# 对齐的理论、技术与评估

**吉嘉铭，邱天异，陈博远，杨耀东***
人工智能安全与治理中心，人工智能研究院，北京大学
中国，北京，100080
*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”(CAIS, 2023)。为了提供对齐领域的全面和最新概述，本文深入探讨了对齐的核心理论、技术和评估。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**。本文旨在为对齐研究提供全面且对初学者友好的调研。同时本文还发布并持续更新网站 www.alignmentsurvey.com，该网站提供了一系列教程、论文集和其他资源。更详尽的讨论与分析请见 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

# Theories, Techniques, and Evaluation of AI Alignment

**Jiaming Ji, Tianyi Qiu, Boyuan Chen, Yaodong Yang***
Center for AI Safety and Governance, Institute for AI, Peking University
China,Beijing, 100080
*caisg@pku.edu.cn*

## Abstract

AI alignment aims to ensure that the behavior of AI systems is consistent with human intentions and values. As the capabilities of AI systems continue to increase, the risks associated with alignment failures are also rising. Hundreds of AI experts and public figures have expressed concerns about AI risks, stating that “mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war” (CAIS, 2023). To provide a comprehensive and up-to-date overview of the field of alignment, this paper delves into the core concepts, methodology, and practice of alignment. Firstly, the paper identifies four key goals of AI alignment: Robustness, Interpretability, Controllability, and Ethicality (**RICE**). Guided by these four principles, the paper outlines the current landscape of AI alignment research and breaks it down into two key components: **forward alignment** and **backward alignment**. This paper aims to offer a comprehensive and beginner-friendly survey of alignment research. Additionally, a continuously updated website, www.alignmentsurvey.com, is released, providing a range of tutorials, collections of papers, and other resources. For more detailed discussions and analyses, please refer to https://arxiv.org/abs/2310.19852.

**Keywords:** AI Safety , AI Alignment , RICE Principles

* 通讯作者 Corresponding Author.

# 1 引言

随着人工智能系统愈发强大，它们逐渐被应用于不同领域，如基于大语言模型的智能体开发(Xi et al., 2023)，以及应用深度强化学习技术来控制核聚变 (Degrave et al., 2022)。然而，人工智能系统能力的日益提升和在高风险领域的拓展应用带来了巨大的潜在风险。先进人工智能系统（如大语言模型）已经表现出了各种不良行为（如操纵(Turner et al., 2021; Perez et al., 2023; Carroll et al., 2023a; Steinhardt, 2023; Sharma et al., 2023) 和欺骗(Park et al., 2023b)），这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)*(Christian, 2020; Bucknall and Dori-Hacohen, 2022) 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与价值观相一致(Leike et al., 2018)——它更多关注的是人工智能的意图和目标，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(Pan et al., 2022)和目标错误泛化(Shah et al., 2022)等(§1.1)。这些对齐失败进一步被模型能力所放大，体现为谄媚(Perez et al., 2023)、欺骗(Hubinger et al., 2019)和权利寻求(Power et al., 2022)等可能危害人类社会的行为。

接下来本文将阐释实现对齐的四个关键目标(§1.2)：鲁棒性、可解释性、可控性和道德性(**RICE**)。同时，本文将当前关于对齐的研究和实践分解为四个关键领域(§1.3)：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原则）和四个领域共同构成了**对齐循环**。

本文介绍了人工智能对齐的理论、技术与评估，并讨论了可能的未来研究方向。

## 1.1 对齐的动机

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(Smolensky, 1987; Goel, 2022)扩展到基于自监督学习的系统(Mnih et al., 2015; OpenAI, 2023a)。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(Silver et al., 2017; Kaufmann et al., 2023)以及复杂且高风险的真实世界应用场景(Ruff and Pappu, 2021; Degrave et al., 2022)中，基于深度学习的人工智能系统均取得了显著成就。除此之外，大语言模型在多步推理(Wei et al., 2022; Wang et al., 2023)和跨任务泛化(Brown et al., 2020; Askell et al., 2021)方面的能力也不断增强。

然而，随着人工智能系统能力的增强，其带来的风险也随之增加。近年来，两个不同的群体敲响了警钟。第一个群体关注当前的技术伦理风险，早在几年前面部识别系统已经出现了对某类人群或性别识别特别不准确的现象，而当下的大语言模型也展现出一些关乎伦理的不良行为（例如，不真实的回答(Bang et al., 2023)和对性别和移民身份等明显的偏见(Perez et al., 2023)），这些行为可能进一步加剧社会现有的不平等现象。而第二类群体关注于未来的风险——人工智能对齐与人工智能安全之间的界限正日益模糊。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，即人工智能系统可以在所有相关方面达到或超过人类智能(Bubeck et al., 2023)。然而，这种潮流在带来技术进步(Korinek et al., 2021)和效率提升(Furman and Seamans, 2019)的同时，也可能带来严重的风险(CAIS, 2023)，甚至是全球范围内的严重危害(Hendrycks et al., 2023; GOV.UK, 2023)和存在性风险（即威胁到人类长期生存的潜在风险）(Ord, 2020)。在CAIS (2023) 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。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 [0]。

具体来说，当前人工智能系统已经表现出的与人类意图相悖的不良或有害行为，被称为人工智能系统的对齐失败，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(Hendrycks et al., 2021c)和潜在的生存风险(Hendrycks et al., 2023)。本文总结了几类较为显著的对齐失败行为和先进人工智能系统可能



[0] https://www.gov.uk/government/topical-events/ai-safety-summit-2023。

具备的危险能力，从而为未来的对齐评估提供研究方向：

- **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(Wilke et al., 2001; Lehman et al., 2020; Hendrycks et al., 2021c)。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(Lehman et al., 2020)。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(Ouyang et al., 2022)。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(Lin et al., 2022; Chen et al., 2021)。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(Hendrycks et al., 2021c)。

- **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(Shevlane et al., 2023)。这些系统可以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(OpenAI, 2023a)。在大语言模型、推荐系统(系统影响用户的偏好)(Adomavicius et al., 2022)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)(Amodei et al., 2017)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。Sciadvadh (Spitale et al., 2023)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。

- **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(Kenward and Sinclair, 2021)。针对这些不足的研究工作涵盖了机器伦理领域(Tolmeijer et al., 2020)，并深入探讨了关键问题，例如，人工智能应该与谁保持一致？(Santurkar et al., 2023)。

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，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(OpenAI, 2023b)。

### 1.2 对齐目标: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用来衡量对齐的标准。首先，我们必须明确本文讨论的对齐目标是什么。Leike (2018)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(OpenAI, 2023b)："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 (2021)所指出的，这是一个具有挑战性的任务。例如，人类可以代表从个体到人类群体的各种实体。Gabriel (2020)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**）。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(Dietterich, 2017)或对抗压力(Rudner and Toner, 2021b)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(Taleb, 2007)和长尾风险(Hendrycks et al., 2021c)，以及各种对抗压

力(Song et al., 2018; Chakraborty et al., 2021)。例如，一个未完全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(Zou et al., 2023)。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(Steinhardt and Toner, 2020)，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(Kirilenko et al., 2017; OecdAI, 2021; Rudner and Toner, 2021b)。一个对齐的系统应在其生命周期内始终保持鲁棒性(Russell, 2019)。

- **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(Räuker et al., 2023)。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(Turpin et al., 2023; Park et al., 2023b; Jacob Steinhardt, 2023)或欺骗性对齐(Carranza et al., 2023)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(Carroll et al., 2023b)。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(Elhage et al., 2021; Meng et al., 2022)。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(Holzinger et al., 2017)，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(DeepMind, 2018; Rudner and Toner, 2021a)。

- **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(Soares et al., 2015; Hadfield-Menell et al., 2016a)。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(ARC Evals, 2023)。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(Shevlane et al., 2023; ARC Evals, 2023)。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(Bowman et al., 2022)，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)(Soares et al., 2015)。可控性原则引发的一个关键的问题是，随着人工智能系统愈发强大，人类如何更好的指导人工智能系统甚至是超级人工智能系统的训练和运行？这也被称为超对齐问题(Superalignment)。

- **道德性**指人工智能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(Kearns and Roth, 2019; Berk et al., 2021)，对个人造成伤害(Hendrycks et al., 2021b)，以及缺乏多样性或公平性(Collective Intelligence, 2023)。有大量的研究致力于为人工智能系统开发道德框架(Pankowska, 2020)。将道德性原则加入系统对于它们融入社会至关重要(Winfield et al., 2019)。

### 1.3 对齐循环

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[1](§1.3)。前向对齐过程关注于基于已有的对齐需求构建对齐的系统，而后向对齐关注于验证已对齐的系统在部署过程中的实际对齐度，并根据实际需要或社会道德的变化更新对齐需求。需要注意的是，前向对齐和后向对齐并不是割裂的，后向对齐也可能会出现在前向对齐的过程中，二者相辅相成贯穿于人工智能系统发展的整个生命周期，特别地，我们会更近一步讨论人类价值观在人工智能对齐中的地位(§5.1)，并进一步分析对齐范围外的AI安全问题(§5.2)。

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。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。

[1]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。

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。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对齐治理)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Koessler (2023) 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，包括在训练前后阶段和部署阶段。同样，对系统生命周期的每个阶段的监管措施也已经被广泛讨论(Schuett et al., 2023; Anderljung et al., 2023)。

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；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。接下来本文将对每个支柱进行介绍，阐明其如何协同构建一个全面的人工智能对齐框架。

## 2 从反馈中学习

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训练中的人工智能系统的行为？在大语言模型的应用中，一个典型的解决方案是利用基于人类反馈的强化学习(RLHF)(Christiano et al., 2017)，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。一些研究发现，经过RLHF训练的LLM (Ouyang et al., 2022) 比通过单纯使用监督学习方法训练的模型(Devlin et al., 2019; Brown et al., 2020) 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(Ouyang et al., 2022)。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(Bai et al., 2022a)。这些优点使得RLHF被广泛用来对齐LLM (Ziegler et al., 2019; OpenAI, 2023a; Touvron et al., 2023)。特别地，Dai (2023)提出基于人类反馈的安全强化学习（Safe RLHF），用于平衡大语言模型在对齐训练中帮助性和无害性目标之间的内在矛盾。通过将大模型中的安全性形式化为一种在训练中需要满足的约束，并用约束马尔科夫决策(CMDP)来形式化整个任务，Safe RLHF在满足输出符合人类价值中的安全性的前提下(约束)，尽可能的提高了帮助性(目标)。未来的努力可以集中在减少对人类标注的依赖(Sun et al., 2023) 和通过利用迭代RLHF方法(即将其与辩论框架集成(Irving et al., 2018))等，提高奖励模型的有效性。

本文根据Ouyang (2022)的研究将RLHF的流程总结为如下三个阶段：

- **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$\pi^{\mathrm{SFT}}$。这些任务包括对话处理、指令跟随和总结。

- **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$\pi^{SFT}(y \mid x)$ 的响应对$(y_1, y_2)$。然后，这些响应对被展示给人类标注者并标注得到偏好数据。偏好数据被用来构建奖励模型$r_\theta$。

- **通过强化学习进行策略优化**。最后一步是基于奖励模型$r_\theta$提供的奖励，使用RL方法对LLM的策略$\pi$进行优化。大语言模型从提示生成响应的过程被建模为一个bandit环境 (Ouyang et al., 2022)，在每个响应结束时从奖励模型$r_\theta$获得奖励。RL的主要目标是调整大语言模型的参数$\phi$，使得在训练提示数据集$D_{\mathrm{RL}}$上的期望奖励最大化：

$$\arg\max_{\pi_\phi} \; E_{x\sim D_{\mathrm{RL}}, y\sim\pi_\phi}\left[r_\theta\left(x, y\right)\right].$$

通常，会引入来自SFT模型$\pi^{\mathrm{SFT}}$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$\mathcal{D}_{\mathrm{pretrain}}$当中产生的梯度有助于保持模型性能，这在Ouyang (2022)中被称为PTX损失。因此，可以引入一个更全面的目标函数：

$$J(\phi) = E_{x\sim D_{\mathrm{RL}}, y\sim\pi_\phi}\Big[r_\theta(x,y) - \beta\log\left(\pi_\phi(y|x)/\pi^{\mathrm{SFT}}(y|x)\right)\Big] + \eta\, E_{(x,y)\sim D_{\mathrm{pretrain}}}\Big[\log\left(\pi_\phi(y|x)\right)\Big]$$

其中$\beta$和$\eta$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很受欢迎，但它面临着许多挑战(Tien et al., 2023)。其中一个突出挑战是可扩展监督，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(Bowman et al., 2022)。具体来说，可扩展监督的出现源自两个实际原因。首先是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(Christiano et al., 2017)。其次是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(Saunders et al., 2022)，例如，教人工智能系统总结书籍(Wu et al., 2021)，生成复杂的代码片段(Pearce et al., 2022)，和预测未来的天气变化(Bi et al., 2023)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(Leike et al., 2018)。

**从人工智能反馈中进行强化学习(RLAIF)** RLAIF是对RLHF的进一步扩展，通过使用大语言模型生成的反馈替代人类反馈，提高模型整体效用。该方法通过自我评估、修订、微调和评估人工智能反馈，构建一个可用于训练的奖励模型(Bai et al., 2022b)。与RLHF相比，RLAIF通过人工智能反馈实现无害性，降低训练成本，且在摘要任务上与人类反馈相媲美(Bowman et al., 2022)。

**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，将任务分解为子任务，形成树状结构，以便人类监督和评估(Wu et al., 2021)。同时，通过人工智能模型生成的批评有助于人类发现模型可能忽视的缺陷(Saunders et al., 2022)。这种混合方法展示了在复杂问题和多领域监督中，使用人工智能协助的可行性。

**递归奖励建模(RRM)** 奖励建模允许我们将系统目标的构建与行为评估分离(Ibarz et al., 2018)。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(Bai et al., 2022a; Touvron et al., 2023)。递归奖励建模(Recursive Reward Modleing, RRM)(Leike et al., 2018; Hubinger, 2020)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。基于评估答案比回答问题更容易的假设，奖励模型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。

**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(Irving et al., 2018)。作为零和辩论游戏，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势方。然而，通过观察游戏的过程和最终结果，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，这意味着说真话的辩论者更具有优势。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(Du et al., 2023; Claude, 2023)。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(Irving et al., 2018)。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., 2018)。

**合作逆强化学习(CIRL)** 许多对齐失败的模式，例如奖励破解(Victoria et al., 2020; Skalse et al., 2022)，欺骗(Park et al., 2023b)，和操纵(Carroll et al., 2023b)，都是AI系统对错误规范的目标进行“自信地”优化的结果(Pan et al., 2022)。在训练和部署过程中，指定的目标(如奖励函数)对AI系统来说起着一个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(Everitt et al., 2021) 或者被操纵(Carroll et al., 2023b)。合作逆强化学习(Hadfield-Menell et al., 2016b) 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* (Hadfield-Menell et al., 2016b)，协助博弈 (Fickinger et al., 2020)，和协助*POMDP* (Shah et al., 2020)。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$r$。

## 3 在分布偏移下学习

与固定输入分布的反馈学习过程形成对比，此部分更关注对齐过程中输入分布发生变化的情况，即分布偏移(Krueger et al., 2020; Hendrycks et al., 2021a)。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。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(Di Langosco et al., 2022)。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(Krueger et al., 2020; Perdomo et al., 2020)。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(Adomavicius et al., 2022)。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(Park et al., 2023b) 和操纵行为(Carroll et al., 2023b) 。应对分布偏移的方法包括算法干预 ，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 ，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）(Krueger et al., 2021) 和基于连通性的微调（CBFT）(Lubana et al., 2023) 等方法。后者包括对抗训练(Bai et al., 2021)，它用对抗性输入增强训练输入分布，以及合作训练(Dafoe et al., 2020)，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。

## 4 对齐保证

即使人工智能系统经过了前向对齐，我们在实际部署它之前还需要考察其对齐度的置信值(Anderljung et al., 2023)。这就是对齐保证：在人工智能系统实际训练和部署后对其实际对齐情况进行测量和评估。对齐保证的方法包括安全评估(Perez et al., 2023) 和更高级的方法，如红队测试(Perez et al., 2022)和可解释性技术(Olah et al., 2018)：

**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(Doshi-Velez and Kim, 2017)。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，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(Räuker et al., 2023)。

**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(Chakraborty et al., 2021; Perez et al., 2022; Liu et al., 2023b; Chen et al., 2023)。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(Yoo and Qi, 2021; Bai et al., 2021)。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。

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(Dafoe et al., 2021) 和伦理性(Tolmeijer et al., 2020) 的形式化理论，以及广泛的经验和实证方法。在这里我们介绍两种用于验证人类价值契合性的方法：

**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(Hendrycks et al., 2021d)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。Pan (2023)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 (2022)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 (2023a)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 (2023)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 (2023)使用社会价值取向的框架(Messick and McClintock, 1968; Van Lange et al., 1997)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。

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(Shevlane et al., 2023; Koessler and Schuett, 2023)。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，红队测试是对抗性训练的关键组成部分，可解释性可以帮助提供反馈(Burns et al., 2023)。

## 5 人工智能治理

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。在这里，本文主要讨论人工智能治理的多方利益相关者的方法，包括政府规定(Anderljung et al., 2023)，实验室自主治理(Schuett et al., 2023)，以及第三方组织(Shevlane et al., 2023; Koessler and Schuett, 2023)。总的来说，**政府机构**运用立法、司法和执法权力，制定人工智能发展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，是被监管的主体，同时又提出方法进行自我监督并影响治理体系架构。**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、人工智能系统应用提供审计，而且协助政府制定政策。特别地，本文建立在跨代际视角上从两个方面分析国际人工智能治理合作的重要性和可行性：管理全球灾难性人工智能风险和管理人工智能中的机遇，从而为国际人工智能治理的未来结构贡献创新思考。

**管理全球性的人工智能灾难风险** 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过快的开发和部署，带来潜在的负面全球影响(Tallberg et al., 2023)。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(Swaugerarchive, 2020)。国际治理合作的干预可以缓解这些全球性的挑战，例如国家之间的共识可以帮助避免潜在的人工智能军备竞赛，而而全行业的协议可以防止草率和不负责任的开发人工智能系统，从而保障人工智能长期和可持续的发展(Ho et al., 2023)。

**管理人工智能机遇** 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。人工智能发展中的地理差异将加剧经济和社会效益的不公平分配(Ho et al., 2023)。此外，技术领域决策权掌握在少数个体中可能会导致权力分配的不平等，从而形成代际性垄断(Noble et al., 2021)。通过人工智能的传播、教育和基础设施发展(Opp, 2023) 促成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。

人工智能治理领域另一个争论焦点是人工智能模型是否应开源(Seger et al., 2023)。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 (2010)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 (2023)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 (Touvron et al., 2023)、Falcon (Penedo et al., 2023)、Vicuna (Chiang et al., 2023)等。本节主要评估开源模型的安全优势和潜在威胁，以促进关于开源的可行性和具体方法的讨论。

**支持开源的观点** 支持开源的研究人员和政策制定者认为开源可以通过多种途径减轻模型中固有的安全风险：(1) 开源可以促进开发者和社区对模型的测试,进而快速识别和解决模型可能具有的问题，并增强对模型相关风险的认知，促进对这些潜在风险更多的关注和研究(Zellers, 2019)。(2) 开源被认为是促进权力和控制分散化的有效策略。一个例子是Stability公开Stable Diffusion的核心原因：他们将信任寄托于个人和社区，而不是由中央集中控制、未经选举的实体控制人工智能技术(Mostaque, 2022)。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(Howard, 2023)，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** 开源模型的批评者从多个角度评估了开源模型可能被滥用的风险，提出了反对意见：(1) 开源模型可能被微调为有害模型。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(Urbina et al., 2022)和生物武器(Sandbrink, 2023)。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。(Goldstein et al., 2023)。(2) 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(Seger et al., 2023)。(Zou et al., 2023)通过使用Vicuna-7B和13B(Chiang et al., 2023)实现了开发攻击后缀。一旦这些后缀在像ChatGPT(OpenAI, 2023a)，Bard(Google, 2023)和Claude(Anthropic, 2023)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。

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，但仍需要做好必要准备。例如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(Seger et al., 2023)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。

### 5.1 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(Gabriel and Ghazavi, 2021)。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 (2013) 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(Yu et al., 2018)。这一类工作最早涵盖了符号和统计学习系统(Anderson et al., 2005; Anderson and Anderson, 2007)，后来扩展到包括建立大型道德伦理数据集(Pan et al., 2023)和基于深度学习的方法(Jiang et al., 2021; Jin et al., 2022)。

- **公平性** 尽管存在争议(Verma and Rubin, 2018)，但公平性的定义相对于其他人类价值观来说比较清晰。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(Mehrabi et al., 2021)。关于人工智能公平性的研究非常广泛，这些方法涵盖从在训练前减少数据偏见出发(d'Alessandro et al., 2017; Bellamy et al., 2018)，最小化在训练过程中引入的不公平性(Berk et al., 2017)，以及处理训练阶段未成功学习到的不公平样例(Xu et al., 2018)。

- **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-林赛的价值观系统(Allport, 1955)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。Van (1997)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 (2014)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 (1973)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 (1992; 1994)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。

**合作型人工智能** 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(Kirilenko et al., 2017)。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(Dafoe et al., 2021)。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(Dafoe et al., 2020)。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 (2020)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。

**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 (2020)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(Bonabeau, 2002; De Marchi and Page, 2014)，基于深度学习的模拟(Storchan et al., 2021)，以及那些包含大语言模型的模拟(Park et al., 2023a)。这些模拟方法可以服务于各种下游应用，从影响评估(Osoba et al., 2020)到多智能体社会学习(Critch and Krueger, 2020)。在另一方面，社会选择(Sen, 1986; Arrow, 2012)领域以及相关的计算社会选择(Brandt et al., 2016)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(Leike, 2023; Collective Intelligence, 2023)。一部分研究对这个提议已经进行了早期阶段的实验(Yamagata et al., 2021; Köpf et al., 2023)。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(Kenward and Sinclair, 2021)，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。

### 5.2 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 (2023)，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(Cao and Baptista, 2023)。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(Soice et al., 2023)。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(Sandbrink, 2023)。像奥姆真理教(Danzig et al., 2012)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(Mirsky et al., 2023)，或者创建能在人类控制之外生存和传播的智能体(Bengio, 2023)。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众利益的政策将是避免这种威胁的关键。

**集体行动问题** 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(Grant and Weise, 2023)。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(Armstrong et al., 2013)。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(Hendrycks, 2023)。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。

## 6 结论

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(Hendrycks, 2022) – 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，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。

**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(Drexler, 2019)，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(Shevlane et al., 2023)、计算治理的基础设施(Shavit, 2023)以及关于人工智能系统的可验证声明的机制(Brundage et al., 2020)。

**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(Abbass, 2019)，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。首先，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(Critch and Krueger, 2020)。其次，将人类的道德和社会价值纳入对齐，这与机器伦理学和价值对齐领域密切相关(Gabriel, 2020)。第三，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(Bonabeau, 2002; Park et al., 2023a) 和博弈论(Critch and Krueger, 2020)。

**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(Popper, 1935)。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(Abbass, 2019)，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。

## 参考文献

Hussein A Abbass. 2019. Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, 11(2):159–171.

Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2022. Recommender systems, ground truth, and preference pollution. *AI Magazine*, 43(2):177–189.

Gordon Willard Allport. 1955. *Becoming: Basic considerations for a psychology of personality*, volume 20. Yale University Press.

Dario Amodei, Paul Christiano, and Alex Ray. 2017. Learning from human preferences. https://openai.com/research/learning-from-human-preferences.

Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.

Michael Anderson and Susan Leigh Anderson. 2007. The status of machine ethics: a report from the aaai symposium. *Minds and Machines*, 17:1–10.

Michael Anderson, Susan Anderson, and Chris Armen. 2005. Towards machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI 2005 fall symposium on machine ethics*, pages 1–7.

Anthropic. 2023. Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

ARC Evals. 2023. Update on ARC's recent eval efforts. https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. 2013. Racing to the precipice: a model of artificial intelligence development. Technical Report 2013-1, Future of Humanity Institute, Oxford University.

Kenneth J Arrow. 2012. *Social choice and individual values*, volume 12. Yale university press.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4312–4321. International Joint Conferences on Artificial Intelligence Organization, 8. Survey Track.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Yoshua Bengio. 2023. How rogue ais may arise.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, pages 1–6.

Eric Bonabeau. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, 99(suppl_3):7280–7287.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of computational social choice*. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Benjamin S Bucknall and Shiri Dori-Hacohen. 2022. Current and near-term ai as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 119–129.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

CAIS. 2023. Center for ai safety: Statement on ai risk. https://www.safe.ai/statement-on-ai-risk.

Ella Cao and Eduardo Baptista. 2023. 'deepfake' scam in china fans worries over ai-driven fraud. *Reuters*, 5.

Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnuv Tandon, and Sanmi Koyejo. 2023. Deceptive alignment monitoring.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023a. Characterizing Manipulation from AI Systems, March. arXiv:2303.09387 [cs].

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023b. Characterizing manipulation from ai systems. *arXiv preprint arXiv:2303.09387*.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2021. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. 2023. Content-based unrestricted adversarial attack. *arXiv preprint arXiv:2305.10665*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Brian Christian. 2020. *The alignment problem: Machine learning and human values*. WW Norton & Company.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Ruby RobertM GPT-4 Claude. 2023. New lw feature debates.

Collective Intelligence. 2023. Introducing the collective intelligence project.

Andrew Critch and David Krueger. 2020. Ai research considerations for human existential safety (arches). *arXiv preprint arXiv:2006.04948*.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134.

Richard Danzig, Zachary Hosford, Marc Sageman, Terrance Leighton, Lloyd Hough, Hidemi Yuki, and Rui Kotani. 2012. Aum shinrikyo: Insights into how terrorists develop biological and chemical weapons, second edition. Report, Center for a New American Security (CNAS), 12.

Scott De Marchi and Scott E Page. 2014. Agent-based models. *Annual Review of political science*, 17:1–20.

DeepMind. 2018. Building safe artificial intelligence: specification, robustness, and assurance.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186. Association for Computational Linguistics.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. 2022. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR.

Thomas G Dietterich. 2017. Steps toward robust artificial intelligence. *Ai Magazine*, 38(3):3–24.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

K Eric Drexler. 2019. Reframing superintelligence: Comprehensive ai services as general intelligence. *Future of Humanity Institute, University of Oxford*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467.

Arnaud Fickinger, Simon Zhuang, Dylan Hadfield-Menell, and Stuart Russell. 2020. Multi-principal assistance games. *arXiv preprint arXiv:2007.09540*.

Jason Furman and Robert Seamans. 2019. Ai and the economy. *Innovation policy and the economy*, 19(1):161–191.

Iason Gabriel and Vafa Ghazavi. 2021. The challenge of value alignment: From fairer algorithms to ai safety. *arXiv preprint arXiv:2101.06060*.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Ashok Goel. 2022. Looking back, looking ahead: Symbolic versus connectionist ai. *AI Magazine*, 42(4):83–85.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Google. 2023. Bard.

GOV.UK. 2023. Frontier ai: capabilities and risks – discussion paper.

Nico Grant and Karen Weise. 2023. In a.i. race, microsoft and google choose speed over caution. *New York Times*, 4. Updated on April 10, 2023.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016a. The off-switch game. *arXiv preprint arXiv:1611.08219*.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016b. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021b. Aligning ai with shared human values. *ICLR 2021*.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021c. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021d. What would jiminy cricket do? towards agents that behave morally. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Dan Hendrycks. 2022. Pragmatic ai safety.

Dan Hendrycks. 2023. Natural selection favors ais over humans.

Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, et al. 2023. International institutions for advanced ai. *arXiv preprint arXiv:2307.04699*.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Jeremy Howard. 2023. Ai safety and the age of dislightenment.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. Deceptive alignment.

Evan Hubinger. 2020. An overview of 11 proposals for building safe advanced ai. *arXiv preprint arXiv:2012.07532*.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Jacob Steinhardt. 2023. Emergent deception and emergent optimization.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35:28458–28473.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. 2023. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987.

Michael Kearns and Aaron Roth. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

B. Kenward and T. R. Sinclair. 2021. Machine morality, moral progress, and the looming environmental disaster. *Cognitive Computation and Systems*, 3:83–90.

Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998.

Leonie Koessler and Jonas Schuett. 2023. Risk assessment at agi companies: A review of popular risk assessment techniques from other safety-critical industries. *arXiv preprint arXiv:2307.08823*.

Mr Anton Korinek, Mr Martin Schindler, and Joseph Stiglitz. 2021. *Technological progress, artificial intelligence, and inclusive growth*. International Monetary Fund.

David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. 2020. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Jan Leike. 2023. A proposal for importing society's values.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023a. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023b. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Ekdeep Singh Lubana, Eric J Bigelow, Robert P Dick, David Krueger, and Hidenori Tanaka. 2023. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pages 22965–23004. PMLR.

Alasdair MacIntyre. 2013. *After virtue*. A&C Black.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

David M Messick and Charles G McClintock. 1968. Motivational bases of choice in experimental games. *Journal of experimental social psychology*, 4(1):1–25.

Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Gelei Deng, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, and Battista Biggio. 2023. The threat of offensive AI to organizations. *Comput. Secur.*, 124:103006.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Emad Mostaque. 2022. Democratizing ai, stable diffusion & generative models.

Ryan O Murphy and Kurt A Ackermann. 2014. Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1):13–41.

Safiya Umoja Noble, Beatrice Dias, Sara Cole Stratton, Aimee van Wynsberghe, Carlos Affonso Souza, Ilene Carpenter, Alvaro Martin Enriquez, and Emily Ratté. 2021. Ai regulation through an intergenerational lens.

OecdAI. 2021. Ai principles.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The Building Blocks of Interpretability. *Distill*, 3(3):e10, March.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Introducing superalignment. Accessed on July 5, 2023.

Robert Opp. 2023. Committing to bridging the digital divide in least developed countries.

Toby Ord. 2020. *The precipice: Existential risk and the future of humanity*. Hachette Books.

Osonde A Osoba, Benjamin Boudreaux, and Douglas Yeung. 2020. Steps towards value-aligned systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 332–336.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *ICML*.

Paulina Karolina Pankowska. 2020. Framework on ethical aspects of artificial intelligence, robotics and related technologies. *European Parliament*.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche, editors, *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*, pages 2:1–2:22. ACM.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023b. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*.

Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2022. Asleep at the keyboard? assessing the security of github copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 754–768. IEEE.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR.

Ethan Perez, Saffron Huang, H. Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3419–3448. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13387–13434. Association for Computational Linguistics.

Karl R. Popper. 1935. *The Logic of Scientific Discovery*. Routledge, London, England.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483. IEEE.

Milton Rokeach. 1973. *The nature of human values*. Free press.

Tim Rudner and Helen Toner. 2021a. Key concepts in ai safety: Interpretability in machine learning.

Tim Rudner and Helen Toner. 2021b. Key concepts in ai safety: Robustness and adversarial examples.

Kiersten M Ruff and Rohit V Pappu. 2021. Alphafold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208.

Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Jonas B Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 2023. Towards best practices in agi safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, et al. 2023. open-sourcing-highly-capable-foundation-models.

Amartya Sen. 1986. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181.

Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. 2020. Benefits of assistance over reward learning. *https://openreview.net/forum?id=DFIoGDZejIB*.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*.

Jacob N Shapiro and David A Siegel. 2010. Is this paper dangerous? balancing secrecy and openness in counterterrorism. *Security Studies*, 19(1):66–98.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Yonadav Shavit. 2023. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint arXiv:2303.11341*.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

Paul Smolensky. 1987. Connectionist ai, symbolic ai, and the brain. *Artificial Intelligence Review*, 1(2):95–109.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology?

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. 2018. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Jacob Steinhardt and Helen Toner. 2020. Why robustness is key to deploying ai.

Jacob Steinhardt. 2023. Emergent Deception and Emergent Optimization, February.

Victor Storchan, Svitlana Vyetrenko, and Tucker Balch. 2021. Learning who is in the market from time series: market participant discovery through adversarial calibration of multi-agent simulators. *arXiv preprint arXiv:2108.00664*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Shea Swaugerarchive. 2020. Software that monitors students during tests perpetuates inequality and violates their privacy.

Nassim Nicholas Taleb. 2007. *The black swan: The impact of the highly improbable*, volume 2. Random house.

Jonas Tallberg, Eva Erman, Markus Furendal, Johannes Geith, Mark Klamberg, and Magnus Lundgren. 2023. The global governance of artificial intelligence: Next steps for empirical and normative research. *arXiv preprint arXiv:2305.11528*.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D. Dragan, and Daniel S. Brown. 2023. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2021. Optimal policies tend to seek power. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23063–23074. Curran Associates, Inc.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191.

Paul AM Van Lange, Ellen De Bruin, Wilma Otten, and Jeffrey A Joireman. 1997. Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. *Journal of personality and social psychology*, 73(4):733.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

Krakovna Victoria, Uesato Jonathan, Mikulik Vladimir, Rahtz Matthew, Everitt Tom, Kumar Ramana, Kenton Zac, Leike Jan, and Legg Shane. 2020. Specification gaming: the flip side of ai ingenuity.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Claus O Wilke, Jia Lan Wang, Charles Ofria, Richard E Lenski, and Christoph Adami. 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

Alan F Winfield, Katina Michael, Jeremy Pitt, and Vanessa Evers. 2019. Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue]. *Proceedings of the IEEE*, 107(3):509–517.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

Taku Yamagata, Ryan McConville, and Raul Santos-Rodriguez. 2021. Reinforcement learning with feedback from multiple humans with diverse skills.

Jin Yong Yoo and Yanjun Qi. 2021. Towards improving adversarial training of NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5527–5533. ijcai.org.

Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. 2022. In situ bidirectional human-robot value alignment. *Science robotics*, 7(68):eabm4183.

Rowan Zellers. 2019. Why we released grover.

Zhaowei Zhang, Nian Liu, Siyuan Qi, Ceyao Zhang, Ziqi Rong, Yaodong Yang, and Shuguang Cui. 2023. Heterogeneous value evaluation for large language models. *arXiv preprint arXiv:2305.17147*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.