arXiv:2301.03740v1 [cs.CY] 10 Jan 2023

人工智能对齐问题的多层级框架

Betty Li Hou[1, 2], **Brian Patrick Green**[1]
Markkula 应用伦理中心[2] 纽约大学
blh9134@nyu.edu, bpgreen@scu.edu

摘要

人工智能对齐关注的是如何以符合人类价值观的方式对人工智能系统进行编码。这一问题的规范性层面探讨的是：我们是否应在人工智能中编码某些道德价值观或原则；如果应当，具体应编码哪些。为此，我们提出一个框架，从四个层面考量这一问题：个人、组织、国家和全球。我们旨在说明，人工智能对齐由各层面的价值对齐问题构成，各层面的价值观会相互影响，影响也可能双向流动。我们概述了每个层面的关键问题和考量，并展示如何将这一框架应用于人工智能内容审核这一主题。

1 引言

人工智能在全球范围内以多种方式得到应用，从社交媒体算法和网络安全，到智能家居设备以及日益自主的自动驾驶车辆。这会对我们的政治、经济和社会结构造成直接和间接的负面影响风险。面对这一新兴技术领域，我们必须充分了解并努力应对相关风险，才能驾驭这一领域并明智地使用这项技术。这就是人工智能伦理领域，本文具体讨论的是人工智能安全。

1.1 人工智能对齐

人工智能对齐问题探讨的是如何对人工智能系统进行编码，使其符合人类道德价值观。当系统中存在多项需要优先考虑的价值观时，这个问题就会变得复杂。例如，对于在线内容审核这类具有道德意义的任务，我们可能希望系统兼具速度和准确性。如果这些价值观存在任何程度的冲突，就不可能同时将二者最大化。当人工智能系统在这样的规模上运行，以至于人类无法切实地评估其作出的每项决策、确认其是否负责任且合乎伦理时，人工智能对齐就显得更加重要。

对齐问题包含两个部分 [1]。第一部分是技术层面，重点在于如何将价值观和原则正式编码到人工智能中，使其能够可靠地按照应有的方式行事。如果处理不当，可能会产生意料之外的负面副作用和奖励黑客行为 [2]。第二部分是规范性问题，探讨我们是否应该将某些道德价值观或原则编码到人工智能中；如果应该，具体应编码哪些。为此，我们提出一个框架，从四个层面来考量这一问题。[1] 尽管其他近期的重要论文更侧重于问题解决方案的内容

[1]关于这个问题的另一种观点，请参阅 Max Tegmark 接受 The 80,000 Hours Podcast 采访时的“多尺度对齐”部分 [3]。Tegmark 的框架似乎尚未发表，因此我们无法确切了解他的框架与我们的框架在哪些方面相似或不同。

机器学习安全研讨会，第36届神经信息处理系统大会（NeurIPS 2022）。

（例如 Hendrycks 等人， 2021, [4] 以及 Hendrycks 等人， 2022, [5]），但这一框架关注的是这些以内容为重点的解决方案必须置身其中的社会背景，以及该背景包含多个层面，解决方案应在这些层面上加以整合并保持连贯。

2 分解人工智能对齐问题

人工智能对齐由多个不同层面的价值对齐问题构成，而不仅仅涉及技术本身、技术的构建方式以及设计方法。为了使人工智能真正与人类道德价值观相一致，所有层面也必须彼此对齐。以下是一种人工智能对齐方法：各层面的价值观相互影响，其影响既向下传递，也向上传递。在每个层面，都有需要回答的关键问题。

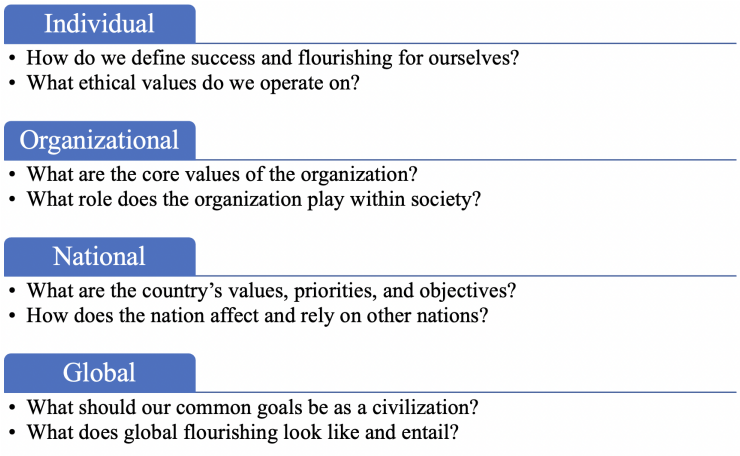


图 1：对齐的四个层面

个人与家庭 在个人层面，框架邀请个人和家庭思考有关价值观与幸福的问题。在日常行动中，我们正在塑造自己对个人幸福的定义——什么使生活充实并带来满足感。我们必须思考希望效仿哪些榜样和生活方式、如何定义自己的成功、愿意作出哪些牺牲，以及优先重视哪些道德价值观。

组织 组织层面指企业、州和地方政府、大学、教会、社会运动以及公民社会中的其他各种团体。在考虑这一层面的对齐时，我们必须确定组织依据哪些价值观运作、其产品和服务中灌输哪些价值观，以及该组织在社会中扮演什么角色。对于各类机构而言，重要的考量包括什么构成成功、使用哪些指标评估成功，以及它们如何参与更广泛的人工智能对齐运动。

国家层面 下一个层面是国家层面。每个国家都以隐含或明确的方式定义了价值观，这些价值观决定了该国在人工智能方面的目标。一个旨在确立全球强国地位的国家，可能会投入资源建设本国的人工智能产业，并规范人工智能的使用，以引导和推动用户的行为朝着特定观点发展。另一方面，一个旨在促进自由的国家，可能会采取去中心化的人工智能生产方式，赋予企业自由和隐私，同时促进企业间的竞争。或者，各国可能会尝试以一种不仅确保自身与道德价值观保持一致，也鼓励或要求其他国家这样做的方式来建设人工智能计划。

全球 从全球视角来看，人类必须思考我们希望拥有怎样的未来。最近提出的联合国可持续发展目标（可持续发展目标）提供了一个良好的起点，但这些目标仅仅是生存与繁荣所必需的前提条件，因此还远远不够 [6]。我们还需要进一步确定作为一个文明共同追求的目标；更具哲学意味地说，还要思考人类存在的目的，以及人工智能将如何融入其中。我们追求的是生存、养育

儿童、生活在社会中、寻求真理，等等吗？与此相关的，还有经济和政治结构的终极目标，以及强国和企业需要作出哪些让步，才能顾及贫困人口和地球的需求。

2.1 将各个层面结合起来

所有这些层面彼此相互作用。由于人工智能通常起源于组织层面，往往来自以营利为导向的企业，其首要动机通常只是赚钱。不过，置于其他层面的背景下，还应能看出更多目标：1）人工智能的发展应符合个人及家庭的需求；2）人工智能的发展应符合国家利益；3）人工智能的发展应有助于人类在全球层面的生存与繁荣。

不过，框架中的其他层次也会通过输入和输出彼此相互作用。例如，从输入的角度审视同一个组织层面，个人可以选择是否购买某些类型的技术，国家可以通过制定法律和法规来管控科技公司能够采取的行动；在全球层面，国际压力（例如联合国通过可持续发展目标等理念施加的压力）也会影响科技公司的行为。值得注意的是，这些层面之间还可能存在中间层面，例如欧盟，其层级高于国家层面、低于全球层面，并通过《通用数据保护条例》（GDPR） [7],对互联网和数据产生了重大影响，进而影响人工智能。

考察个人层面，我们已经看到它如何影响组织层面，又如何受到组织层面的影响。个人层面可以通过选举影响国家层面，也可以通过联合国等组织影响全球层面，不过这些影响尚未充分发展。同样，全球层面可以通过国际条约影响个人，而国家显然会通过法律及其他行为预期，对其公民施加重大控制。

最后，国家层面与全球层面会相互作用。国家会影响地球的全球状况，例如通过战争以及其他会产生全球影响的国家政策（如可能加剧或缓解气候变化的能源政策）。全球层面也能反过来施加影响，无论是通过联合国，还是通过其他关于国家行为的国际预期。

为了更实际地理解框架，我们来看看社交媒体内容审核这一问题。

3 内容审核实例

围绕互联网上某些类型内容所带来风险的全球辩论已然兴起。用户生成内容不像传统媒体内容那样受到同等程度的编辑审查，这使用户能够发布可能伤害他人的内容，尤其是可能伤害儿童或弱势群体的内容。这类内容包括但不限于宣扬恐怖主义的内容、儿童虐待材料、仇恨言论、色情内容，以及暴力或极端主义内容。然而，与此同时，限制此类内容的尝试在一些人看来，可能侵犯用户的言论自由以及听取某些表达的自由。组织和政府一直在努力权衡通过内容审核减轻这些潜在危害的可行性与伦理问题，同时又试图留住那些认为自身自由受到限制的用户。

人工智能辅助内容审核的速度和规模都远超人工审核。Google 的一份透明度报告显示，2022年4月至6月期间，YouTube 上被移除的视频中，超过 90% 是因自动标记而接受审核的 [8]。然而，这些方法会影响人们未来使用和看待在线内容分享的方式，因此，确保这些流程中采用的人工智能在多个层面上符合人类价值观至关重要。

3.1 使用该框架

组织层面 ⇒ **个人**第一个问题出现在组织层面，即企业与个人之间存在严重的目标错位。采用内容审核的企业会受到激励，力求最大化股东价值，因此将利润置于社会公益之上。例如，公司通常会根据“互动度”来制定算法：某个话题或帖子获得的点赞、评论和分享越多，就越可能出现在人们的新闻推送中。对于每个个人资料，公司也会

根据互动度追踪用户的行为和习惯，然后向他们推送他们想看的内容。这样一来，用户就会在网站上花费更多时间，为企业带来更多广告收入，从而提升股东价值。然而，这会导致回音室和两极分化，因为用户接触不到与自己不同的观点，最终影响的不仅是个人和家庭，还包括整个国家，甚至全球话语。组织与个人之间的目标错位已被证明具有危险性，缅甸袭击少数群体的事件便说明了其可能造成的后果 [9]。

国家 ⇒ **组织** ⇒ **个人**国家法规塑造了组织审核内容的方式，因为组织必须在这些法规的界限内构建人工智能。一个国家的内容审核立法通常体现了其大多数公民的文化价值观，而这些价值观往往与其领导层的文化价值观相似，但并非总是如此。尽管这些法规由个别立法者制定，可能体现许多个人公民的价值观，但也会影响组织和其他个人。例如，公共利益视角可能倾向于加强内容审核，以尽量减少社会危害，但代价是个人的言论自由。

于是，文化价值观与人工智能内容审核之间的对齐问题便随之而来。我们或许能够识别国家价值观与组织价值观之间的不一致，而这种不一致又会影响个人。例如，在美国，个人自由备受重视，因此内容审核法规很少，且仅要求公司审核诸如非法分享受版权保护的内容，以及分享儿童性虐待材料等犯罪活动。因而，尽管公司可能遵守所有相关政府法规，社会仍然受到有害影响，这表明美国政府的内容审核立法与社会需求并不对齐。缅甸等案例也表明，这些美国立法可能与全球需求并不对齐，因为其他国家同样面临这些问题，并承受其后果。

基于以上内容，人工智能对齐的首要目标似乎应当是实现国家层面与组织层面的对齐（假设该组织也与个人福祉保持一致）。然而，这还不够——我们还必须考虑这些国家价值观是否在全球层面保持一致，也就是说，它们是否有助于实现全球人类繁荣。宣扬分裂、颠覆政府、极端主义、阴谋论以及其他具有社会破坏性和反社会行为的内容，不仅会引发地方或国家层面的问题，而且如果这些群体获得广泛赋权，还会造成全球性问题。尽管如此，在我们这个多元化世界中，在国际层面管控内容审核根本行不通。不过，即使问题目前还无法解决，这一框架至少有助于诊断问题。

个人 ⇒ **组织** ⇒ **国家** ⇒ **全球** 影响会双向流动。负责内容审核的组织有时最重视个人用户反馈，而实力足够强大的组织能够影响国家利益；一个国家或一组国家也可能改变国际秩序，例如欧盟的 GDPR。

总而言之，内容审核是价值对齐如何在当今社会发挥作用的典型例子。要迅速或轻易地协调好所有四个层级或许并不可行，但借助这一框架，我们可以找出造成当前复杂不一致的一些原因——而且未来这类不一致还会更多——并探讨一些可能的解决方案雏形。

4 结论

如果要在人工智能对齐的规范层面取得良好进步，我们就必须考虑所有层级：个人、组织、国家和全球，并理解它们如何协同运作，而非只协调其中一个或少数几个部分。本文提出了一个用于考量这些问题的框架。在附录A中，我们分析了这项工作与人工智能存在风险的关系。

框架的多功能性意味着它可以应用于许多其他领域，包括但不限于自动驾驶车辆、人工智能辅助临床决策支持系统、监控和刑事司法工具。在这些争议激烈、没有明确答案的领域，通过从四个层面分析这些问题，我们能够看清其中相互作用的诸多要素，从而打造更符合伦理且对齐得更好的人工智能。

参考文献

[1] I ason Gabriel.《人工智能、价值观与对齐》。载于：*Minds & Machines*，30:411–437，2020 .

[2] Dario Amodei、Chris Olah、Jacob Steinhardt、Paul Christiano、John Schulman 和 Dan Mané。《人工智能安全中的具体问题》。载于：*arXiv* 预印本 *arXiv:1606.06565*，2016年。

[3] Robert Wiblin 和 Keiran Harris。《Max Tegmark 谈一项"要么付诸行动，要么闭嘴"的决议如何促使他投身人工智能和算法新闻选择》。The 80,000 Hours Podcast，2022年7月1日，1:13:13–1:51:01。网址：https://80000hours.org/podcast/episodes/max-tegmark-ai-and-algorithmic-news-selection/。

[4] Dan Hendrycks、Collin Burns、Steven Basart、Andrew Critch、Jerry Li、Dawn Song 和 Jacob Steinhardt。《让人工智能与共同人类价值观保持一致》。载于：*arXiv* 预印本 *arXiv:2008.02275*，2021年。

[5] Dan Hendrycks、Nicholas Carlini、John Schulman 和 Jacob Steinhardt。《机器学习安全中的未解问题》。载于：*arXiv* 预印本 *arXiv:2109.13916v5*，2022。

[6] 《17 项目标——可持续发展目标》。联合国。网址：https://sdgs.un.org/ goals。

[7] 《通用数据保护条例》。欧盟。网址：https://gdpr-info.eu。

[8] "YouTube 社区准则执行情况"。Google。网址：https://transparencyreport.google。com/youtube-policy/removals。

[9] Nathaniel Persily 和 Joshua A. Tucker。《社交媒体与民主：研究现状及改革前景》。剑桥大学出版社，2020。

A 存在风险分析

人工智能 X 风险分析模板，摘自：DanHendrycks 和 MantasMazeika。《人工智能研究的 X 风险分析》。载于：*arXiv* 预印本 *arXiv:2206.05862v7*，2022。

A.1 先进人工智能系统的长期影响

在本节中，我们将分析这项工作如何影响通往先进人工智能系统的进程，以及它如何引导这一进程朝着更安全的方向发展。

1. **概述**。 这项工作旨在如何降低先进人工智能系统带来的生存风险？**回答：** 先进人工智能系统必须置身于由不同国家、组织和个人组成的世界之中。为了避免全球灾难，我们必须审视各个层面。通过使这些层面对齐，并考虑它们之间的互动方式，我们可以更有把握地确保先进人工智能系统能够与这个世界相兼容，或者至少更清楚地认识到它们可能引发冲突的方式。这样一来，开发者就能更清楚地了解他们希望系统如何在世界中运行。但若不采取上述做法，未来的先进人工智能（包括通用人工智能）就不可能在任何意义上实现全球范围的对齐，从而为生存风险敞开大门。恶意、妄想或仅仅是无知的行为者利用人工智能内容审核时，已经可以看到这些错位所造成的影响，并导致了数不胜数的不良社会影响。

2. **直接影响**。 如果这项工作直接降低了生存风险，它直接影响的主要危害、脆弱性或失效模式是什么?

   回答： 这项工作通过阐明应对存在性风险所带来的问题、实现任何切实解决方案的先决条件，直接应对生存风险问题。任何未考虑个人、组织、国家和全球层面如何相互整合的解决方案都会存在缺陷，因此算不上切实可行的解决方案。这项工作阐明了我们应如何直接降低风险：首先，了解各层级如何实现对齐或发生目标错位；然后，考虑用于应对风险的人工智能解决方案可能如何存在于这一多层级结构之中。这种做法的一种失效模式是，这些知识也可能被用于制造目标错位，使人们得以优化多层级目标错位，而非实现对齐。一些国家和团体已经以这种方式利用人工智能审核系统，因此从负面角度看，这绝非什么新发现；而积极的一面似乎直到现在仍未被明确提出。希望通过将其明晰化，我们能够开始重新掌控已经逃逸的目标错位人工智能之精灵，并扭转其影响。

3. **间接影响**。 如果这项工作以间接或分散的方式降低了生存风险，那么它主要影响了哪些促成因素?

回答： 本文的主要作用是通过间接方式降低存在性风险：使那些提出其他人工智能对齐存在性风险解决方案的人，更好地理解解决方案必须存在于其中的社会的多层次性质，并进一步了解如何将其解决方案融入这一多层次社会。如果忽视本文的思想，那么人工智能对齐存在性风险解决方案可能会变得不那么有效和全面，因此无法像充分整合本文思想后那样，更好地解决人工智能对齐问题。仅仅在一个层面上实现对齐是不够的，例如政府层面，因为世界各国政府仍将处于目标错位状态。目前，我们可以在专制政体与民主政体、自由国家与压迫性国家之间看到这种情况。仅仅使国家能够让其民众与政府保持对齐的人工智能对齐存在性风险解决方案，只会增强压迫性的专制政体，使全球失调进一步恶化，而不是改善。其他层面也是如此。对齐必须从上到下一致，否则问题只会被挤压到其他层面，就像一个即将爆裂的水球一样。

4. **利害攸关之处是什么？** 未来可能出现什么情景，使这一研究方向能够避免大规模人员突然丧生？如果不适用，那么这一研究方向可能在哪种未来情景下带来巨大益处？

回答： 如前所述，这一框架有助于防止缅甸因人工智能辅助的社交媒体内容审核而发生的类似事件。如果创建这些系统的企业与个人、国家和全球需求保持一致，我们就能开始解决两极分化、回音室等问题，并杜绝人工智能像在缅甸那样被利用来针对少数群体。若没有这项工作，理论上的通用人工智能可能会按照一个国家的价值观构建，却与另一个国家的价值观直接冲突，从而造成重大伤害；或者不得不将其限制在较小的领域内，例如仅限于一个国家。然而，如果一个国家的组织对成功的定义达成一致，各国也对全球繁荣的定义达成共识，那么就有可能创造出能够带来更大社会效益的通用人工智能。

5. **结果脆弱性**。 这些发现是否依赖强理论假设；是否未通过前沿任务或模型进行验证；或者是否对超参数高度敏感？ ☒

6. **问题难度**。 是否不太可能有任何实用系统能够在这项任务上显著超越人类？ ☐

7. **人类不可靠性**。 这种方法是否严重依赖手工设计特征、专家监督或人类可靠性？ ☐

8. **竞争压力**。 推进这种方法是否会严重牺牲原始智能、其他通用能力或经济效用？ ☐

A.2 安全性与能力的平衡

本节将分析这项工作与通用能力的关系，以及它如何影响安全与通用能力所带来危害之间的平衡。

9. **概述**。 这如何在提升通用能力的同时，更大程度地提升安全？

回答： 本文提出的框架强调，任何提升通用能力的努力，都应首先考虑更广泛社会背景下的安全，以及如何实现这一点。由于这是一个思考通用能力的框架，其弊端在于，这些能力可能会被更全面地用于不良目的——然而，似乎在国际事务中起破坏作用的国家早已明白这一点。因此，世界上其他希望改善世界现状的国家、组织和个人，需要更加努力地吸取这一教训，并以更全面、更整合的方式推动人工智能对齐。因此，这一框架对安全的促进作用大于其阻碍作用。

10. **红队测试**。 这会以何种方式加速通用能力的发展或存在性风险的出现？ **回答：** 这只会以与任何全面的社会模型可能有助于更好地思考如何将技术融入社会相同的方式，加速人工智能通用能力的发展。不过，恶意行为者或许会利用它，试图在各种人工智能之间播下相互冲突的对齐目标，或暗中引导人工智能走向一种极其严重的目标错位。如果通用人工智能以这些方式构建，那么安全就会受到损害——但这似乎几乎就是当前人工智能研究的默认模式（常常混乱、不同层级之间未能对齐，或被恶意行为者蓄意滥用）。因此，更明确地阐明问题的真实形式，似乎会……

善意者可能的进步方向将是改进，而不是退化。 11. **一般任务**。 这项工作是否推动了此前被视为通常能力研究主题的任务的进步？ □12. **一般目标**。 这是否改进或促进了以下方向的研究：一般预测、分类、状态估计、效率、可扩展性、生成、数据压缩、执行清晰指令、助人性、信息丰富性、推理、规划、研究、优化、（自）监督学习、序贯决策、递归自我改进、开放式目标、访问互联网的模型，或类似能力？ ☒13. **与一般能力的相关性**。 已分析的能力是否被认为可以通过一般认知能力或受教育程度进行高度预测？ □14. **通过能力实现安全**。 这是否随着其他能力或人工智能研究的发展而推进安全，或作为其结果推进安全？ ☒

A.3 补充说明与其他考量

15. **其他**。关于这项工作和存在性 -值得一提的风险？ **回答：**本文描述了人工智能相关风险出现的背景，并为其构建了一个框架。作为一种描述和框架，它让人们能够以一种新的方式理解这一问题，也希望能为寻求解决人工智能对齐问题的人带来启发。由于对齐问题贯穿一个连续谱，从内容审核等当代问题一直延伸到通用人工智能和 X 风险，这一框架应能为寻求打造全面安全的人工智能的人提供有益见解。而对于那些想要反其道而行之的人，他们早已知道如何播下纷争与邪恶的种子。这个框架是帮助行善者做得更好的工具。对于心怀恶意的人来说，双重用途本质上毫无用处，因为他们早已在战术、作战和战略层面熟知这些内容。真正能从中获得最大帮助的，应是那些致力于全面向善的人；他们可以在战术、作战和战略层面运用这一框架，让人工智能对所有人都更加安全。