在此处查找所有Alignment Newsletter资源。特别是，您可以注册或查看此电子表格中所有摘要中的摘要。

强调

人工智能安全需要社会学家（*Geoffrey Irving*等人）（Richard 总结）：人工智能安全的一种方法是“向人类提出大量关于他们想要的问题，训练他们的价值观的 ML 模型，并根据学到的价值优化人工智能系统的行为“。然而，人类给出的答案是有限的，有偏见的，而且经常彼此不一致，因此人工智能安全需要社会科学家来弄清楚如何改进这些数据 —— 最终可能会从数千或数百万人那里进行收集。特别重要的是能够设计严谨的实验，从对人类认知和行为的跨学科理解中汲取经验。作者讨论了 debate （ AN#5）作为安全技术的案例研究，其成功取决于经验问题，例如：默认情况下人类作为法官的技能如何？我们能培养人才成为更好的法官吗？有没有办法限制 debate，以便更容易判断？

这个论点背后有几个关键的前提。首先，尽管存在人为偏见，但对于人类价值观的问题却有正确的答案 —— 如果给出所有相关信息和无限时间思考，我们可能会认可这个答案。然而，只要他们能够识别出他们不确定并且什么都不做的情况（尽管在某些情况下不采取行动会造成伤害，例如自动驾驶），AI 并不一定总能找到答案。汽车在中途停止转向，似乎可以通过无所作为来避免最令人担忧的长期灾难。乐观的另一个原因是，即使是社会科学实验的不完整或负面结果，也可能有助于为未来的技术安全研究提供信息。但是，在某些情况下，我们尝试取推理的系统完全不同于现在能够测试的情况 —— 例如 AI debater 比人类强太多。

Richard的观点： 这篇文章及其随附的论文对我来说似乎非常明智。虽然我对如何提供关于超越人类辩论者的人类辩论数据有些怀疑，但看起来值得尝试获得更多经验信息似乎值得。请注意，虽然本文主要讨论辩论，但我认为它的许多论点都适用于包含任何人在内的安全方法（也可能是其他方法）。目前我认为 Ought 是最关注收集人类数据的安全小组，但我期待看到其他研究人员这样做。

技术AI对齐

技术议程和优先次序

FLI播客：2018年人工智能的突破和挑战与David Krueger和Roman Yampolskiy（*Ariel Conn*，*David Krueger*和*Roman Yampolskiy*）：David 和 Roman 总结回顾了 2018 年人

工智能的进展并推测其影响。Roman 确定了一种模式，我们看到像 AlphaZero（AN#36），AlphaStar（AN#43）和AlphaFold（AN#36）这样的 突破现在这么频繁，当一个新的出现时，它似乎不再令人印象深刻。另一方面，David 对 Dota 和星际争霸的进展印象不那么深刻，因为两个 AI 系统都能够执行人类永远不会做的动作（Dota 的快速反应时间和星际争霸的每分钟高动作）。他还认为这些项目并没有像 AlphaZero 那样产生任何明确的一般算法见解。

在深度强化学习结合机器人方面，David 提到了 Dactyl（AN#18）和 QT-Opt（我记得阅读和喜欢了的，但显然我没有记住放入 AN 中）的重大进展 。他还指出 GAN 已经有了显着的改进，并特别谈到了特征转换。Roman 注意到进化算法的性能正在提高。

David 还指出了如何通过创建可扩展的算法，然后使用大量计算，引用 AI 和 Compute（AN#7），解释 AI 计算趋势（AN#15）和 重新解释 AI 和计算（AN#38）。

在政策方面，他们谈到了 Deep Fake 以及人工智能可能正在快速推进以保持其安全隐患的总体趋势。他们确实发现研究人员开始接受他们的研究确实具有安全性和安全性。

在 AI 安全方面，David 指出，主要的进展似乎是使用 超越人类反馈的方法，包括 辩论（AN#5），迭代放大（在本期简报中经常讨论，但该论文在 AN#30中）和 递归奖励建模（AN#34）。他还将 未受限制的对抗性例子（AN#24）确定为未来需要关注的领域。

Rohin的观点：我大致同意这里确定的人工智能进展领域，尽管我也可能会加入 NLP，例如 BERT。我不同意细节 - 例如，我认为 OpenAI Five（AN#13）比我当时预期的要好得多，如果我还没有看过 OpenAI Five，那么 AlphaStar 也是如此。事实上，他们做了一些人类做不到的事情，几乎没有削弱成就。（我的观点非常类似于 Alex Irpan 在AlphaStar上发表的文章。）

诡计转弯，模拟和脑机接口（*Michaël Trazzi*）

学习人的意图

AI Alignment播客：人类认知和智力的本质（*Lucas Perry* 和 *Joshua Greene*）（由 Richard 总结）：Joshua Greene 的实验室有两个研究方向。首先是我们如何将概念结合起来形成思想：一个允许我们理解任意新情景的过程（即使是我们认为没有发生过的情景）。他讨论了他最近的一些研究，该研究使用脑成像来推断人类在考虑复合概念时

会发生什么。虽然 Joshua 认为思想的组合性质很重要，但他认为要建立 AGI，有必要从“基础认知”开始，其中表征来自感知和物理行为，而不仅仅是学习操纵符号（如语言）。

Joshua 还致力于道德的心理学和神经科学。他讨论了他最近的工作，其中提示参与者考虑罗尔斯的无知面纱论证（当做出影响许多人的决定时，我们应该这样做，好像我们不知道我们是哪一个），然后要求评估道德困境如手推车问题。Joshua 认为，公正的概念是道德的核心，它推动人们走向更多的功利主义思想（尽管他想将功利主义重新塑造为“深刻的实用主义”来解决其公关问题）。

从不完美的示范中学习模仿（*Yueh-Hua Wu* 等）

通过空间接口评估强化学习来学习用户偏好（*Miguel Alonso Jr*）

解释性

规范黑盒模型以提高可解释性（*Gregory Plumb*等）

稳健性

对抗性示例是噪声中测试误差的自然后果（*Nicolas Ford*，*Justin Gilmer*等人）（由Dan H 总结）：虽然之前在AN＃32中对此进行了总结 ，但该草案更具可读性。

用合成噪声提高机器翻译的稳健性（*Vaibhav*，*Sumeet Singh*，*Craig Stewart*等）（由 Dan H 总结）：通过将噪声（如拼写错误，单词遗漏，俚语）注入机器翻译模型的训练集中，作者能够提高自然数据的性能。虽然这个技巧通常不适用于计算机视觉模型，但它可以用于 NLP 模型。

推动学生正确学习：元学习者对腐败标签的渐进式渐变校正（*Jun Shu*等）

杂项（对齐）

人工智能安全需要社会科学家（*Geoffrey Irving* 等）：参见强调！

人工智能战略和政策

不集中的人不是一般情报（莎拉康斯坦丁）：这篇文章认为人类可以浏览 GPT-2 制作的故事 （AN＃46）无法判断它们是由机器生成的，因为在略读时我们无法注意到其写作中明显的逻辑不一致。关键引用：“OpenAI 已经实现了通过自动驾驶仪对人类进行图灵测试的能力”。这表明虚假新闻，社交操纵等将变得更加容易。然而，它也可能迫使人们

学习检测人类和机器人之间差异的技能，这可以让他们学会分辨他们何时积极关注某些东西并“实际学习”而不是略读“低阶相关”。

Rohin 的观点：我在阅读 GPT-2 结果的时候注意到了这种效果的变种 —— 我的大脑很快就陷入了撇脂的模式而没有吸收任何东西，尽管感觉更像是我做了评估，没有什么可以从中获益内容，如果目标是避免假新闻，这似乎没关系。我还发现这是关于我们的低级别，轻松模式匹配之间的差异以及我们更加努力和准确的“逻辑推理”的特别有趣的证据。

AI的其他进展

勘探

InfoBot：通过信息瓶颈转移和探索（*Anirudh Goyal* 等）

强化学习

AlphaStar 上的一篇逾期文章（*Alex Irpan*）：这两篇文章的 第一篇文章讨论了 AlphaStar（AN#43）对星际争霸社区和广大公众的影响。我专注于第二个，它讨论了AlphaStar的技术细节和含义。这篇文章的一些内容与我对 AlphaStar 的总结重叠，但这些部分更好地充实并有更多细节。

首先，模仿学习是一个令人惊讶的良好基础政策，达到了黄金级别的水平。这是令人惊讶的，因为你可能会认为 DAgger 问题是极端的：因为星际争霸游戏中有太多的动作，你的模仿学习策略会产生一些错误，然后这些错误会在回合的很长一段时间内复杂化策略使得进一步远离正常人类进入没有接受过训练的状态的策略。其次，基于人群的训练可能至关重要，将来也很重要，因为它可以探索整个策略空间。

第三，主要的挑战是使 RL 达到良好的性能，之后它们很快变得很好。经过多年的研究，让 Dota 和星际争霸机器人得到了不错的体验，然后经过几天的培训让他们成为世界级的。有趣的话：“尽管 OpenAI 的 DotA 2 智能体在专业团队中失利， 但他们能够在 80% 的时间内通过 10 天的训练击败他们的旧版的智能体”。

第四，AlphaStar 有很多研究成果。这表明通过将大量技术放在一起并观察它们的工作情况，可以获得巨大的收益，而目前这种情况并未发生。这有很好的理由：如果一种技术建立在一个简单的标准算法之上，而不是必须考虑与其他技术的相互作用，而这些技术可能会或可能无法正确地进行比较，那么评估技术要容易得多。如果我们只是把正确的

东西放在一起，我们现在可以做一些很酷的结果，这种工作也让我们在新设置中测试技术，看看哪些实际上是一般的，而不是只有在原来的评价中。

Rohin的观点：我非常喜欢这篇文章，并且几乎同意其中的所有内容。在模仿学习点上，我也发现模仿学习的效果令人惊讶。亚历克斯表示，可能是人类数据有足够的变化，代理人可以学习如何从错误的决策中恢复。我认为这是最好的部分解释 - 存在巨大的组合爆炸，不清楚为什么你不需要更大的数据集来覆盖整个空间。也许在任何现实的复杂环境中都有“自然”的表示，你开始在他们正在使用的计算水平上准确学习，一旦你拥有那些，那么模仿学习和足够的变化就可以很好地运作。

关于将技术放在一起的最后一点，我认为这有时候值得做，但往往可能不是。任何真正的任务都可以做到这一点，因为这是对现实技术的测试。（这里星际争霸算作“真正的”任务，而 Atari 则没有;标准就像“如果任务成功自动化，无论如何解决，我们都会留下深刻印象”。）我不太热衷于将技术拼凑在一起基准。我认为通常这些技术通过添加类似于良好的归纳偏差的东西，通过恒定的乘法因子来提高样本效率; 在这种情况下，将它们放在一起可能会让我们更快地解决人工基准，但它并没有给我们很好的证据证明“归纳偏差”对于现实任务是有益的。

学习从稀疏和未指定的奖励中推广（*Rishabh Agarwal*等）

通过元学习奖励塑造（*Haosheng Zou*，*Tongzheng Ren et al*）

研究连续深层强化学习的推广（*Chenyang Zhao*等）

深度学习

神经架构搜索的随机搜索和再现性（*Liam Li*等）

新闻

MIRI暑期研究员项目（*ColmÓRiain*）：CFAR 和 MIRI 将于 8月9日至24日举办 MIRI暑期研究员项目。申请截止日期为 3 月 31日。

RAISE正在推出他们的MVP（*Toon Alfrink*）：人工智能安全卓越之路将于周一开始发布反强化学习和迭代放大课程。他们正在为他们的测试小组寻找志愿者，他们将在 RAISE 的指导下每周研究这些材料大约一整天，并提供有关材料的反馈，特别是任何混淆的来源。