ALIGNMENT NEWSLETTER 中文版

对齐周报第 96 期

对齐周报是每周出版的出版物，其中包含与全球人工智能对齐有关的最新内容。在此处找到所有对齐周报资源。特别是，你可以浏览此电子表格，查看其中的所有摘要。此处的 音频版本（可能尚未启用）。

强调

人工智能联盟播客：2018 年和 2019 年技术性人工智能对齐发展概述（Lucas Perry，Buck Shlegeris 和 Rohin Shah）*（由 Rohin 总结）：这个与 Buck 和我的播客大致围绕我撰写的评论构建（AN＃84），但还有更多的辩论，并探讨了悲观主义和乐观主义的具体观点。我怀疑每个读者都会有他们感兴趣的部分。由于很多讨论本身都是摘要，所以我不会尝试进一步总结。以下是涵盖的主题列表：

- 我们对采用不同方式研究人工智能对齐的乐观和悲观
- 将人工智能视为高风险的传统论点
- 将智能体建模为预期的效用最大化器
- 雄心勃勃的价值学习和规格学习/狭义的价值学习
- 代理与优化
- 健壮性
- 扩展到超人的能力
- 普遍性
- 影响正则化
- 因果模型、神谕和决策论
- 不连续和连续起飞场景
- 人工智能引发的存在风险的可能性
- 通用人工智能的时间表
- 信息危害

技术性人工智能对齐

技术议程和优先级

AI服务作为研究范式（*Vojta Kovarik*）（由Rohin总结）：CAIS报告（AN＃40）建议，未来的技术发展将由*AI*服务系统驱动，而不是单个整体式AGI代理。但是，自该报告发表以来，没有太多的后续研究。该文档认为这是因为报告中引入的任务和服务的概念不适合形式化，因此很难对其进行研究。因此，它为服务系统提供了一个简单的抽象模型，其中每个服务获取一些输入，产生一些输出，并使用某些损失函数进行训练，并且这些服务彼此互连。（某些服务可以由人类提供，而不是由AI系统提供。）然后列出了可以用这种通用观点解决的几个研究问题。

Rohin 的观点：我原本希望有一个特定于AI的问题列表，但实际列表非常广泛，在我看来，这就像一个子问题列表：“你如何面对技术发展来设计一个好的社会”。例如，它包括失业，系统维护，潜在的勒索，旁道攻击等。

学习人类意图

使AI与人类价值观保持一致意味着选择正确的指标（*Jonathan Stray*）（由 Rohin 总结）：最近，人们对推荐系统的缺陷引起了广泛关注，特别是在优化诸如参与度之类的简单指标时 —— 我们称之为“狭义价值对齐”的一个例子。这篇文章重构了 Facebook 和 YouTube 自 2015 年和 2017 年以来如何将更好的指标纳入其算法。例如，Facebook 发现学术研究表明，“有意义的社交互动”改善了幸福感，但被动地消耗了内容却使幸福感恶化。结果，他们更改了推荐算法的指标以更好地对此进行跟踪。他们是如何测量的？看来，他们只是询问了成千上万的人（在 Facebook 内外）最有意义的内容是什么，并以此为训练模型来预测“

然后，作者将这种狭义价值对齐方式与通用人工智能对齐方式进行了对比。他的主要观点是，狭义价值对齐应该更容易解决，因为我们可以从现有系统在现实世界中的行为中学习，而我们获得的见解可能对通用人工智能对齐至关重要。我将在结论结尾引用一句话：“我的论点并不是那么多人应该使用人工智能来优化幸福感。相反，我们生活在一个已经在进行大规模优化的世界中。我们可以选择不评估或调整这些系统，但没有理由想象无知和不作为会更好。”

Rohin 的观点：尽管我经常对对齐动机感到乐观（AN＃80），但即使我惊讶地发现Facebook 似乎在努力使其推荐算法与幸福感保持一致时也感到惊讶。在某种程度上，推荐算法仍然是有害的（idk，可能是对还是错，idk），这向我暗示，鉴于你得到的反馈稀少，可能很难给出好的推荐。当然，还有更多的愤世嫉俗的解释，例如 Facebook 只是想看起来像他们在乎幸福，但如果他们真的关心他们，他们会做得更好。我倾向于第一种解释，但是很难区分这些假设。

尽管这篇文章声称狭义价值对齐应该比通用人工智能对齐容易，但实际上我不确定。使用通用人工智能对齐，你将有一个非常有力的假设，即你要对齐的人工智能系统是智能的：这似乎很有帮助。例如，Facebook 正在使用的推荐系统可能无法预测将改善人类福祉的方面，而在这种情况下，狭义对齐注定会失败。通用人工智能并非如此（取决于你对通用人工智能的定义）—— 它应该至少能够像人类一样做。挑战在于确保人工智能系统的真正动力（AN＃33）这样做，而不是他们是否有能力这样做；窄对齐需要解决两个问题。

更少的东西是：人类行为的概率模型的反思（*Andreea Bobu*，*Dexter RR Scobee*等）（由 Asya 总结）：本文介绍了一种用于推断人类偏好的机器人新模型，称为 LESS。传统的玻尔兹曼噪声理性决策模型假设人们大致优化了奖励函数，并根据其指数奖励选择了轨迹。玻尔兹曼模型在对不同离散选项之间的决策进行建模时效果很好，但在对连续空间中的人类轨迹进行建模（例如路径查找）时会遇到问题，因为它对轨迹数非常敏感，即使它们是相似的-使用玻尔兹曼模型的机器人必须预测人类是通过左侧的一条路径还是右侧的三个非常相似的路径之一来绕过障碍物导航，因此默认情况下，它将为每个路径分配相同的概率。

为了解决这个问题，LESS 通过将每条轨迹视为连续空间的一部分并将每条轨迹映射到特征向量来预测人类行为。选择一条轨迹的可能性与其与其他轨迹的特征空间相似度成反比，这意味着对相似的轨迹进行了适当的加权。

本文在几个实验环境中测试了 LESS 与玻尔兹曼模型的预测性能，其中包括人为构造的任务，要求人类在障碍物周围导航的相似路径之间进行选择，以及现实世界的任务，人类对自由度为 7 的机械臂进行合适的行为演示。通常，当给定少量人类行为样本时，LESS 的性能优于玻尔兹曼模型，但随着样本量的增加，其表现也同样好。在机械臂任务中，当将演示汇总到一个批次中并且一次对整个批次进行推断时，玻尔兹曼模型的表现会更好，这表示尝试近似“平均”用户而不是为每个用户定制行为。该论文声称，发生这种情况是因为玻尔兹曼在稀疏地区的示威活动中学习过多，而在密集的示威活动中则学习不足。随着样本数量的增加，你对“真实”轨迹空间的逼近会越来越好，因此 10 个轨迹集的变化越来越小，这意味着玻尔兹曼模型的表现不会太差。由于单批演示汇总了演示，因此它在逼近“真实”轨迹空间方面具有相似的效果。

本文指出，此方法的局限性在于依赖于一组预先指定的机器人功能，尽管少数实验结果表明，当添加少量无关功能时，LESS 的性能仍优于玻尔兹曼模型。

Asya的观点：这似乎是一篇很好的论文，并且非常像玻尔兹曼模型的自然扩展，包括了对相似轨迹的解释。如论文所述，我在很大程度上担心是否依赖于一组预先指定的机器人功能-在更复杂的推理情况下，手动指定相关功能可能不切实际，并且很难让机器人进行推断。在最坏的情况下，似乎错误指定的功能可能会通过暗示无关的相似性而使性能比玻尔兹曼模型差。

Rohin 的观点：（请注意，本文来自 InterACT 实验室，我也是其中一员。）

人类行为的玻尔兹曼模型具有几个理论上的依据：它是最大熵（即最小编码信息）在轨迹上的分布受到特征期望与观察到的人类行为的期望相匹配的约束；它是最大的熵分布，假设人类满足某个阈值以上的预期奖励，等等。我从未发现这些引人注目，而是将其视为简单得多的东西：你希望模型编码这样一个事实，即人类更有可能采取好的行动要比采取坏的行动多，并且你希望模型为所有轨迹分配非零概率；玻尔兹曼模型是满足这些条件的最简单模型。（你可以想象将模型中的指数删除为“甚至更简单”，但这等效于奖励函数的单调变换。）

我认为本文提出的模型之前符合我的两个标准，并增加了第三个标准：当我们可以基于相似性对轨迹进行聚类时，我们应该将人类视为在聚类之间进行选择，而不是在轨迹之间进行选择。给定一个良好的相似性指标，这似乎是人类行为的一个更好的模型-如果我正在行走并且路径中有一棵树，我会选择绕树的那一边，但我不会仔细思考我的脚步将落在哪里。

我发现玻尔兹曼在稀疏区域过学习（overlearn）的说法是不直观的，因此我在此评论中对其进行了更深入的研究。我的总体看法是，该主张通常会在实践中成立，但不能保证。

防止不良行为

好奇心杀死了猫和渐近最优（*Michael Cohen* 等人）（由 Rohin 总结）：在没有重置的环境中，渐近最优因子是最终发挥最佳作用的一种。（可能是智能体首先以决定性的次优方式进行自我干预，但鉴于当前的受干扰位置，最终它将推出最佳策略。）本文指出，此类智能体必须进行很多探索：毕竟，下一步很可能永远是砍断手臂永远给你最大的回报-你怎么知道不是这样吗 由于它必须进行大量探索，因此极有可能陷入“陷阱”，而不再能获得高额回报：例如，其执行器被破坏了。

更正式地讲，本文证明了当渐近最优主体起作用时，对于任何事件，该事件都会发生，或者在一定的时间之后，即使发生概率很小，也没有可识别的机会导致该事件发生。将其应用于“智能体被销毁”事件，我们可以看到该智能体最终被销毁了，或者在物理上也无法销毁该智能体，即使是本身也无法销毁-鉴于后者似乎不太可能，我们希望最终智能体被摧毁。

作者建议安全探索不是一个明确的问题，因为你永远都不知道探索时会发生什么，因此他们建议智能体应该由导师或父母指导探索（AN#53）（另请参见delegative RL（AN#57），避免了经由人为干预灾难，和屏蔽更多的例子）。

Rohin 的观点：在我对Safety Gym（AN#76）的意见中，我提到了对安全探索的零违规约束将如何要求已经满足该约束的导师或父母；所以从这个意义上讲，我同意本文的意思，只是使该声明更为正式和准确。

尽管如此，我仍然认为可以安全地进行有意义的探索：一旦你了解了一个有合理信心的好的模型，便可以找到模型中不确定的部分，但至少可以有信心它不会带来永久性的负面影响，你可以在那里进行探索。例如，我经常“探索”我喜欢的食物，不确定我会喜欢哪种食物，但是我非常有信心食物不会毒害我。（但是，这种探索概念与强化学习中通常使用的探索概念完全不同，最好将其称为“基于模型的探索”或类似的东西。）

杂项（对齐）

贝叶斯进化到灭绝（*Abram Demski*）（由 Rohin 总结）：考虑一个贝叶斯学习者，它使用贝叶斯规则更新各种假设的权重。如果假设可以影响未来的事件和预测（例如，它可以写出日志，从而影响将来要问的问题），那么影响未来的假设只能以他们可以预测的方式进行选择贝叶斯规则，而不是直接预测未来而又不试图影响未来的假设。从某种意义上讲，这是贝叶斯更新方面的“近视”行为：贝叶斯规则仅优化每个假设，而没有考虑对整体未来准确性的影响。如果彩票假设，这种现象也可能适用于神经网络（第4号侦探）成立：在这种情况下，每个“门票”都可以被视为竞争假设。

人工智能战略与政策

再谈“天网”：核指挥自动化的危险诱惑（*Michael T. Klare*）（由 Rohin 总结）（H / T Jon Rodriguez）：虽然我不会在这里完整地总结本文，但我发现了解如何一些学者正在考虑军事自动化的风险，以及了解当前自动化工作的实际情况。我发现一句话特别有趣：

联合人工智能中心（JAIC）主任杰克·沙纳汉中将在 2019 年 9 月在乔治城大学举行的会议上说：“国防部将人工智能功能整合到国防部中，你将找不到更强大的支持者。”是我暂停的地方，这与核指挥与控制有关。” 他说：“我读到了。我的直接回答是，“不。我们不。”’

人工智能对齐播客：关于致命自主武器（*Lucas Perry* 和 *Paul Scharre*）（由 Flo 总结）：《无敌军：自主武器与战争的未来》一书的作者 Paul Scharre 谈论了有关致命自主武器（LAW）的各种问题），包括当不同的人通过“军备竞赛”表达不同的意思并且自治程度不同时，谈论围绕自动武器进行军备竞赛的困难，军方对人工智能系统缺乏对分配转移的鲁棒性的可靠性的需求以及对抗性攻击，战争法是否正确处理法律，以及使人陷入困境的优缺点。

尽管自动武器不太可能直接造成生存风险，但通过建立网络并为未来人工智能问题的合作与合作准备机构，建立限制武器的努力可能是有价值的。

人工智能的其他进展

深度学习

具有神经切线的快速简便的无限宽网络（*Roman Novak*，*Lechao Xiao*，*Samuel S. Schoenholz* 等）（由 Zach 总结）：深度学习的成功促使研究人员探索为什么它们如此有效的函数逼近器。一个关键的见解是，增加网络层的宽度可以使其更容易了解。更精确地讲，随着宽度被发送到无穷大，网络的学习动态可以用泰勒展开近似，并成为内核问题。该核的极限形式精确，称为神经正切核（NTK）。最终，这使我们能够使用称为高斯过程的简单模型对网络进行建模。不幸的是，很难通过分析来表明这一点，而创建有效的实现则很麻烦。作者通过引入**“神经切线”**解决了这个问题，该库使创建无限宽的网络像使用 **PyTorch** 或 **TensorFlow** 之类的有限对等网络一样容易。它们包括通过全填充，残差连接，前馈网络支持卷积，并支持多种激活功能。此外，还提供对 CPU、GPU和TPU的现成支持。此外，可以通过精确的贝叶斯推断对有限合奏进行不确定性比较。

阅读更多：论文：神经切线：Python中的快速简便的无限神经网络

> **Zach** 的观点：我看了一下代码库，发现那里有足够的文档，这使我很容易尝试训练自己的无限宽网络。作者得出一种实用的方法来计算精确的卷积 NTK，这使我印象深刻，这似乎是本文的主要技术贡献。尽管作者指出，进入所谓的“内核机制”需要一定的条件，但实际上，似乎你往往只能摆脱较大的网络宽度。如果没有别的，我建议至少仔细阅读他们可用的笔记本电脑，或者看

看他们呈现的神经网络的可视化结果，这些神经网络收敛于高斯过程，这依赖于微妙的大数定律应用。