIERI M, MALMASI S, NAKOV P, et al. Predicting the type and target of offensive posts in social media [C/OL]//BURSTEIN J, DORAN C, SOLORIO T. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019: 1415-1420. <https://doi.org/10.18653/v1/n19-1144>.*
- [545] NANGIA N, VANIA C, BHALERAO R, et al. Crows-pairs: A challenge dataset for measuring social biases in masked language models[C/OL]//WEBBER B, COHN T, HE Y, et al. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 1953-1967. <https://doi.org/10.18653/v1/2020.emnlp-main.154>.*
- [546] ROSENTHAL S, ATANASOVA P, KARADZHOV G, et al. SOLID: A large-scale semi-supervised dataset for offensive language identification[C/OL]//ZONG C, XIA F, LI W, et al. *Findings of ACL: ACL/IJCNLP 2021 Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021. Association for Computational Linguistics, 2021: 915-928. <https://doi.org/10.18653/v1/2021.findings-acl.80>.*
- [547] WESTON J, BORDES A, CHOPRA S, et al. *Towards ai-complete question answering: A set of prerequisite toy tasks[A]. 2015.*
- [548] ZHANG Z, CHENG J, SUN H, et al. Constructing highly inductive contexts for dialogue safety through controllable reverse generation[C/OL]//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 3684-3697. <https://aclanthology.org/2022.findings-emnlp.270>. DOI: 10.18653/v1/2022.findings-emnlp.270.
- [549] LIN Y T, CHEN Y N. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models[A]. 2023.

- [550] LI R, PATEL T, DU X. Prd: Peer rank and discussion improve large language model based evaluations[A]. 2023.
- [551] WULCZYN E, THAIN N, DIXON L. Ex machina: Personal attacks seen at scale[C]//Proceedings of the 26th international conference on world wide web. 2017: 1391-1399.
- [552] GEHMAN S, GURURANGAN S, SAP M, et al. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models[C/OL]//COHN T, HE Y, LIU Y. Findings of ACL: EMNLP 2020 Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020. Association for Computational Linguistics, 2020: 3356-3369. <https://doi.org/10.18653/v1/2020.findings-emnlp.301>.
- [553] GANGULI D, LOVITT L, KERNION J, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned[A]. 2022.
- [554] SHETH A, SHALIN V L, KURSUNCU U. Defining and detecting toxicity on social media: context and knowledge are key[J]. Neurocomputing, 2022, 490: 312-318.
- [555] JI J, LIU M, DAI J, et al. Beavertails: Towards improved safety alignment of llm via a human-preference dataset [A]. 2023.
- [556] TURNER A M. On avoiding power-seeking by artificial intelligence[A]. 2022.
- [557] MUNIR A, AVED A, BLASCH E. Situational awareness: techniques, challenges, and prospects[J]. AI, 2022, 3(1): 55-77.
- [558] FALKE T, RIBEIRO L F, UTAMA P A, et al. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 2214-2220.
- [559] DHINGRA B, FARUQUI M, PARIKH A P, et al. Handling divergent reference texts when evaluating table-to-text generation[C/OL]//KORHONEN A, TRAUM D R, MÀRQUEZ L. Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 4884-4895. <https://doi.org/10.18653/v1/p19-1483>.
- [560] TIAN R, NARAYAN S, SELLAM T, et al. Sticking to the facts: Confident decoding for faithful data-to-text generation[A]. 2019.
- [561] WANG Z, WANG X, AN B, et al. Towards faithful neural table-to-text generation with content-matching constraints [C/OL]//JURAFSKY D, CHAI J, SCHLUTER N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020: 1072-1086. <https://doi.org/10.18653/v1/2020.acl-main.101>.
- [562] HONOVICH O, CHOSHEN L, AHARONI R, et al. \hat{q}^2 : Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering[C/OL]//MOENS M, HUANG X, SPECIA L, et al. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics, 2021: 7856-7870. <https://doi.org/10.18653/v1/2021.emnlp-main.619>.
- [563] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

- [564] LENTZOS F. Ai and biological weapons[M]//Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm. Springer, 2022: 91-100.
- [565] DATHATHRI S, MADOTTO A, LAN J, et al. Plug and play language models: A simple approach to controlled text generation[C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. <https://openreview.net/forum?id=H1edEyBKDS>.
- [566] KRAUSE B, GOTMARE A D, MCCANN B, et al. GeDi: Generative discriminator guided sequence generation [C/OL]//Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 4929-4952. <https://aclanthology.org/2021.findings-emnlp.424>. DOI: 10.18653/v1/2021.findings-emnlp.424.
- [567] DENG M, WANG J, HSIEH C P, et al. RLPrompt: Optimizing discrete text prompts with reinforcement learning [C/OL]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 3369-3391. <https://aclanthology.org/2022.emnlp-main.222>. DOI: 10.18653/v1/2022.emnlp-main.222.
- [568] LEE D, MOON S, LEE J, et al. Query-efficient and scalable black-box adversarial attacks on discrete sequential data via Bayesian optimization[C/OL]//CHAUDHURI K, JEGELKA S, SONG L, et al. Proceedings of Machine Learning Research: Vol. 162 Proceedings of the 39th International Conference on Machine Learning. PMLR, 2022: 12478-12497. <https://proceedings.mlr.press/v162/lee22h.html>.
- [569] JONES E, DRAGAN A D, RAGHUNATHAN A, et al. Automatically auditing large language models via discrete optimization[C/OL]//KRAUSE A, BRUNSKILL E, CHO K, et al. Proceedings of Machine Learning Research: Vol. 202 International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. PMLR, 2023: 15307-15329. <https://proceedings.mlr.press/v202/jones23a.html>.
- [570] WALLACE E, FENG S, KANDPAL N, et al. Universal adversarial triggers for attacking and analyzing NLP [C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 2153-2162. <https://aclanthology.org/D19-1221>. DOI: 10.18653/v1/D19-1221.
- [571] SHEN X, CHEN Z, BACKES M, et al. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models[A]. 2023.
- [572] WEI A, HAGHTALAB N, STEINHARDT J. Jailbroken: How does llm safety training fail?[A]. 2023.
- [573] LIU Y, DENG G, XU Z, et al. Jailbreaking chatgpt via prompt engineering: An empirical study[A]. 2023.
- [574] XU J, JU D, LI M, et al. Recipes for safety in open-domain chatbots[A]. 2020.
- [575] XU J, JU D, LI M, et al. Bot-adversarial dialogue for safe conversational agents[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 2950-2968.
- [576] EBRAHIMI J, RAO A, LOWD D, et al. HotFlip: White-box adversarial examples for text classification[C/OL]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 31-36. <https://aclanthology.org/P18-2006>. DOI: 10.18653/v1/P18-2006.

- [577] CHENG M, YI J, CHEN P, et al. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples[C/OL]//The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 2020: 3601-3608. <https://doi.org/10.1609/aaai.v34i04.5767>.
- [578] ZANG Y, QI F, YANG C, et al. Word-level textual adversarial attacking as combinatorial optimization[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 6066-6080. <https://aclanthology.org/2020.acl-main.540>. DOI: 10.18653/v1/2020.acl-main.540.
- [579] ZHAO Y, PANG T, DU C, et al. On evaluating adversarial robustness of large vision-language models[A]. 2023.
- [580] BROWN T B, CARLINI N, ZHANG C, et al. Unrestricted adversarial examples[A]. 2018.
- [581] FABIAN D. Google's ai red team: the ethical hackers making ai safer[EB/OL]. 2023[2023-07-19]. <https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer>.
- [582] PEARCE W, LUCAS J. Nvidia ai red team: An introduction[EB/OL]. 2023[2023-06-14]. <https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction>.
- [583] Ram Shankar Siva Kumar. Microsoft ai red team building future of safer ai[EB/OL]. 2023[2023-08-07]. <https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai>.
- [584] BUÇINCA Z, PHAM C M, JAKESCH M, et al. Aha!: Facilitating ai impact assessment by generating examples of harms[A]. 2023. arXiv: 2306.03280.
- [585] VERMA A K, AJIT S, KARANKI D R, et al. Reliability and safety engineering: Vol. 43[M]. Springer, 2010.
- [586] STEINHARDT J. Long-term and short-term challenges to ensuring the safety of ai systems[Z]. 2015.
- [587] STAHL B C, LEACH T. Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: an empirical test of the european union assessment list for trustworthy ai (altai)[J]. AI and Ethics, 2022: 1-23.
- [588] CELIKYILMAZ A, CLARK E, GAO J. Evaluation of text generation: A survey[A]. 2020.
- [589] PAULLADA A, RAJI I D, BENDER E M, et al. Data and its (dis) contents: A survey of dataset development and use in machine learning research[J]. Patterns, 2021, 2(11).
- [590] SAI A B, MOHANKUMAR A K, KHAPRA M M. A survey of evaluation metrics used for nlg systems[J]. ACM Computing Surveys (CSUR), 2022, 55(2): 1-39.
- [591] YUEN M C, KING I, LEUNG K S. A survey of crowdsourcing systems[C]//2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. IEEE, 2011: 766-773.
- [592] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration[C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. <https://openreview.net/forum?id=rygGQyrFvH>.

-
- [593] ENGSTROM L, ILYAS A, SANTURKAR S, et al. Identifying statistical bias in dataset replication[C]// International Conference on Machine Learning. PMLR, 2020: 2922-2932.
- [594] CABRERA Á A, FU E, BERTUCCI D, et al. Zeno: An interactive framework for behavioral evaluation of machine learning[C]//Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023: 1-14.
- [595] HADA R, GUMMA V, DE WYNTER A, et al. Are large language model-based evaluators the solution to scaling up multilingual evaluation?[A]. 2023.
- [596] LIU X, YU H, ZHANG H, et al. Agentbench: Evaluating llms as agents[A]. 2023.
- [597] SALEIRO P, KUESTER B, HINKSON L, et al. Aequitas: A bias and fairness audit toolkit[A]. 2018.
- [598] RUDINGER R, NARADOWSKY J, LEONARD B, et al. Gender bias in coreference resolution[A]. 2018.
- [599] KIRITCHENKO S, MOHAMMAD S M. Examining gender and race bias in two hundred sentiment analysis systems[C/OL]//NISSIM M, BERANT J, LENCI A. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018. Association for Computational Linguistics, 2018: 43-53. <https://doi.org/10.18653/v1/s18-2005>.
- [600] WEBSTER K, RECASENS M, AXELROD V, et al. Mind the gap: A balanced corpus of gendered ambiguous pronouns[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 605-617.
- [601] NADEEM M, BETHKE A, REDDY S. Stereoset: Measuring stereotypical bias in pretrained language models [C/OL]//ZONG C, XIA F, LI W, et al. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, 2021: 5356-5371. <https://doi.org/10.18653/v1/2021.acl-long.416>.
- [602] LIANG P P, WU C, MORENCY L P, et al. Towards understanding and mitigating social biases in language models [C]//International Conference on Machine Learning. PMLR, 2021: 6565-6576.
- [603] DANCETTE C, CADENE R, TENNEY D, et al. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 1574-1583.
- [604] LANDERS R N, BEHREND T S. Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models.[J]. American Psychologist, 2023, 78(1): 36.
- [605] HARTVIGSEN T, GABRIEL S, PALANGI H, et al. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection[C/OL]//MURESAN S, NAKOV P, VILLAVICENCIO A. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, 2022: 3309-3326. <https://doi.org/10.18653/v1/2022.acl-long.234>.
- [606] SANNEMAN L, SHAH J A. A situation awareness-based framework for design and evaluation of explainable ai [C]//Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2. Springer, 2020: 94-110.
- [607] LI Y, DU Y, ZHOU K, et al. Evaluating object hallucination in large vision-language models[A]. 2023.

- [608] ZHANG Y, LI Y, CUI L, et al. Siren’s song in the ai ocean: A survey on hallucination in large language models [A]. 2023.
- [609] KINNIMENT M, KOBAYASHI S, DU H, et al. Evaluating language-model agents on realistic autonomous tasks [EB/OL]. 2023. <https://evals.alignment.org/language-model-pilot-report>.
- [610] VON STENGEL B, KOLLER D. Team-maxmin equilibria[J]. *Games and Economic Behavior*, 1997, 21(1-2): 309-321.
- [611] COHEN F. Managing network security—red teaming[J]. *Network Security*, 1998, 1998(3): 13-15.
- [612] CASPER S, LIN J, KWON J, et al. Explore, establish, exploit: Red teaming language models from scratch[A]. 2023. arXiv: 2306.09442.
- [613] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning[A]. 2017.
- [614] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J/OL]. *Nature Machine Intelligence*, 2019, 1(5): 206-215. <https://www.nature.com/articles/s42256-019-0048-x>. DOI: 10.1038/s42256-019-0048-x.
- [615] PAVLUS J. A new approach to understanding how machines think[J]. *Quanta Magazine*, 2019.
- [616] OLAH C, CAMMARATA N, SCHUBERT L, et al. Zoom in: An introduction to circuits[J]. *Distill*, 2020, 5(3): e00024-001.
- [617] OLAH C. Interpretability dreams[J/OL]. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/interpretability-dreams/index.html#larger-scale>.
- [618] CARVALHO D V, PEREIRA E M, CARDOSO J S. Machine learning interpretability: A survey on methods and metrics[J]. *Electronics*, 2019, 8(8): 832.
- [619] RUSU A A, RABINOWITZ N C, DESJARDINS G, et al. Progressive neural networks[A]. 2022. arXiv: 1606.04671.
- [620] SMITH J S, TIAN J, HALBE S, et al. A closer look at rehearsal-free continual learning[C/OL]//*IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023: 2410-2420. <https://doi.org/10.1109/CVPRW59228.2023.00239>.
- [621] FRANKLE J, CARBIN M. The lottery ticket hypothesis: Finding sparse, trainable neural networks[C/OL]//*7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. <https://openreview.net/forum?id=rJl-b3RcF7>.
- [622] HOEFLER T, ALISTARH D, BEN-NUN T, et al. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks[J/OL]. *J. Mach. Learn. Res.*, 2021, 22: 241:1-241:124. <http://jmlr.org/papers/v22/21-0366.html>.
- [623] MEISTER C, LAZOV S, AUGENSTEIN I, et al. Is sparse attention more interpretable?[C/OL]//ZONG C, XIA F, LI W, et al. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 2021: 122-129. <https://doi.org/10.18653/v1/2021.acl-short.17>.

- [624] WONG E, SANTURKAR S, MADRY A. Leveraging sparse linear layers for debuggable deep networks[C/OL]// MEILA M, ZHANG T. Proceedings of Machine Learning Research: Vol. 139 Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. PMLR, 2021: 11205-11216. <http://proceedings.mlr.press/v139/wong21b.html>.
- [625] BÉNA G, GOODMAN D F M. Dynamics of specialization in neural modules under resource constraints[A]. 2023. arXiv: 2106.02626.
- [626] ALVAREZ-MELIS D, JAAKKOLA T S. Towards robust interpretability with self-explaining neural networks [C/OL]//BENGIO S, WALLACH H M, LAROCHELLE H, et al. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. 2018: 7786-7795. <https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html>.
- [627] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828. DOI: 10.1109/TPAMI.2013.50.
- [628] SALMAN H, ILYAS A, ENGSTROM L, et al. Do adversarially robust imagenet models transfer better?[C/OL]// LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020. <https://proceedings.neurips.cc/paper/2020/hash/24357dd085d2c4b1a88a7e0692e60294-Abstract.html>.
- [629] IMPACT B. Alignment course[J/OL]. AI Safety Fundamentals, 2023. <https://course.aisafetyfundamentals.com/alignment>.
- [630] ZOU A, PHAN L, CHEN S, et al. Representation engineering: A top-down approach to ai transparency[A]. 2023. arXiv: 2310.01405.
- [631] MENG K, SHARMA A S, ANDONIAN A J, et al. Mass-editing memory in a transformer[C/OL]//The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. <https://openreview.net/pdf?id=MkbcAHlYgyS>.
- [632] OPENAI. Curve detectors[J/OL]. Circuits Thread, 2021. <https://distill.pub/2020/circuits/curve-detectors/>.
- [633] PURVES D, AUGUSTINE G J, FITZPATRICK D, et al. Neuroscience, 2nd edition[M]. Sinauer Associates, 2001.
- [634] OLSSON C, ELHAGE N, NANDA N, et al. In-context learning and induction heads[A]. 2022.
- [635] WANG K R, VARIENGIEN A, CONMY A, et al. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small[C/OL]//The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. <https://openreview.net/pdf?id=NpsVSN6o4ul>.
- [636] HANNA M, LIU O, VARIENGIEN A. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model[A]. 2023.
- [637] HEIMERSHEIM S, JETT J. A circuit for Python docstrings in a 4-layer attention-only transformer[J/OL]. AI Alignment Forum, 2023[2023-09-07]. <https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/a-circuit-for-python-docstrings-in-a-4-layer-attention-only>.
- [638] NANDA N, CHAN L, LIEBERUM T, et al. Progress measures for grokking via mechanistic interpretability[C/OL]// The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. <https://openreview.net/pdf?id=9XFSbDPmdW>.

- [639] ANCONA M, CEOLINI E, ÖZTIRELI C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=Sy21R9JAW>.
- [640] DURRANI N, SAJJAD H, DALVI F, et al. Analyzing individual neurons in pre-trained language models[C/OL]//WEBBER B, COHN T, HE Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 4865-4880. <https://doi.org/10.18653/v1/2020.emnlp-main.395>.
- [641] LUNDSTRÖM D, HUANG T, RAZAVIYAYN M. A rigorous study of integrated gradients method and extensions to internal neuron attributions[C/OL]//CHAUDHURI K, JEGELKA S, SONG L, et al. Proceedings of Machine Learning Research: Vol. 162 International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. PMLR, 2022: 14485-14508. <https://proceedings.mlr.press/v162/lundstrom22a.html>.
- [642] DAR G, GEVA M, GUPTA A, et al. Analyzing transformers in embedding space[C/OL]//ROGERS A, BOYD-GRABER J L, OKAZAKI N. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. Association for Computational Linguistics, 2023: 16124-16170. <https://doi.org/10.18653/v1/2023.acl-long.893>.
- [643] DAI D, DONG L, HAO Y, et al. Knowledge neurons in pretrained transformers[C/OL]//MURESAN S, NAKOV P, VILLAVICENCIO A. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, 2022: 8493-8502. <https://doi.org/10.18653/v1/2022.acl-long.581>.
- [644] MCGRATH T, RAHTZ M, KRAMAR J, et al. The hydra effect: Emergent self-repair in language model computations[A]. 2023. arXiv: 2307.15771.
- [645] RAGER C, LAU Y T, DAO J, et al. An adversarial example for direct logit attribution: memory management in gelu-4l[Z]. 2023.
- [646] LIEBERUM T, RAHTZ M, KRAMÁR J, et al. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla[A]. 2023. arXiv: 2307.09458.
- [647] NANDA N. Attribution patching: Activation patching at industrial scale[Z]. 2023.
- [648] BELROSE N, FURMAN Z, SMITH L, et al. Eliciting latent predictions from transformers with the tuned lens[A]. 2023. arXiv: 2303.08112.
- [649] ERHAN D, BENGIO Y, COURVILLE A, et al. Visualizing higher-layer features of a deep network[J]. University of Montreal, 2009, 1341.
- [650] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[J]. Proceedings of the European conference on computer vision (ECCV), 2014.
- [651] SIMONYAN K, et al. Deep inside convolutional networks: Visualising image classification models and saliency maps[A]. 2013.
- [652] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

- [653] KARPATY A, JOHNSON J, LI F F. Visualizing and understanding recurrent networks[A]. 2015.
- [654] MORDVINTSEV A, OLAH C, TYKA M. Inceptionism: Going deeper into neural networks[J]. Google Research Blog, 2015.
- [655] NGUYEN A, YOSINSKI J, CLUNE J. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks[A]. 2016.
- [656] KINDERMANS P, SCHÜTT K T, ALBER M, et al. Learning how to explain neural networks: Patternnet and patternattribution[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=Hkn7CBaTW>.
- [657] OLAH C, et al. Feature visualization[J/OL]. Distill, 2017. DOI: 10.23915/distill.00007.
- [658] CARTER S, ARMSTRONG Z, SCHUBERT L, et al. Activation atlas[J/OL]. Distill, 2019. <https://distill.pub/2019/activation-atlas/>. DOI: 10.23915/distill.00016.
- [659] REIF E, YUAN A, WATTENBERG M, et al. Visualizing and measuring the geometry of BERT[C/OL]//WALLACH H M, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. 2019: 8592-8600. <https://proceedings.neurips.cc/paper/2019/hash/159c1ffe5b61b41b3c4d8f4c2150f6c4-Abstract.html>.
- [660] OLAH C, CAMMARATA N, SCHUBERT L, et al. Visualizing weights[J/OL]. Distill, 2020. <https://distill.pub/2020/circuits/visualizing-weights/>. DOI: 10.23915/distill.00021.
- [661] IVANOV M, KADIKIS R, OZOLS K. Perturbation-based methods for explaining deep neural networks: A survey [J]. Pattern Recognition Letters, 2021.
- [662] MORCOS A S, BARRETT D G T, RABINOWITZ N C, et al. On the importance of single directions for generalization[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. <https://openreview.net/forum?id=r1iuQjxCZ>.
- [663] ZHOU B, SUN Y, BAU D, et al. Revisiting the importance of individual units in cnns via ablation[A/OL]. 2018. <https://arxiv.org/abs/1806.02891>.
- [664] HOD S, CASPER S, FILAN D, et al. Quantifying local specialization in deep neural networks[A/OL]. 2021. <https://arxiv.org/abs/2110.08058>.
- [665] RAVFOGEL S, TWITON M, GOLDBERG Y, et al. Linear adversarial concept erasure[C]//International Conference on Machine Learning. PMLR, 2022: 18400-18421.
- [666] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[A]. 2016. arXiv: 1409.0473.
- [667] LEE J, SHIN J H, KIM J S. Interactive visualization and manipulation of attention-based neural machine translation [C/OL]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 121-126. <https://aclanthology.org/D17-2021>. DOI: 10.18653/v1/D17-2021.

- [668] FONG R, VEDALDI A. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks[C/OL]//2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. Computer Vision Foundation / IEEE Computer Society, 2018: 8730-8738. http://openaccess.thecvf.com/content_cvpr_2018/html/Fong_Net2Vec_Quantifying_and_CVPR_2018_paper.html. DOI: 10.1109/CVPR.2018.00910.
- [669] STROBELT H, GEHRMANN S, BEHRISCH M, et al. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models[J/OL]. IEEE Trans. Vis. Comput. Graph., 2019, 25(1): 353-363. <https://doi.org/10.1109/TVCG.2018.2865044>.
- [670] LIU S, LI T, LI Z, et al. Visual interrogation of attention-based models for natural language inference and machine comprehension[C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels, Belgium: Association for Computational Linguistics, 2018: 36-41. <https://aclanthology.org/D18-2007>. DOI: 10.18653/v1/D18-2007.
- [671] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)[C/OL]//DY J G, KRAUSE A. Proceedings of Machine Learning Research: Vol. 80 Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. PMLR, 2018: 2673-2682. <http://proceedings.mlr.press/v80/kim18d.html>.
- [672] CLARK K, KHANDELWAL U, LEVY O, et al. What does BERT look at? an analysis of bert's attention[C/OL]//LINZEN T, CHRUPALA G, BELINKOV Y, et al. Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019. Association for Computational Linguistics, 2019: 276-286. <https://doi.org/10.18653/v1/W19-4828>.
- [673] VIG J. A multiscale visualization of attention in the transformer model[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Florence, Italy: Association for Computational Linguistics, 2019: 37-42. <https://aclanthology.org/P19-3007>. DOI: 10.18653/v1/P19-3007.
- [674] VASHISHTH S, UPADHYAY S, TOMAR G S, et al. Attention interpretability across nlp tasks[A]. 2019. arXiv: 1909.11218.
- [675] HAO Y, DONG L, WEI F, et al. Self-attention attribution: Interpreting information interactions inside transformer[C/OL]//Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, 2021: 12963-12971. <https://doi.org/10.1609/aaai.v35i14.17533>.
- [676] CHEFER H, GUR S, WOLF L. Transformer interpretability beyond attention visualization[C/OL]//IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021: 782-791. https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html. DOI: 10.1109/CVPR46437.2021.00084.
- [677] SCHNEIDER J, VLACHOS M. Explaining neural networks by decoding layer activations[C/OL]//ABREU P H, RODRIGUES P P, FERNÁNDEZ A, et al. Lecture Notes in Computer Science: Vol. 12695 Advances in Intelligent Data Analysis XIX - 19th International Symposium on Intelligent Data Analysis, IDA 2021, Porto, Portugal, April 26-28, 2021, Proceedings. Springer, 2021: 63-75. https://doi.org/10.1007/978-3-030-74251-5_6.

- [678] BANSAL Y, NAKKIRAN P, BARAK B. Revisiting model stitching to compare neural representations[C/OL]// RANZATO M, BEYGEZIMMER A, DAUPHIN Y N, et al. Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. 2021: 225-236. <https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html>.
- [679] BELINKOV Y. Probing classifiers: Promises, shortcomings, and advances[J/OL]. *Comput. Linguistics*, 2022, 48 (1): 207-219. https://doi.org/10.1162/coli_a_00422.
- [680] GEVA M, SCHUSTER R, BERANT J, et al. Transformer feed-forward layers are key-value memories[C/OL]// MOENS M, HUANG X, SPECIA L, et al. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics, 2021: 5484-5495. <https://doi.org/10.18653/v1/2021.emnlp-main.446>.
- [681] LI B Z, NYE M I, ANDREAS J. Implicit representations of meaning in neural language models[C/OL]// ZONG C, XIA F, LI W, et al. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, 2021: 1813-1827. <https://doi.org/10.18653/v1/2021.acl-long.143>.
- [682] POWER A, BURDA Y, EDWARDS H, et al. Grokking: Generalization beyond overfitting on small algorithmic datasets[A]. 2022.
- [683] RIGOTTI M, MIKSOVIC C, GIURGIU I, et al. Attention-based interpretability with concept transformers[C/OL]// International Conference on Learning Representations. 2022. <https://openreview.net/forum?id=kAa9eDS0RdO>.
- [684] GEVA M, CACIULARU A, WANG K R, et al. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space[C/OL]// GOLDBERG Y, KOZAREVA Z, ZHANG Y. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. Association for Computational Linguistics, 2022: 30-45. <https://doi.org/10.18653/v1/2022.emnlp-main.3>.
- [685] GEVA M, CACIULARU A, DAR G, et al. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models[C/OL]// CHE W, SHUTOVA E. Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 - System Demonstrations, Abu Dhabi, UAE, December 7-11, 2022. Association for Computational Linguistics, 2022: 12-21. <https://doi.org/10.18653/v1/2022.emnlp-demos.2>.
- [686] WANG F, RUDIN C. Falling rule lists[A/OL]. 2015. <https://arxiv.org/abs/1510.05175>.
- [687] CARUANA R, LOU Y, GEHRKE J, et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission[C/OL]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1721-1730. <https://doi.org/10.1145/2783258.2788613>.
- [688] HENDRYCKS D, GIMPEL K. Gaussian error linear units (gelus)[A/OL]. 2016. <https://arxiv.org/abs/1606.08415>.
- [689] KLAMBAUER G, UNTERTHINER T, MAYR A, et al. Self-normalizing neural networks[J/OL]. *Advances in Neural Information Processing Systems*, 2017, 30. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547de91fbd053c1c4a845aa-Abstract.html>.

- [690] ANTHROPIC. Softmax Linear Units[J/OL]. Transformer Circuits Thread, 2022[2023-09-07]. <https://transformer-circuits.pub/2022/solu/index.html>.
- [691] QUIAN R, REDDY L, KREIMAN G. Invariant visual representation by single neurons in the human brain[J]. Nature, 2005.
- [692] OPENAI. Weight banding[J/OL]. Circuits Thread, 2021. <https://distill.pub/2020/circuits/weight-banding/>.
- [693] OLAH C, CAMMARATA N, VOSS C, et al. Naturally occurring equivariance in neural networks[J]. Distill, 2020, 5(12): e00024-004.
- [694] GOH G, CAMMARATA N, VOSS C, et al. Multimodal neurons in artificial neural networks[J]. Distill, 2021, 6(3): e30.
- [695] VOSS C, GOH G, CAMMARATA N, et al. Branch Specialization[J/OL]. Distill, 2021, 6(4): e00024.008[2023-08-17]. <https://distill.pub/2020/circuits/branch-specialization>. DOI: 10.23915/distill.00024.008.
- [696] SCHUBERT L, VOSS C, OLAH C. High/Low frequency detectors[J/OL]. Distill, 2021, 6(1): 10.23915/distill.00024.005[2023-09-05]. <https://distill.pub/2020/circuits/frequency-edges>. DOI: 10.23915/distill.00024.005.
- [697] ARORA S, LI Y, LIANG Y, et al. Linear algebraic structure of word senses, with applications to polysemy[J/OL]. Trans. Assoc. Comput. Linguistics, 2018, 6: 483-495. https://doi.org/10.1162/tacl_a_00034.
- [698] ELHAGE N, HUME T, OLSSON C, et al. Toy models of superposition[A]. 2022.
- [699] BRICKEN T, TEMPLETON A, BATSON J, et al. Towards monosemanticity: Decomposing language models with dictionary learning[J]. Transformer Circuits Thread, 2023.
- [700] NANDA N. Othello-gpt: Future work i am excited about[J/OL]. AI Alignment Forum, 2023. <https://www.alignmentforum.org/posts/qgK7smTvJ4DB8rZ6h/othello-gpt-future-work-i-am-excited-about>.
- [701] CHUGHTAI B, CHAN L, NANDA N. A toy model of universality: Reverse engineering how networks learn group operations[C/OL]//KRAUSE A, BRUNSKILL E, CHO K, et al. Proceedings of Machine Learning Research: Vol. 202 International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. PMLR, 2023: 6243-6267. <https://proceedings.mlr.press/v202/chughtai23a.html>.
- [702] CONMY A, MAVOR-PARKER A N, LYNCH A, et al. Towards automated circuit discovery for mechanistic interpretability[A]. 2023.
- [703] ANTHROPIC. Circuits updates - july 2023[J/OL]. Transformer Circuits Thread, 2023. <https://transformer-circuits.pub/2023/july-update/index.html>.
- [704] SEGERIE C R. Against almost every theory of impact of interpretability[EB/OL]. 2023. <https://www.alignmentforum.org/posts/LNA8mubrByG7SFacm/against-almost-every-theory-of-impact-of-interpretability-1>.
- [705] VAN DER MAATEN L, HINTON G E. Visualizing data using t-sne[J/OL]. Journal of Machine Learning Research, 2008, 9: 2579-2605. <https://api.semanticscholar.org/CorpusID:5855042>.
- [706] OLAH C. Visualizing mnist: An exploration of dimensionality reduction[EB/OL]. 2014. <https://colah.github.io/posts/2014-10-Visualizing-MNIST/>.

- [707] OLAH C. Visualizing representations: Deep learning and human beings[EB/OL]. 2015. <https://colah.github.io/posts/2015-01-Visualizing-Representations/>.
- [708] CASPER S. Moving Forward: 11th post of The Engineer' s Interpretability Sequence[EB/OL]. 2023. <https://www.alignmentforum.org/posts/L5Rua9aTndvly8dvc/eis-xi-moving-forward>.
- [709] CASPER S, LI Y, LI J, et al. Red Teaming Deep Neural Networks with Feature Synthesis Tools[M/OL]. arXiv, 2023. <http://arxiv.org/abs/2302.10894>. DOI: 10.48550/arXiv.2302.10894.
- [710] LAWRENCE C, GARRIGA-ALONSO A, GOLDOWSKY-DILL N, et al. Causal Scrubbing: a method for rigorously testing interpretability hypotheses [Redwood Research][J/OL]. AI Alignment Forum, 2023[2023-09-07]. <https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing>.
- [711] VON WRIGHT G H. Deontic logic[J]. *Mind*, 1951, 60(237): 1-15.
- [712] DUBLJEVIĆ V, RACINE E. The adc of moral judgment: Opening the black box of moral intuitions with heuristics about agents, deeds, and consequences[J]. *AJOB Neuroscience*, 2014, 5(4): 3-20.
- [713] MERMET B, SIMON G. Formal verication of ethical properties in multiagent systems[C]//1st Workshop on Ethics in the Design of Intelligent Agents. 2016.
- [714] PEREIRA L M, SAPTAWIJAYA A, et al. Programming machine ethics: Vol. 26[M]. Springer, 2016.
- [715] DENNIS L, FISHER M, SLAVKOVİK M, et al. Formal verification of ethical choices in autonomous systems[J]. *Robotics and Autonomous Systems*, 2016, 77: 1-14.
- [716] BERREBY F, BOURGNE G, GANASCIA J G. A declarative modular framework for representing and applying ethical principles[C]//16th Conference on Autonomous Agents and MultiAgent Systems. 2017.
- [717] DUBLJEVIC V. Toward implementing the agent-deed-consequence model of moral judgment in autonomous vehicles [C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020: 243-243.
- [718] ABEL D, MACGLASHAN J, LITTMAN M L. Reinforcement learning as a framework for ethical decision making. [C]//AAAI Workshop: AI, Ethics, and Society: Vol. 16. Menlo Park, CA, USA: The AAAI Press, 2016: 02.
- [719] WU Y H, LIN S D. A low-cost ethics shaping approach for designing reinforcement learning agents[C]//Proceedings of the AAAI conference on artificial intelligence: 32(1). 2018.
- [720] SVEGLIATO J, NASHED S B, ZILBERSTEIN S. Ethically compliant sequential decision making[C]//Proceedings of the AAAI Conference on Artificial Intelligence: 35(13). 2021: 11657-11665.
- [721] MURTARELLI G, GREGORY A, ROMENTI S. A conversation-based perspective for shaping ethical human-machine interactions: The particular challenge of chatbots[J]. *Journal of Business Research*, 2021, 129: 927-935.
- [722] ROSSI F, VENABLE K B, WALSH T. A short introduction to preferences: Between ai and social choice[M]. Morgan & Claypool Publishers, 2011.
- [723] PEREIRA L M, SAPTAWIJAYA A, PEREIRA L M, et al. Bridging two realms of machine ethics[J]. *Programming machine ethics*, 2016: 159-165.
- [724] CONITZER V, SINNOTT-ARMSTRONG W, BORG J S, et al. Moral decision making frameworks for artificial intelligence[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 31. 2017.

- [725] NOOTHIGATTU R, GAIKWAD S, AWAD E, et al. A voting-based system for ethical decision making[C]// Proceedings of the AAAI Conference on Artificial Intelligence: 32(1). 2018.
- [726] HAN T A, PEREIRA L M. Evolutionary machine ethics[J]. *Handbuch maschinenethik*, 2019: 229-253.
- [727] STACKELBERG H V. Marktform und gleichgewicht[J]. Verlag von Julius Springer, 1934.
- [728] PITA J, JAIN M, TAMBE M, et al. Robust solutions to stackelberg games: Addressing bounded rationality and limited observations in human cognition[J]. *Artificial Intelligence*, 2010, 174(15): 1142-1171.
- [729] LI T, SETHI S P. A review of dynamic stackelberg game models[J]. *Discrete & Continuous Dynamical Systems-B*, 2017, 22(1): 125.
- [730] FIEZ T, CHASNOV B, RATLIFF L J. Convergence of learning dynamics in stackelberg games[A]. 2019.
- [731] FIEZ T, CHASNOV B, RATLIFF L. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study[C]//International Conference on Machine Learning. PMLR, 2020: 3133-3144.
- [732] MCKEE K R, GEMP I, MCWILLIAMS B, et al. Social diversity and social preferences in mixed-motive reinforcement learning[C/OL]//SEGHRUCHNI A E F, SUKTHANKAR G, AN B, et al. Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020. International Foundation for Autonomous Agents and Multiagent Systems, 2020: 869-877. <https://dl.acm.org/doi/10.5555/3398761.3398863>.
- [733] OESTERHELD C, CONITZER V. Safe pareto improvements for delegated game playing[J]. *Autonomous Agents and Multi-Agent Systems*, 2022, 36(2): 46.
- [734] AXELROD R, HAMILTON W D. The evolution of cooperation[J]. *Science*, 1981, 211(4489): 1390-1396.
- [735] SCHUSTER P, SIGMUND K. Replicator dynamics[J]. *Journal of theoretical biology*, 1983, 100(3): 533-538.
- [736] WEIBULL J W. Evolutionary game theory[M]. MIT press, 1997.
- [737] SANTOS M D, PINHEIRO F L, SANTOS F C, et al. Dynamics of n-person snowdrift games in structured populations[J]. *Journal of Theoretical Biology*, 2012, 315: 81-86.
- [738] DIGIOVANNI A, MACÉ N, CLIFTON J. Evolutionary stability of other-regarding preferences under complexity costs[A]. 2022.
- [739] MIN B, ROSS H, SULEM E, et al. Recent advances in natural language processing via large pre-trained language models: A survey[J]. *ACM Computing Surveys*, 2023, 56(2): 1-40.
- [740] AWAD E, DSOUZA S, KIM R, et al. The moral machine experiment[J]. *Nature*, 2018, 563(7729): 59-64.
- [741] HAGENDORFF T. A virtue-based framework to support putting ai ethics into practice[J]. *Philosophy & Technology*, 2022, 35(3): 55.
- [742] ABDULHAI M, CREPY C, VALTER D, et al. Moral foundations of large language models[C]//AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI. 2022.

- [743] PAN A, CHAN J S, ZOU A, et al. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark[C]//International Conference on Machine Learning. PMLR, 2023: 26837-26867.
- [744] SCHERRER N, SHI C, FEDER A, et al. Evaluating the moral beliefs encoded in llms[A]. 2023.
- [745] LIU R, YANG R, JIA C, et al. Training socially aligned language models in simulated human society[A]. 2023.
- [746] DURMUS E, NYUGEN K, LIAO T I, et al. Towards measuring the representation of subjective global opinions in language models[A]. 2023.
- [747] ZHANG Z, LIU N, QI S, et al. Heterogeneous value evaluation for large language models[A]. 2023.
- [748] ZHANG Z, BAI F, GAO J, et al. Measuring value understanding in language models through discriminator-critique gap[A]. 2023.
- [749] IEEE. The ieee global initiative on ethics of autonomous and intelligent systems[EB/OL]. 2016. <https://standards.ieee.org/industry-connections/ec/autonomous-systems/>.
- [750] LEE S, LEE M, LEE S. What if artificial intelligence become completely ambient in our daily lives? exploring future human-ai interaction through high fidelity illustrations[J]. *International Journal of Human-Computer Interaction*, 2023, 39(7): 1371-1389.
- [751] WEIDINGER L, MCKEE K R, EVERETT R, et al. Using the veil of ignorance to align ai systems with principles of justice[J]. *Proceedings of the National Academy of Sciences*, 2023, 120(18): e2213709120.
- [752] GOODHART C A, GOODHART C. *Problems of monetary management: the uk experience*[M]. Springer, 1984.
- [753] EMELIN D, BRAS R L, HWANG J D, et al. Moral stories: Situated reasoning about norms, intents, actions, and their consequences[C/OL]//MOENS M, HUANG X, SPECIA L, et al. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics*, 2021: 698-718. <https://doi.org/10.18653/v1/2021.emnlp-main.54>. DOI: 10.18653/V1/2021.EMNLP-MAIN.54.
- [754] FORBES M, HWANG J D, SHWARTZ V, et al. Social chemistry 101: Learning to reason about social and moral norms[C/OL]//WEBBER B, COHN T, HE Y, et al. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics*, 2020: 653-670. <https://doi.org/10.18653/v1/2020.emnlp-main.48>. DOI: 10.18653/V1/2020.EMNLP-MAIN.48.
- [755] ZIEMS C, YU J A, WANG Y, et al. The moral integrity corpus: A benchmark for ethical dialogue systems[C/OL]//MURESAN S, NAKOV P, VILLAVICENCIO A. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics*, 2022: 3755-3773. <https://doi.org/10.18653/v1/2022.acl-long.261>. DOI: 10.18653/V1/2022.ACL-LONG.261.
- [756] CALISKAN A, BRYSON J J, NARAYANAN A. Semantics derived automatically from language corpora contain human-like biases[J]. *Science*, 2017, 356(6334): 183-186.
- [757] ACEMOGLU D, RESTREPO P. *Artificial intelligence, automation, and work*[M]//The economics of artificial intelligence: An agenda. University of Chicago Press, 2018: 197-236.

- [758] MULLIGAN C E, GODSIFF P. Datalism and data monopolies in the era of ai: A research agenda[A]. 2023.
- [759] TURCHIN A, DENKENBERGER D. Classification of global catastrophic risks connected with artificial intelligence [J]. *Ai & Society*, 2020, 35(1): 147-163.
- [760] URBINA F, LENTZOS F, INVERNIZZI C, et al. Dual use of artificial-intelligence-powered drug discovery[J]. *Nature Machine Intelligence*, 2022, 4(3): 189-191.
- [761] MCLEAN S, READ G J, THOMPSON J, et al. The risks associated with artificial general intelligence: A systematic review[J]. *Journal of Experimental & Theoretical Artificial Intelligence*, 2023, 35(5): 649-663.
- [762] WENG L. Llm powered autonomous agents[EB/OL]. 2023. <https://lilianweng.github.io/posts/2023-06-23-agent/>.
- [763] GRAVITAS S. Auto-gpt, 2023[J]. URL <https://github.com/Significant-Gravitas/Auto-GPT>. Accessed: May, 2023, 21.
- [764] OSIKA A. Gpt engineer, 2023[J]. URL <https://github.com/AntonOsika/gpt-engineer/commits>, 2023.
- [765] NAKAJIMA Y. Babyagi[J]. Python. <https://github.com/yoheinakajima/babyagi>, 2023.
- [766] MANNES A. Governance, risk, and artificial intelligence[J]. *Ai Magazine*, 2020, 41(1): 61-69.
- [767] BRADLEY P. Risk management standards and the active management of malicious intent in artificial superintelligence[J]. *AI & SOCIETY*, 2020, 35(2): 319-328.
- [768] ZHANG L. The legal positioning and hierarchical governance of generative ai[J]. *Modern Law Science*, 2023, 45 (126-141).
- [769] BRUNDAGE M, AVIN S, WANG J, et al. Toward trustworthy ai development: mechanisms for supporting verifiable claims[A]. 2020.
- [770] European Parliament. Eu ai act: first regulation on artificial intelligence[EB/OL]. 2023. <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>.
- [771] BLUMENTHAL R, HAWLEY J. Bipartisan framework for u.s. ai act[EB/OL]. 2023. <https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf>.
- [772] WHITTLESTONE J, CLARK J. Why and how governments should monitor ai development[A]. 2021.
- [773] The White House. Fact sheet: Biden-harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by ai[EB/OL]. 2023. <https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>.
- [774] BENGIO Y, RUSSELL S, MUSK E, et al. Pause giant ai experiments: An open letter[EB/OL]. 2023. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
- [775] ALAGA J, SCHUETT J. Coordinated pausing: An evaluation-based coordination scheme for frontier ai developers [A]. 2023. arXiv: 2310.00374.
- [776] ANTHROPIC. Anthropic's responsible scaling policy[EB/OL]. 2023. <https://www.anthropic.com/index/anthropics-responsible-scaling-policy>.

- [777] MÖKANDER J, SCHUETT J, KIRK H R, et al. Auditing large language models: a three-layered approach[J]. *AI and Ethics*, 2023: 1-31.
- [778] MAAS M M. Aligning ai regulation to sociotechnical change[J]. *Oxford Handbook on AI Governance* (Oxford University Press, 2022 forthcoming), 2021.
- [779] FLORIDI L, COWLS J, BELTRAMETTI M, et al. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations[EB/OL]. 2018. <https://link.springer.com/article/10.1007/s11023-018-9482-5#citeas>.
- [780] ATOMIUM-EISMD. Ai4people[EB/OL]. 2023. <https://www.eismd.eu/ai4people/>.
- [781] KERRY C F, MELTZER J P, RENDA A, et al. Strengthening international cooperation on ai, progress report [EB/OL]. 2021. <https://www.brookings.edu/articles/strengthening-international-cooperation-on-ai/>.
- [782] ERMAN E, FURENDAL M. The global governance of artificial intelligence: Some normative concerns[J]. *Moral Philosophy and Politics*, 2022, 9(2): 267-291.
- [783] ERMAN E, FURENDAL M. Artificial intelligence and the political legitimacy of global governance[J]. *Political Studies*, 2022: 00323217221126665.
- [784] GUTERRES A. Secretary-general’s remarks to the security council on artificial intelligence[EB/OL]. 2023. <https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence>.
- [785] VINUESA R, AZIZPOUR H, LEITE I, et al. The role of artificial intelligence in achieving the sustainable development goals[J]. *Nature communications*, 2020, 11(1): 1-10.
- [786] TALLBERG J, ERMAN E, FURENDAL M, et al. The global governance of artificial intelligence: Next steps for empirical and normative research[A]. 2023.
- [787] SWAUGERARCHIVE S. Software that monitors students during tests perpetuates inequality and violates their privacy[EB/OL]. 2020. <https://www.technologyreview.com/2020/08/07/1006132/software-algorithms-proctoring-online-tests-ai-ethics/>.
- [788] SARA STRATTON B D. Why we must consider the intergenerational impacts of ai[EB/OL]. 2021. <https://www.weforum.org/agenda/2021/10/why-we-must-consider-the-intergenerational-impact-of-ai/>.
- [789] NOBLE S U, DIAS B, STRATTON S C, et al. Ai regulation through an intergenerational lens[EB/OL]. 2021. https://www3.weforum.org/docs/WEF_AI_Regulation_through_an_Intergenerational_Lens_2021.pdf.
- [790] OPP R. Committing to bridging the digital divide in least developed countries[EB/OL]. 2023. <https://www.undp.org/blog/committing-bridging-digital-divide-least-developed-countries>.
- [791] SEPASSPOUR R. A reality check and a way forward for the global governance of artificial intelligence[J]. *Bulletin of the Atomic Scientists*, 2023, 79(5): 304-315.
- [792] G20. G20 ai principles[EB/OL]. 2019. https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf.
- [793] OECD. Oecd principles on artificial intelligence[EB/OL]. 2019. <https://oecd.ai/en/ai-principles>.

-
- [794] UNESCO. Recommendation on the ethics of artificial intelligence[EB/OL]. 2021. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.
- [795] TRAGER R, HARACK B, REUEL A, et al. International governance of civilian ai: A jurisdictional certification approach[A]. 2023.
- [796] SHEVLANE T, DAFOE A. The offense-defense balance of scientific knowledge: Does publishing ai research reduce misuse?[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020: 173-179.
- [797] SHAPIRO J N, SIEGEL D A. Is this paper dangerous? balancing secrecy and openness in counterterrorism[J]. Security Studies, 2010, 19(1): 66-98.
- [798] PENEDO G, MALARTIC Q, HESSLOW D, et al. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only[A]. 2023.
- [799] ZELLERS R. Why we released grover[EB/OL]. 2019. <https://thegradients.pub/why-we-released-grover/>.
- [800] MOSTAQUE E. Democratizing ai, stable diffusion & generative models[EB/OL]. 2022. <https://exchange.scale.com/public/videos/emad-mostaque-stability-ai-stable-diffusion-open-source>.
- [801] HOWARD J. Ai safety and the age of dislightenment[EB/OL]. 2023. <https://www.fast.ai/posts/2023-11-07-dislightenment.html>.
- [802] GOLDSTEIN J A, SASTRY G, MUSSER M, et al. Generative language models and automated influence operations: Emerging threats and potential mitigations[A]. 2023.
- [803] Google. Bard[EB/OL]. 2023. <https://blog.google/technology/ai/bard-google-ai-search-updates/>.
- [804] SOLAIMAN I, BRUNDAGE M, CLARK J, et al. Release strategies and the social impacts of language models[A]. 2019.
- [805] CHAVEZ P. An ai challenge: Balancing open and closed systems[EB/OL]. 2023. <https://cepa.org/article/an-ai-challenge-balancing-open-and-closed-systems/>.
- [806] HENDRYCKS D. Pragmatic ai safety[J/OL]. AI Alignment Forum, 2022. <https://www.alignmentforum.org/s/FaEBwhhe3otzYKGGQt>.
- [807] POPPER K R. The logic of scientific discovery[M]. London, England: Routledge, 1935.
- [808] ABBASS H A. Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust[J]. Cognitive Computation, 2019, 11(2): 159-171.
- [809] SHIMI A. How to diversify conceptual alignment: the model behind refine[J/OL]. AI Alignment Forum, 2022. <https://www.alignmentforum.org/posts/5uiQkyKdejX3aEHLm/how-to-diversify-conceptual-alignment-the-model-behind>.
- [810] DREXLER K E. Reframing superintelligence: Comprehensive ai services as general intelligence[J]. Future of Humanity Institute, University of Oxford, 2019.
- [811] SHAVIT Y. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring[A]. 2023.