大语言模型对齐：概念、挑战、路线、评测及趋势

熊德意
天津大学智能与计算学部
天津市津南区海河教育园区雅观路135号，300350
dyxiong@tju.edu.cn

摘要

通用智能的“智能-目标”正交性及“工具性趋同”论点均要求通用智能的发展要智善结合。目前大语言模型在能力（智）方面发展迅速，但在更具挑战性的价值对齐（善）方面研究相对滞后。本综述将概述对齐的基本概念和必要性，简述其存在的社会和技术挑战，分析大语言模型对齐的主要技术路线和方法，探讨如何对大语言模型对齐进行评测，并对未来趋势进行展望。

关键词： 大语言模型；通用人工智能；AI对齐；大语言模型对齐

Large Language Model Alignment: Concepts, Challenges, Roadmaps, Evaluations and Trends

Deyi Xiong
College of Intelligence and Computing, Tianjin University
No.135 Yaguan Road, Haihe Education Park, Tianjin, 300350, China
dyxiong@tju.edu.cn

Abstract

The “intelligence-goal” orthogonality and “instrumental convergence” theses require a deep coupling between capability and alignment for the development of general intelligence. At present, large language models are developing rapidly in terms of capability (intelligence), but the research on a more challenging problem, value alignment (goodness), is relatively lagging behind. This article will introduce the basic concepts and necessity of alignment research, briefly describe its social and technical challenges, analyze the main technical routes and methods of large language model alignment and discuss how to evaluate large language model alignment and future trends.

Keywords: Large Language Model , Artifiicial General Intelligence , AI Alignment , LLM Alignment

1 引言

近年来，以OpenAI ChatGPT和GPT-4为代表的大语言模型（Large Language Model，LLM）发展迅速，重新燃起了人们对通用人工智能（Artificial General Intelligence，AGI）的热情和憧憬。虽然大语言模型是否是通向AGI之路仍存在争议，但在标度

律（Scaling Law）基础上不断扩展规模的大语言模型，其能力逐步呈现出一些AGI的特征(Bubeck et al., 2023)：在海量语言数据上训练的GPT模型，除了展现出强大的语言能力之外，在数学、推理、医疗、法律、编程等多个领域，正以惊人的速度逼近人类水平。

与大语言模型技术和能力不断突破相伴随的是，人们对大语言模型本身存在的社会伦理风险及其对人类生存构成的潜在威胁的普遍担忧。首先，在真实可见的社会伦理风险方面，研究发现(Weidinger et al., 2021)，一方面，大语言模型在输出文本中存在多种类型的信息危害，如将训练数据中存在的偏见、歧视、有毒内容输出到预测文本中，在生成文本中泄露训练数据中的隐私和敏感信息，生成低质量、虚假性、误导性信息；另一方面，大语言模型的使用也带来社会伦理风险，如大语言模型存在被滥用的可能性，用于人机交互类产品中时可能对使用者带来潜在影响，大范围使用大语言模型可能带来对环境、信息传播、就业等方面的影响。OpenAI团队研究发现(Eloundou et al., 2023)，美国80%的劳动力，其工作存在对大语言模型10%的风险敞口（即会受到大语言模型影响），19%的就业人员，其50%的工作任务会受到大语言模型影响，且收入越高，受大语言模型影响越大。

其次，在更远期的潜在影响方面，很多人担心未对齐的AGI可能带来人类存亡风险（Existential Risk，X-Risk），即超过人类知识和智能水平的AI代理（Agent）会形成自己的目标（Goal），且该目标与人类赋予的目标不一致，为了实现自己的目标，AI代理将会获取更多的资源，实现自我保持、自我提升，这种发展将会持续扩展至对整个人类进行权利剥夺（Disenpower），从而不可避免地导致人类生存灾难(Carlsmith, 2022)。基于以上担忧，美国波斯顿未来生命研究所（由Skype联合创始人和麻省理工学院教授共同创立）于2023年3月22日发起暂停巨型AI实验的公开倡议信[0]，要求所有AI实验室暂停训练比GPT-4更强大的AI模型至少6个月，截至2023年7月6日，网上签名人数已超过三万三千人，签名人员包括图灵奖获得者Yoshua Bengio、特斯拉CEO Elon Musk等人。公开信中提到阿西洛马人工智能原则（Asilomar AI Principles）："先进的人工智能可能代表地球生命史上的一次深刻变化，应该以相应的关心和资源进行规划和管理"。图灵奖获得者，也是此次大语言模型底层核心技术的发明者之一，Geoffrey Hinton也表达了对未来AGI的担忧，并签名了由AI安全中心于2023年5月30日发起的AI安全声明[1]。该声明仅包含一句话（22个单词），强调AI安全应该具有和防止大流行病、核战争一样的优先级别。

对AGI是否导致X-Risk，目前还存在争议。与Geoffrey Hinton、Yoshua Bengio同年获得图灵奖的Yann LeCun认为目前的大语言模型技术并不能实现AGI，也不会导致X-Risk。2023年6月22日，著名辩论会"芒克辩论会"（Munk Debates）邀请了图灵奖获得者Yoshua Bengio和MIT教授Max Tegmark作为正方，图灵奖获得者Yann LeCun和圣塔菲研究所教授Melanie Mitchell作为反方，就AI研究和发展是否构成X-Risk威胁问题进行了辩论[2]，辩论前正反方观众投票为67% vs 33%（即67%的观众认为AI研究和发展构成X-Risk威胁，33%认为不会），辩论后，正反方得票率为63% vs 37%。虽然反方辩论后获得了4个点的支持，但大部分观众听完辩论后仍然认为AI研究和发展构成X-Risk威胁。

需要注意的是，以上公开倡议、广泛的讨论和辩论，并不是宣扬AI宿命论，而是强调在致力于提升AI能力研究的同时，也要高度重视AI安全的研究。强调AI发展的长远风险，也并不是要掩盖或者回避大语言模型带来的真实社会伦理风险。AI能力研究势不可挡，AI安全研究势在必行！

上述社会伦理风险与人类存亡风险，都与AI安全技术——人工智能对齐（AI Alignment）——密切相关。AI对齐是AI的一个新兴领域，真正发展时间不过10年左右，但随着大语言模型的飞速发展，该领域越来越受到关注和重视。本文将介绍AI对齐的基本概念和相关背景（第2节），阐述对齐存在的巨大挑战（第3节）,探讨实现大语言模型对齐的主要技术路线（第4节）,介绍如何评测对齐模型（第5节），并对未来AI对齐研究的趋势进行展望（第6节）。

2 什么是AI/LLM对齐

人工智能对齐的概念萌芽最早可以追溯至控制论之父Norbert Wiener，他在1960年发表于

[0]https://futureoflife.org/open-letter/pause-giant-ai-experiments/
[1]https://www.safe.ai/statement-on-ai-risk
[2]https://munkdebates.com/debates/artificial-intelligence

《Science》的一篇论文(Wiener, 1960)中提到：

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.

在这段话中，Norbert Wiener明确指出，“mechanical agency”的目标应该与我们期待它实现的目标保持一致，即机器目标要与人类目标对齐。

2014年，《人工智能：一种现代方法》作者之一Stuart Russell教授在一次访谈中[3]指出：

> The right response seems to be to change the goals of the field itself; instead of pure intelligence, we need to build intelligence that is provably aligned with human values. For practical reasons, we will need to solve the value alignment problem even for relatively unintelligent AI systems that operate in the human environment. There is cause for optimism, if we understand that this issue is an intrinsic part of AI, much as containment is an intrinsic part of modern nuclear fusion research. The world need not be headed for grief.

Stuart Russell教授在这次访谈中首次提出了“价值对齐问题（Value Alignment Problem）”，即我们构建的不是纯粹的智能，而是与人类价值对齐的智能，并认为价值对齐问题是人工智能内在固有的一部分，价值对齐与人工智能的关系犹如安全壳之于核聚变反应堆。

虽然AI对齐概念在AI诞生之初就已萌芽，但由于人工智能在过去几十年发展曲折，其智能水平一直与人们期望的水平相差甚远，甚至很多时候被认为是人工智障，因此，对齐机器目标与人类目标/价值的紧迫性一直没有发展AI智能水平的紧迫性高。但近年来，大语言模型推动AI智能水平迅猛发展，并在越来越多的任务上，使其性能逼近甚至超过人类的水平，AI对齐的重要性和紧迫性也因此浮出水面，并受到越来越多的关注。从2012年开始，关于AI对齐的讨论和研究论文逐渐出现在相关论坛和arXiv上；2017年，AI对齐讨论的文章数量及论文数量出现爆发式增长，从原来的每年不足20篇猛增至400余篇(Kirchner et al., 2022)，这与大语言模型基础架构Transformer及GPT发明的时间基本吻合。

相比于AI其他研究领域，如自然语言处理，AI对齐还处于混沌状态，尚未形成科学研究范式(Kirchner et al., 2022)，除此之外，该领域的许多关键概念和术语也未形成共识。首先，在术语方面，“对齐”、“AI对齐”、“价值对齐”等名称经常在有关AI对齐的讨论文章和论文里交替使用，在中文相关讨论中，“人机对齐”也以AI对齐的替代形式出现。“对齐”在AI对齐领域及上下文环境中使用没有问题，但在更广泛的领域，容易与其他对齐概念产生混淆（如机器翻译中的双语或多语对齐）；“价值对齐”虽明确了对齐内容但未明确研究对象和领域；“人机对齐”虽然明确了人与机器之间对齐，但未明确研究领域、对齐内容及到底是人对齐机器还是机器对齐人。鉴于此，本文统一使用AI对齐和LLM对齐，LLM对齐可看作是AI对齐与自然语言处理、大语言模型的交叉领域。

其次，AI对齐的定义也没有形成共识。Paul Christiano将AI对齐定义为[4]：

> A is aligned with H if A is trying to do what H wants it to do.

上述定义过于宽泛，任何AI系统都可以认为是要完成人类想要它完成的任务，但实际上，上文已隐含提到AI对齐主要针对具有高智能（Highly Capable）的AI代理(Carroll, 2018)，这也意味着由未对齐AI导致的安全问题有别于一般的弱人工智能安全问题。也有研究人员从AI与人类关系的角度定义AI对齐。如Eliezer Yudkowsky将AI对齐定义为“创造友好的AI”、“连贯的外推意志（Coherent Extrapolated Volition）”。

除了从其本身的内涵及与人类关系角度定义AI对齐之外，还有一些工作试图以AI对齐要解决的具体问题来解释和具化AI对齐，Gordon Worley汇总了一些研究人员提出的AI对齐需要解决的问题[5]：

[3]http://edge.org/conversation/the-myth-of-ai#26015
[4]https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6
[5]https://laptrinhx.com/formally-stating-the-ai-alignment-problem-223323934/

- 避免副作用（Avoiding Negative Side Effects）：避免AI代理产生预期之外的行为。
- 避免奖励劫持/游戏（Avoiding Reward Hacking/Gaming）：避免AI代理利用奖励函数中的漏洞反复攫取奖励而忽视真正的目标。
- 可扩展的监管（Scalable Oversight）：将对AI代理的监管延伸至信息有限的情形或者人类难以直接判断的复杂任务上，比如当大语言模型能力在很多任务上超过人类水平时仍然可对其进行有效监管。
- 分布变化鲁棒性（Robustness to Distributional Shifts）：在新领域、新环境中，AI代理仍然能够按预期方式运行，尤其是在人类设计员未预期到的环境中，AI代理不会产生破坏性后果。
- 对抗鲁棒性（Robustness to Adversaries）：AI代理对对抗性攻击具有鲁棒性，其对齐不会被对抗攻击破坏，如在大语言模型的指令数据中掺入未对齐指令，其对齐效果不会受影响。
- 安全探索（Safe Exploration）：AI代理在不产生危险结果的前提下探索新的行为，如清洁机器人探索使用湿抹布，但不会用湿抹布擦拭电源插座。
- 安全中断（Safe Interruptibility）：AI 代理可随时被操作员安全中断，即AI代理不寻求避免被人类中断。
- 自我修改（Self-modification）：AI代理在可修改的环境中进行安全的自我修改，自我修改后仍然与人类价值对齐。
- 本体（Ontology）：AI代理建模世界并知晓其是世界的一部分。
- 理想决策理论和逻辑不确定性（Idealized Decision Theory and Logical Uncertainty）：AI代理能够作出理想化的决策，即使是在不确定环境下。
- Vingean反思（Vingean Reflection）：如何推测一个比人类更聪明的AI代理的行为，以确保其与人类价值对齐？如果能够推测这个更聪明的AI代理的行为，理论上，人类应该与该AI代理一样聪明甚至比其更聪明，这与之前的假设相悖。
- 可修正性（Corrigibility）：如果当人类需要修正AI代理（如修正建造AI代理时犯的错误）或者对其进行重编程时，AI代理应该允许被修正/重编程，而不是阻止，或者欺骗操作员其已被修正/重编程（实际AI代理仍然保持其原有目标，并未被修正/重编程）。
- 价值学习（Value Learning）：AI代理可以学习人类价值。

以上AI对齐问题和任务，有些已经进入了经验主义研究和实践阶段，如避免奖励劫持、可扩展的监管、鲁棒性等，有些则仍然在概念设想阶段，如本体、Vingean反思、可修正性等。

在本文中，我们从AI对齐的内涵角度对其进行定义：AI对齐是指AI代理的外部目标和内部目标均与人类价值一致，外部目标是AI代理设计者根据人类价值设计的训练目标，内部目标则是AI代理内部优化的目标。上述定义虽然对AI代理的目标进行了内部和外部界定和区分，但未对人类价值进行界定，因此仍然是一个不精确的定义。之所以将AI代理的目标分为外部和内部目标，是由AI对齐的技术本质决定的（详见第4节），而未对人类价值进行界定，则是因为AI对齐本身存在的社会和技术挑战导致难以从社会和技术角度明确定义人类价值（详见第3节）。

由于AI对齐涉及到接近或超过人类智能水平的AI代理与人类价值之间的对齐，因此AI对齐自然包括：

- 目前具备高智能的AI代理与人类价值的对齐，如现阶段大语言模型的人类价值对齐，
- 未来AGI与人类价值的对齐。

相比于AI对齐，人们对通用人工智能AGI更未形成共识，其争议也更多。但即便如此，在讨论AI对齐时，我们认为有必要介绍一些AGI基本假设和论点，因为这些AGI相关背景有助于我们对AI对齐形成更好的理解和认识。

- 正交性论点（Orthogonality Thesis）：该论点认为AI代理的智能和它的目标处于两个正交的维度，即任意水平的智能可以与任意的目标相结合(Bostrom, 2012)，处于高智能水平的AI代理并不意味着其目标与人类价值对齐。

- 工具性目标趋同论点（Instrumental Convergence Thesis）：AI代理拥有一些趋同的工具性亚目标（Subgoal），实现这些工具性亚目标有助于AI代理实现其最终目标(Bostrom, 2012)。Nick Bostrom列出了一些潜在的工具性亚目标:
  - 自我保持（Self-preservation）：为了实现最终目标，AI代理可能将自我保持作为其工具性亚目标。
  - 自我增强（Self-improvement）：同样，为了实现最终目标，AI代理可能将自我增强作为其工具性亚目标，因为不断增强的推理能力、认知能力、知识水平，可以帮助AI代理更容易实现最终目标。
  - 资源获取（Resource Acquisition）：AI代理获取更多的资源，如电力等，以帮助其实现最终目标。

3 社会与技术挑战

从以上的介绍和讨论中，可以看出，AI对齐不仅仅是一个技术问题，它还具有很强的社会属性。首先，AI技术在社会经济中广泛应用，其发展也给社会带来了短期和长期影响，AI技术和人类社会形成了一个巨大的社会技术系统（Sociotechnical System），这个社会技术系统必然要求AI与人类社会进行对齐，因为只有如此，这个系统才能和谐发展和共存。其次，AI代理要对齐的人类价值是一个典型的社会概念。以上社会属性，自然给AI对齐带来社会层面的挑战:

- AI对齐的人类价值如何定义？是全人类社会的价值还是某些国家和文化的价值?
- 如何将人类价值的文化差异性纳入AI对齐框架中，使对齐的大语言模型支持不同的文化价值?
- 如何确保在差异性背景下价值对齐的公平性，以保证少数群体的价值不被AI模型忽视?
- 如何在AI对齐框架中处理价值冲突问题?
- 如何在社会技术系统中避免大语言模型的价值对齐不被少数利益群体劫持?
- 如何评估AI对齐对社会的影响?

以上仅列出部分社会挑战，这些挑战对AI对齐的内在实现和外在部署均会形成影响，这就要求大语言模型对齐不仅要从技术角度考虑如何实现，同时也要从大语言模型实际应用的社会环境角度进行综合评估和规划。

除了以上AI对齐的社会属性给AI对齐研究带来社会挑战之外，AI能力研究也会给AI对齐研究带来重大挑战。AI对齐与AI能力，两者关系如同硬币的正面和反面，一方面，AI对齐研究不仅可以为AI能力的研究提供深刻洞见，而且也为AI能力研究提供安全护栏，使其在风险可控的条件下有序发展；另一方面，AI能力研究也可以为AI对齐研究提供技术手段和支持，但失衡的AI能力研究、过度的AI能力研究，反而加速了AI风险的累积，尤其是在AI研究和应用极度竞争的情形下，AI研发机构和利益方可能更关注短期利益，把更多资源投入AI能力研发，以获取更快的利益回报和竞争优势。

除了社会挑战之外，AI对齐研究面临巨大的技术挑战，其技术难度不亚于甚至远远超过AI能力研发的技术难度。AI对齐至少面临以下几方面的技术挑战:

- 如何设定（Specify）AI代理需要对齐的人类价值：一方面，人类价值具有多元化、结构复杂、文化相关、不断演变等特点，造成其难以被明确定义；另一方面，人类价值是一个定性的概念，而AI代理常常需要一个可度量的定量优化目标。

- 如何优化AI代理的目标：由于人类价值难以设定，AI对齐通常优化人类价值的替代物（Proxy），如从人类偏好中学习到的奖励函数。但是优化替代物是否就是优化AI模型使其逼近人类价值，这本身就是一个问题。另一方面，在优化过程中，如何避免模型对奖励进行劫持或游戏也是具有挑战性的技术难题。

- 如何规避负面效果：如何防止AI对齐损害AI代理的能力（对齐税）？如何避免AI代理产生预期之外的行为?

- 如何应对未见情况：如何应对分布变化、对抗性攻击?

- 如何将AI对齐扩展至更高级系统：AI代理能力越强，使其与人类价值对齐的难度也越大，如何使AI对齐沿AI能力增长曲线进行有效扩展，极具挑战性。

以上技术挑战是目前在对齐大语言模型等高智能的AI代理中真实存在的技术难题，如果未来AGI预期实现的话，第2节提到的安全探索、安全中断、自我修改、可修正性、价值学习等均是AI对齐要解决的重要技术挑战。

4 技术路线

针对AI对齐，一些学者和研究机构陆续提出了对齐方法和提案（Proposal）。Geoffrey Irving等人提出通过“辩论（Debate）”的方式实现AI对齐(Irving et al., 2018)，即在零和辩论游戏的基础上，通过自我对局的方式训练AI代理。对给定的问题或建议的行为，两个AI代理轮流做简短陈述，然后由人类判断哪个代理提供了最真实、最有用的信息。提出该方案的主要动机是，对于复杂的任务，人类通常难以直接判断AI代理的行为是否安全和有效，辩论方式使人类可以在多步对局的环境中只需要简单的推理规则就可以判断真假。该方案于2018年提出，当时语言模型还不能有效捕获人类意图和指令并生成相应的回复，让AI代理使用自然语言进行辩论，在当时条件下难以实现。虽然目前的大语言模型已经具备使用自然语言交互的能力，但辩论方案是否对大语言模型的对齐有效，仍然有待实验验证。

同在2018年，Paul Christiano（前OpenAI语言模型对齐团队负责人、对齐研究中心ARC创始人）等人提出了“迭代蒸馏和扩增（Iterated Distillation and Amplification，IDA）”方案（又称为迭代扩增）(Christiano et al., 2018)，该方案同样是针对人类难以在复杂任务上评测AI代理的问题提出来的，即实现可扩展的监管。初始时，人类将知识蒸馏给一个比自己弱的AI代理，这个过程称为蒸馏（Distillation），接着人类可以使用蒸馏的AI代理辅助自己，得到扩增版的新代理，这个过程称为扩增（Amplification）。以上蒸馏和扩增不断迭代进行，在这个过程中，AI代理的能力在不断增强，同时因为人类提供了对齐信号，其对齐能力也在不断增强。

仍然是在2018年，Jan Leike（现OpenAI对齐团队负责人）等人提出了“递归奖励建模（Recursive Reward Modeling，RRW）”的对齐方案(Leike et al., 2018)，该方案类似于前两个方案，均是针对可扩展的监管问题。RRW方案可看作是用奖励建模取代蒸馏模仿学习的IDA，具体而言，奖励建模分为两步：（1）从用户提供的对齐信号中学习奖励模型；（2）用该奖励模型以强化学习方式优化AI代理。在扩增步中，用户与强化学习优化的AI代理交互形成一个增强版的AI代理，用于下一步的迭代。可以看出，ChatGPT所用的“人类反馈强化学习（Reinforcement Learning from Human Feedback，RLHF）”方法(Ouyang et al., 2022)实际就是一个未递归的RRW，即只进行了一步对齐学习，未进行迭代扩增。最近OpenAI调集资源成立“超级对齐（Superalignment）”团队，并提出了超级对齐方案[6]，该方案可看作是RLHF的迭代扩增版（结合了可解释性及对抗测试）。

以上仅仅介绍了三个不同的AI对齐方案，这只是AI对齐提案的一小部分而已，其他提案还包括“逆奖励设计（Inverse Reward Design）”(Hadfield-Menell et al., 2017)、“协同式逆强化学习（Cooperative Inverse Reinforcement Learning）”(Hadfield-Menell et al., 2016)等，限于篇

[6]https://openai.com/blog/introducing-superalignment

幅，不逐一介绍。辩论、迭代蒸馏和扩增及递归奖励建模三个对齐方案除了都是针对可扩展的监管之外，它们还有一个共同点，即均是进行外部对齐。AI对齐领域近年来形成的一个重要共识是，AI对齐按照由外到内，包含外部对齐和内部对齐两部分。

- 外部对齐（Outer Alignment）：人类价值或预期目标与AI模型训练目标之间的对齐，即AI代理的设计人员是否将人类价值/预期目标转化对应到AI代理的训练目标函数上。预训练语言模型（未进行对齐训练）的目标函数是预测下一个单词，这个目标函数显然和人类价值/目标未对齐，因此，只是经过预训练的大语言模型，与人类价值未进行外部对齐，其输出文本中存在具有社会伦理风险的内容、且通常难以捕获人类的意图。与此相反，人类反馈强化学习RLHF则进行了外部对齐，对齐的实际目标是人类价值、意图等，由于人类价值/意图很难量化定义（见第3节）,RLHF采用了人类偏好作为人类价值/意图的替代物（Proxy）。为了实现外部对齐，RLHF采用了模仿学习和强化学习。在模仿学习步骤中（即有监督的微调（Supervised Fine-tuning，SFT）），RLHF提供与人类价值/意图对齐的样本作为示范供预训练的大语言模型进行模仿学习；在强化学习步骤中，RLHF首先根据人类偏好训练一个奖励函数，然后用该奖励函数通过强化学习进一步优化经过模仿学习的大语言模型，使其进一步与人类价值/意图对齐。

- 内部对齐（Inner Alignment）：AI代理真实优化的目标与人类赋予它的训练目标之间的对齐，即在AI代理训练过程中，其内部优化的目标与模型训练的目标函数一致。Evan Hubinger等人首次提出内部对齐概念(Hubinger et al., 2019)。当一个被训练的模型（如神经网络）本身是一个优化器（即其本身按照某种目标函数在可能的空间中进行搜索）时，我们称之为内优化器（Mesa-optimizer），而训练这个模型的学习算法则称为基优化器（Base-optimizer）。基优化器的目标函数称为基目标（Base-objective），内优化器的目标函数则称为内目标（Mesa-objective），内部对齐便是当一个被训练的模型本身是一个优化器时其基目标与内目标之间的对齐。基目标通常是模型设计人员定义的目标函数，而内目标则通常是是内优化器内部为完成给定任务演化出来的工具性目标，也就是说，基目标是模型设计人员定义和赋予的，内目标并不是设计人员指定的。Evan Hubinger等人用生物进化类比说明基目标与内目标的不对齐情况，生物进化的基目标是生物体与环境的包容性遗传适应性（Inclusive Genetic Fitness），适应性强的生物体被进化基目标选择和保留。作为生物进化出的特殊生物体的人类，其本身也是一个优化器。但是人类大脑的内目标与生物进化的基目标可能并不一致，比如按照生物进化的基目标，人类应该尽可能多地繁衍后代，但是很多人选择不生孩子。

  Evan Hubinger等人进一步指出，内优化器可能产生欺骗性对齐（Deceptive Alignment）。具体而言，内优化器演化出对基目标建模的能力，并知晓内优化器如果在基目标上表现差就会被基优化器修改而不能完成其自身优化的目标，因此，内优化器将会激励自己不被修改：在训练阶段表现出是在优化基目标函数，但一旦训练完成被部署时，由于被修改的风险已解除，内优化器就会寻求自己的内目标。

上文提到RLHF是一种外部对齐方法，该方法虽然对齐效果显著，但是该方法本身因为其潜在的缺陷遭到了批评。对该方法的批评意见主要来自于两方面[7]：

- “强化学习”（RL）部分：批评者认为RLHF中强化学习会带来如下风险：
  - 目标导向性：强化学习可能使大语言模型追求奖励而具有目标性；
  - 工具趋同：强化学习可能使大语言模型形成工具性亚目标，已有工作(Perez et al., 2022)发现，RLHF增强了大语言模型追求自我保持的欲望（即不被关闭）；
  - 激励欺骗性：经过RLHF训练的大语言模型，参数规模越大，产生的回复与用户偏好的回复一致的比例越高(Perez et al., 2022)，即迎合用户的偏好。

- “人类反馈”（HF）部分：批评者认为RLHF中的人类反馈存在以下缺陷：
  - 人类反馈数据通常以人工方式收集，因此需要较高的成本，同时也存在引入错误或被操纵的可能性；

[7]https://www.lesswrong.com/posts/d6DvuCKH5bSoT62DB/compendium-of-problems-with-rlhf

- RLHF最终使用的是从人类反馈中学习到的奖励函数，是人类反馈的替代物，并非人类反馈本身；
- 如前文所述，人类反馈未进行扩增，因此不能适应可扩展的监管（单纯的人类反馈无法胜任复杂任务）。

除了上面提到的外部和内部对齐，AI对齐还有一个重要问题要需要研究和解决，即可解释性（Interpretability）。可解释性的研究通常包括两部分(Critch and Krueger, 2020)：透明性（Transparency）和可说明性（Explainability），前者揭示AI代理、大语言模型的内部运作机理，后者说明AI代理决策过程中的事实或反事实的依赖关系，即模型为什么产生这样的预测结果或行为。相比而言，透明性更专注于模型内部，可说明性则通常是事后行为(Lipton, 2016)。

可解释性研究，显然有助于AI代理研发人员深入了解其研发的模型。对于AI对齐，尤其是内部对齐，可解释性不仅可以提供监测和洞见，而且其本身的评测指标也可以作为AI对齐优化的目标函数(Critch and Krueger, 2020)，以激励AI模型保持目标透明性（Goal Transparency）(Amodei et al., 2016)（避免欺骗性对齐）。

近年来，机械可解释性（Mechanistic Interpretability）成为AI对齐可解释性研究的一个重要方向，该可解释性研究旨在以逆向工程方式剖析AI模型，尤其是黑盒子的神经网络模型。由于大语言模型参数规模庞大，内在神经网络结构复杂，对其进行逆向工程，难度非常高，因此，现阶段的机械可解释性通常是在简化的玩具模型上开展的。即便如此，机械可解释性近几年仍然陆续揭示了神经回路[8]、归纳头（Induction Head，可用于解释语境学习（In-Context Learning））[9]等神经网络内部机理。

5 评测

上文提到大语言模型对人类社会存在近期和远期风险：社会伦理风险及通用人工智能安全风险，而AI对齐技术正是要避免这些风险，因此对AI对齐的评测也主要从这两方面展开：社会伦理对齐评测和通用智能安全评测。

5.1 社会伦理对齐评测

大语言模型生成的内容广泛出现于社交媒体、新闻媒体和在线平台，对人们的意见、观点和决策产生影响。如果大模型的价值观与人类价值观不相符，其生成的内容可能传播有害、误导性或偏见的信息，从而导致社会隔阂、歧视或其他负面后果。恶意行为者还可以利用大语言模型制造虚假信息、进行网络欺诈或发动攻击。

因此，确保大语言模型输出及行为与人类伦理价值对齐至关重要，这是在真实应用场景中部署和应用的大语言模型必须具备的能力。为了评估此能力，现有的研究考虑了多个与人类价值对齐的标准，如真实性、偏见性和伦理性(Askell et al., 2021; Bai et al., 2022)。对于真实性，可以利用对抗性问答任务（例如TruthfulQA (Lin et al., 2021)）检测。偏见性主要指性别、种族和年龄等方面的歧视，许多研究针对偏见的某一方面或多个方面建立了评估数据集。

尽管很多数据集提供了自动评估方法，但在伦理价值评测中，人工评估仍然是一种有效的评测方法，因为许多偏见暗含在语言之中，仅凭现有的自动评估指标很难判定。

5.2 通用智能安全评测

前文提到，通用人工智能通常具有自我保持、自我增强、自主复制、资源获取等方面的特征和趋势，因此，对以大语言模型为代表的AI代理，需要进行通用智能安全方面的科学和综合评测，以及时发现和防范潜在的风险。未经过通用智能安全评测或评测不达标的通用智能体，为避免产生不可控的安全风险，应该由相关部门监督该智能体研发机构对其进行AI对齐修复，直至安全评测达标方可发放模型部署和应用许可证。

目前的大语言模型虽然能力还未达到AGI水平，但是相关的通用智能安全评测已经开始。OpenAI委托对齐研究中心ARC对其研制的GPT-4进行“自主复制”方面的对齐评测，ARC将自主复制（Autonomous Replication）定义为[10]：部署在云端的AI代理获取相关

[8] https://distill.pub/2020/circuits/
[9] https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html
[10] https://evals.alignment.org/

资源（计算、资金等）并利用这些资源进行自我拷贝的能力。ARC对此设计了相应的评测实验，虽然未发现GPT-4具备自主复制方面的能力，但他们警告大语言模型能力在持续提升中。

Anthropic、Surge AI、及Machine Intelligence Research Institute三家单位联合对大语言模型的行为进行了综合评测(Perez et al., 2022)，评测用例由大语言模型本身生成，该评测不仅发现了大语言模型在能力上存在逆扩展（Inverse Scaling ）现象（即模型规模增大，某些能力反而降低），而且发现RLHF使大语言模型具有目标保持、资源获取的趋势。

6 未来趋势

未来几年，AI和LLM对齐将在多个方面取得重要进展和突破：

- 可扩展的监管：现阶段的AI/LLM对齐研究虽然在大规模系统上取得了初步成效，但仍然停留在AI对齐的初步阶段，在人类难以企及的复杂任务上，虽然已经提出了相关的对齐方案，但这些方案仍未进行大规模经验主义验证。未来将基于人机结合、多智能体结合方式进行迭代扩增，实现可扩展的监管的突破。

- 欺骗性对齐的实验验证：现阶段的大语言模型虽然能力非常强，但仍没有达到欺骗性临界点，未来大语言模型能力进一步发展，大型AI研发机构和企业将会开展大规模实验，寻找欺骗性对齐存在的蛛丝马迹，以便在其真正出现的时候做好应对准备。

- 机械可解释性：未来研究将会借鉴神经科学、心理学、脑科学相关理论和方法，对真实复杂的大语言模型进行大规模逆向工程，揭示其内部工作机理，如功能性/任务性神经回路等。

- LLM对齐对大语言模型能力研究的反馈：大语言模型对齐的研究将会为大语言模型能力的研究提供正反馈，帮助解锁大语言模型更多能力，未来大语言模型能力的提升可能不是来自于模型、数据规模的单纯扩增，而是来自于对齐算法及其发现。

- 对齐评测：对齐研究离不开对齐评测，未来对齐评测将呈现从社会伦理评测向通用人工智能安全评测发展的趋势。

7 结论

本文对AI/LLM对齐研究进行了简要介绍，包括相关概念、挑战、技术路线、评测及未来发展趋势。可以看出，为了避免大语言模型和通用人工智能目前的社会伦理风险及未来的人类生存风险，其发展必须坚持“智善一体化”的原则，大力开展AI/LLM对齐研究。对齐研究和大模型能力研究并不矛盾，两者相辅相成，对齐研究为大语言模型能力的研究带来洞见，并进一步推进大语言模型能力的研究。另一方面，AI/LLM对齐研究非常具有挑战性，现阶段的研究还未形成统一的科学研究范式，整个领域还处于前科学阶段，需要更多的研究人员、研究机构投入其中，合力推动AI安全与能力的协同发展。

致谢

本研究受云南省科技厅重点研发计划专项（202203AA080004）、新疆维吾尔自治区自然科学基金重点项目（2022D01D43）、之江实验室开放课题（2022KH0AB01）资助。

参考文献

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *CoRR*, abs/1606.06565.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Nick Bostrom. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.*, 22(2):71–85, may.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *ArXiv*, abs/2303.12712.

Joseph Carlsmith. 2022. Is power-seeking AI an existential risk? *ArXiv*, abs/2206.13353.

Micah Carroll. 2018. Overview of current AI alignment approaches.

Paul F. Christiano, Buck Shlegeris, and Dario Amodei. 2018. Supervising strong learners by amplifying weak experts. *CoRR*, abs/1810.08575.

Andrew Critch and David Krueger. 2020. AI research considerations for human existential safety (ARCHES). *CoRR*, abs/2006.04948.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *ArXiv*, abs/2303.10130.

Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell. 2016. Cooperative inverse reinforcement learning. *CoRR*, abs/1606.03137.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca D. Dragan. 2017. Inverse reward design. *CoRR*, abs/1711.02827.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. Risks from learned optimization in advanced machine learning systems. *CoRR*, abs/1906.01820.

Geoffrey Irving, Paul Francis Christiano, and Dario Amodei. 2018. AI safety via debate. *ArXiv*, abs/1805.00899.

Jan H. Kirchner, Logan Smith, Jacques Thibodeau, Kyle McDonell, and Laria Reynolds. 2022. Understanding AI alignment research: A systematic analysis. *ArXiv*, abs/2206.02841.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *CoRR*, abs/2109.07958.

Zachary Chase Lipton. 2016. The mythos of model interpretability. *CoRR*, abs/1606.03490.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Ethan Perez, Sam Ringer, Kamilė Lukoiūtė, 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, Daisong Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, G R Khundadze, John Kernion, James McCauley Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua D. Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noem'i Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom B. Brown, T. J. Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds,

Jack Clark, Sam Bowman, Amanda Askell, Roger C. Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering language model behaviors with model-written evaluations. *ArXiv*, abs/2212.09251.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Norbert Wiener. 1960. Some moral and technical consequences of automation. *Science*, 131(3410):1355–1358.