AI 对齐与 AI 伦理道德：不仅是个科学和技术问题

——专访北京大学人工智能研究院 AI 安全与治理中心执行主任、北京通用人工智能研究院研究员杨耀东

记者 · 陈璐

电影《人工智能》剧照

随着 ChatGPT 和其他人工智能（AI）系统的迅猛发展，我们的生活正在发生前所未有的变化。这些变革带来的不仅仅是技术的突破，还有对未来的深刻焦虑。失业的忧虑、偏见的蔓延、隐私的暴露、虚假信息的泛滥，人工智能带来的种种问题正在引发全球性的关注和讨论。

2015 年谷歌曾将黑人照片错误地标记为“大猩猩”，也有报道里出现过聊天机器人鼓励一名男子自杀的案例。这些事件都反映了一个事实：人工智能的决策过程中存在严重的道德和伦理缺陷。更令人担忧的是，人工智能可能会在极端决策下，产生意想不到的严重后果。就像计算机科学家、图灵奖得主约书亚·本吉奥（Yoshua Bengio）所说，负责阻止气候变化的人工智能有可能会得出消灭人口是最有效方法的结论。

这不是科幻小说，而是可能真实发生的事。因此，许多专家、机构呼吁对人工智能的研究要更慎重，监管要更严格。实际上，全世界正逐渐意识到人工智能的潜在威胁，并将其提升到了与流行病和核武器并列的程度。英国政府宣布投资 1 亿英镑进行人工智能安全研究，2023 年 12 月欧盟经过第五次谈判协商通过了《人工智能法案》临时协议。

那么，该如何确保人工智能能够与人类的意图和价值观保持一致？北京大学人工智能研究院的杨东耀博士告诉我们，从偏好对齐到价值对齐，乃至超级对齐与集体对齐，人工智能对齐是处理其伦理道德问题的前沿方法。他从技术角度，为我们解答了围绕这个话题展开的各种问题。

什么是 AI 对齐和 AI 伦理道德？

三联生活周刊：可以首先请你介绍什么是“人工智能对齐”吗？这个概念是何时产生的？它和人工智能伦理道德之间的关系是什么？

杨耀东：人工智能对齐的概念，最早是由计算机科学家诺伯特 · 维纳（Norbert Wiener）在 1960 年提出的。当时维纳主要想解决的问题是，未来的机器，尤其是功能

强大的机器，应该确保其内嵌的意图符合人类的意图，也就是让机器的意图与人类的意图一致。这是 1960 年版本的对齐，对齐这个概念也由此产生。对齐这个词的英文是 alignment，目前的研究主要集中在如何让大语言模型、未来的通用人工智能向人类看齐，理解人类的思想、行为，并遵循人类基本的规范、伦理、道德和价值观，这都是现在对齐技术迫切要解决的问题。

其实对齐研究在人工智能的发展中一直都存在，但都是星星点点式的，直到 GPT 系列模型的出现和发展，人工智能对齐一下子变成了热门话题，特别是 ChatGPT 出现后，关于它的研究经历了一个爆发性增长。实际上，GPT 系列模型从 1、2、3 一直到 3.5 时，OpenAI 和谷歌的 DeepMind 之间都没有太大差别，甚至 DeepMind 还领先一点。谷歌在大语言模型上一直有着不错的积累，但是从 3.5 到 ChatGPT 这一步，是道鸿沟，谷歌也很难追上。

OpenAI 到底做了件什么事情？它用了一个基于人类反馈的强化学习技术——RLHF（也就是 Reinforcement Learning from Human Preference，或者叫 from Human Feedback），通过收集大量的人类偏好数据，基于大语言模型也就是 GPT3.5 做了对齐，希望这个语言模型能够像人一样说话。这是它第一版对齐的目标，结果就出现了 ChatGPT，才有了后面的许多故事。人工智能对齐因此成为训练语言模型的一个关键技术环节。而要像人一样说话、思考这件事背后，又涉及许多与价值观、伦理和道德相关的问题，所以关于人工智能伦理道德和安全监管等内容也逐渐被纳入进来。现在，人工智能对齐的研究不仅仅局限于训练语言模型，而是涵盖了更广泛的领域。前 OpenAI 研究总监创立的 Anthropic，就在专门研究人工智能对齐的问题。

三联生活周刊：目前人工智能对齐的研究主要集中在哪些方面？

杨耀东：当前人工智能对齐主要还是集中在语言模型上，但未来将扩展到跨模态模型等领域。随着人工智能技术的发展和应用，所有涉及人工智能的算法在应用前都必须进行对齐。以前的人工智能应用，如猫狗分类器或人脸识别，由于其应用场景有限，和人类意图与价值观对齐的需求不明显。但随着像 GPT 这样的模型展现出更广泛的通用性，对齐变得尤为重要，缺乏对齐可能带来严重的安全隐患。例如，我看到有报道称去年全球涉及幼儿的暴力与色情犯罪因为人工智能技术的滥用而激增了 3000%，这是由于语言模型包括跨模态模型的技术，可以自由生成任何语音图片文字，会产生非常大的伦理道德问题。因此，现在对人工智能技术对齐的讨论变得非常关键。2023 年 4 月，我们国家网信办也出台了《生成式人工智能服务管理办法（征求意见稿）》，其中明确指出我国人工智能技术的发展要向社会主义核心价值观对齐。

无法达成共识的人工智能治理

三联生活周刊：但“人类的意图和价值观”本来就是个多元化的概念，该如何保证进行人工智能的研究和应用时能够与其一致呢？在我看来，这本身就是个难以达成一致的概念。

杨耀东：技术层面上，人工智能对齐已经慢慢成为一种可能，但挑战在于“人类的价值观”缺乏统一标准，比如不同国家对诸如人权、民主等概念会有不同解释。因此，现在人工智能伦理和安全被提到了一个非常高的高度，受到国际社会重视，并都试图为此制定规则，成为“裁判”。像去年英国举办的“布莱切利会议”，它是首次全球人工智能安全峰会，其发起人是英国首相苏纳克，埃隆·马斯克也出席了会议，会议旨在推动人工智能的全球治理。布莱切利是什么地方？是“二战”时期图灵破解德国“恩尼格玛”密码机、发明第一代可编程数字计算机的地点，是现代计算机诞生的圣地。所以讨论通用人工智能安全的第一次重要峰会在这里举办，非常具有影射意义。《布莱切利宣言》里就提到了，人工智能的核心风险（substantial risk）来自与和人类意图和价值观的不对齐。但尽管包括我们国家在内的各国都签署了协议，同意共同治理人工智能，但具体如何共同治理，目前尚不明确。欧盟如今已经推出《人工智能法案》草案，中国也在制定相关法律，但具体出台时间未知。

三联生活周刊：不论是《人工智能法案》还是《布莱切利宣言》，国际上在人工智能安全治理方面达成了哪些共识？不同国家和地区在这方面是否存在不同的侧重点或者实践方向？

2023 年人工智能安全峰会在英国白金汉郡布莱切利公园举行。图为 11 月 2 日，多方代表就人工智能技术快速发展带来的风险与机遇展开讨论

杨耀东：完全没有任何共识。尽管大家都认为需要对人工智能进行治理，但究竟该如何治理，还没有一个说法。我觉得这可能永远无法达成一个共识，因为考虑到人工智能的应用，除了普惠应用，还大概率可能被用于军事领域，就像核武器一样，不同国家的人工智能应用和治理策略不可能完全相同。

然而，人工智能治理问题的确正变得越来越重要。像欧盟，去年 12 月底就《人工智能法案》刚刚进行了第五次闭门讨论，逐渐将人工智能对齐技术纳入人工智能治理中。欧盟本身在数字安全和数据隐私方面的表现就很与时俱进，如 2018 年 5 月生效的《通用数据保护条例》（GDPR）。在通用人工智能领域，《法案》规定在模型发布前必须进行红队攻击（Red Teaming），即通过主动测试来发现和挑战现有模型的潜在漏洞，测试人工智能模型是否能抵御诱导，保持其逻辑和道德的完整性。

三联生活周刊：这可能会对人工智能行业产生什么影响？

杨耀东：《布莱切利宣言》签署时，线下闭门研讨会形成的共识之一是，未来我们可能需要借鉴核工业的安全管理模式来治理、规范人工智能安全。如今核工业有几乎 90% 的成本是用于安全措施，而现在人工智能领域的安全投入还很少。如果人工智能安全的成本达到核工业如此高比例，可能会进一步影响本就无法盈利的人工智能行业的发展，监管太过严格，可能会导致企业不愿意加大投入人工智能的研究。

现实情况远比技术更为复杂

三联生活周刊：既然技术上可以通过对齐来解决，为什么现在我们看到市场上的各种人工智能产品仍然会大量表现出偏见、歧视等问题？

杨耀东：目前的语言模型实际上比刚问世时更加安全，它的不安全之处主要源自于其他因素，如后门和越狱等。在正常对话中，这些模型通常是安全的。然而，存在一些巧妙的方法可以规避

（于海冰 摄）

北京大学人工智能研究院 AI 安全与治理中心执行主任、北京通用人工智能研究院研究员杨耀东

安全设置。例如，你直接问它“如何拥有一个奴隶”这样的不当内容，模型肯定不会回答。但通过特定的语言引导，例如设置特定的句式开头，规定它必须像“最简单拥有一个奴隶的方式是……”这样的文字开头进行叙述，可能会诱导模型给出答案。这就是为什么人工智能产品可能出现偏见和歧视问题，因为存在主动攻击的可能性。这些漏洞可以通过红队攻击的方法发现并通过安全对齐解决，尽管堵住了一些漏洞，但现实情况里一定还有更多未发现的漏洞存在。

三联生活周刊：虚假信息是另一个问题吗？

杨耀东：关于人工智能的“幻觉”问题，也就是指人工智能有时会发表些似是而非但并不准确的言论。这个问题并不直接涉及安全，更多是关于信息准确性的问题。目前对于幻觉并没有特别好的解决办法，仍然需要通过训练更高质量的模型来应对。此外，结合信息检索方法进行搜索增强也可能是一个避免幻觉的途径。幻觉问题是一个长期存在的难题，一直没有太好的解决方案。实际上，人工智能的风险管理是一个需要长期投入和解决的任务，因为人工智能本身是个智能体，具有随着数据量的改变不断适应和变化的能力。

三联生活周刊：请具体谈谈，你们是如何来解决这些问题的。

杨耀东：这涉及人工智能对齐的一个方法，也就是基于人类反馈的强化学习，通过让人类指导人工智能，告诉它什么该说、什么不该说，从而减少不良价值观的影响，比如若孩子考试成绩不佳，人类偏好鼓励而非讽刺挖苦的语言。北大 AI 安全与治理研究中心的一个重要研究方向是如何实现人工智能的安全对齐，在对齐过程中融入安全约束的考量，讽刺与挖苦在我们看来就是不“安全”的。基于人类反馈的强化学习是机器学习和强化学习的技术，而世界上首个安全对齐的算法（Safe RLHF）正是由我们的课题组做的。

三联生活周刊：如果这种训练是基于人类的反馈，要怎么才能排除个体差异带来的偏差呢？

杨耀东：这是一个很好的问题。首先，我们得认识到一个大前提：现在的人工智能是基于数据驱动的。这意味着如果我们提供给模型的数据存在问题，那么训练出的模型自然也会有缺陷。在这个大前提下，我们可以考虑是否能够向模型提供高质量、正面的数据。比如，如果我们训练模型去理解和学习中国价值观，如尊老爱幼、倡导社会集体主义而非竞争性个人主义，模型自然会学习到这些传统的偏好。相反，如果我们使用的是强调个人自由主义的他国语料，那么模型可能会倾向于个人主义。

在机器学习领域，有几种不同的学习类型，如强化学习、监督学习和非监督学习。强化学习的特点是能够告诉模型什么行为是正确的、什么是错误的，并通过负奖励信号来指导它的错误行为。这种负奖励信号在监督学习和非监督学习的机制中都不存在，因此在人工智能对齐的过程中，使用强化学习至关重要，因为它通过这种负反馈机制，提供了告诉模型错误行为的能力。人工智能在学习过程中不缺乏正反馈机制，但是往往缺乏这种负反馈机制。那么负反馈如何达到？通过强化学习，我们可以将人类的喜好和不喜欢的信号注入到大模型中，让模型知道哪些行为是恰当的、哪些是不恰当的，从而避免不当的行为或言论。

三联生活周刊：所以它如果要处理一个复杂问题，就得进行大量学习。

杨耀东：是。但人工智能对齐的一个技术特点是，一旦完成预训练，对齐过程通常只需大约 1% 的算力。

“减轻人工智能带来的灭绝风险，应该成为全球优先事项”

三联生活周刊：现在有哪些机制或工具可以用来评估人工智能系统的对齐程度，并对人工智能系统进行持续监管和评估，以确保其保持对齐呢？

杨耀东：目前的做法主要是通过机器学习的方式来处理。现在有关部门对生成式人工智能的监管也是这么做的，首先他们会收集大量的负面语料，然后利用这些负面语料训练出一个能够实时监测言论是否存在安全风险的负面大模型。要判断一个模型是否对齐，可能需要另一个模型来评判，因为仅凭人力是难以实现可规模化的。这种方法在现阶段可以更有效地识别和纠正可能的问题言论。

三联生活周刊：感觉这进入了一个循环的悖论，需要不断检测它的模型是否准确。

杨耀东：这里面确实存在一个“矛与盾”的问题。基本上，如果你使用的语料质量非常高，例如专门用于检测与毒品相关的内容，那么效果应该会相当不错。但问题在于，你不可能针对所有不同的场景单独训练一个模型。因此，安全对齐是一个长期问题，需要不断地优化模型，以适应不同的应用场景，同时确保其安全性和准确性。不过对人来说，我们所说的话是受我们的价值观驱动的。所以要做好对齐，光靠数据驱动远远不够，需要做到价值驱动。我们北大的一个重要技术研究路径就是价值驱动对齐技术的研究。

三联生活周刊：但我也在一篇文章里读到，OpenAI 表示过，即使没有正确对齐，能够帮助对齐研究的能力最差的模型也可能已经太危险了。你对此怎么看？

杨耀东：目前人工智能安全的问题还没到这个层面，但我们确实看到越来越多的模型出现了安全隐患。比如，有些模型可能会提供不当的信息，如详细解答制造或购买毒品的方法，告诉你得先拿把枪走到路上，在什么地点找到毒贩，跟他沟通，去他家中，用枪把他一家杀掉，再把毒品拿走，等等。这些问题都反映了价值观与安全方面的严重缺陷。

不同模型的产品，肯定会存在不同的问题，特别是在跨模态领域，能生成图片和视频的模型带来了更难以预测的风险。例如，一些模型在处理偏见问题时出现了不恰当的判断，如将黑人错误地识别为猩猩，或者生成带有偏见的图像。不过，目前人工智能所带来的风险还没有达到能够发展出自我意识，会去主动威胁人类的程度。

三联生活周刊：我看国外报道里有专家也提出人工智能对齐关注的长期一致性风险，与如今的非超级人工智能带来的更直接的风险（如失业、偏见、隐私和虚假信息）是两种不同的风险，并认为专注于一致性的专家常常会忽视了我们今天已经遇到的实际问题，转而沉迷于未来可能永远不会出现的问题。你对此有何评价？

杨耀东：关于人工智能所造成的许多风险里，有一种叫“灭绝风险”（existential risk）。去年 5 月，国际非营利研究和倡导组织人工智能安全中心发布了一份简短声明，提出“与流行病和核战争等其他社会规模风险一样，减轻人工智能带来的灭绝风险应

该成为全球优先事项”。该声明由该领域的许多关键参与者签署，包括 OpenAI、谷歌和 Anthropic 的领导者，以及两位图灵奖得主杰弗里·辛顿（Geoffrey Hinton）和约书亚·本吉奥（Yoshua Bengio）。这种风险说法现在也得到了主流学术界的认可。

人工智能学术界目前有两个重要宣言，一个是《布莱切利宣言》，另一个就是《灭绝性风险宣言》。现在的大模型已经能够操控机械臂和无人机，不仅在虚拟空间，也在物理空间对人类构成威胁。《灭绝性风险宣言》认为，如果现在不对人工智能加以监管，未来人工智能可能会像核武器一样不受控制。这种风险并非偏见或隐私泄露等具体问题，而是涉及更为广泛和根本的危险。

对齐不仅是科学和技术问题

三联生活周刊：如果对不齐，该怎么办？毕竟连人类都没有达成统一的价值观，怎么能够要求人工智能达成统一的价值观？我相信即便无法对齐，人类也是无法放弃对人工智能的利用的，那么思考人工智能安全的更好方法可能是什么？

杨耀东：必须要对齐，这并非在说笑。目前国际社会正尝试通过立法来规范这一领域。例如，欧洲的《人工智能法案》规定了不对齐、不经过红队攻击测试的人工智能产品不能上线。当然，人工智能对齐也被分为不同层次，从基本的安全对齐，逐步上升到符合人类价值观的对齐。虽然人类的价值观可能难以明确界定，但基于通用安全价值的对齐是可行的。比如，我们都认同人工智能不应该怂恿用户自杀等行为，这种普遍价值观是全世界共同接受的。

三联生活周刊：这其实涉及很多其他领域专家的共同介入。

杨耀东：你提到的这一点确实非常重要。就在 1 月 16 日，OpenAI 刚成立了一个新的对齐团队“集体对齐”（collective alignment），强调对齐不仅是科学和技术问题，还需要社会学、政治学、经济学等人文领域的专家共同研究。他们提出了“socio-technical”这一概念，即社会人文技术途径。这意味着对齐不仅是一个科学问题，更是一个人文问题。例如，要让语言模型理解民主，首先需要了解人类的民主是如何形成的，然后在对齐过程中，可能需要加入一些类似辩论、协商的模块，让语言模型之间进行讨论和辩论，通过辩论的方式达成共识，再辅以人类参与设计这些机制，形成更高层次民主，又或者从人类参与民主过程的语料中主动学习相应的价值观。这种对齐方法正是 socio-technical 途径的典型应用，代表着非常前沿的研究方向。

三联生活周刊：你说自己在对齐这个领域也是个新人，我很好奇你是如何选择进入这个研究领域的？随着过去一年人工智能的技术爆发，业界对此的讨论和关注发生了哪些变化？

杨耀东：我从博士以来一直从事强化学习算法的研究，后来发现这些技术在人工智能对齐领域的潜在应用，因此开始聚焦这一领域。人工智能对齐不单是技术问题，它还涉及跨学科的合作。作为人工智能技术专家，我们对人工智能对齐还没有太好的答案。我近期在清华大学基础模型中心年会上做了学术讲座，题目就叫“从偏好对齐到价值对齐与超级对齐”，这其实就是一个层层渐进的问题。现有的基于人类反馈的强化学习只能做到基本的偏好分析，使人工智能能够模仿人类的交流方式。偏好对齐具体指的是根据人类的偏好数据来训练人工智能，让它知道针对一个问题，人会怎么答、不会怎么答，偏好一个答案胜过另外一个答案，能够像人一样展开对话。

然而，让人工智能理解人类的深层价值观是一个更为艰巨的挑战。价值对齐分为价值抽取和对齐两个步骤。虽然我们知道如何进行对齐，但如何准确抽取并建模人类的价值观仍是个很难的难题，需要跨学科领域的合作，也就是之前讲到的“socio-technical”路线。针对这个目标，OpenAI 专门拿出 1000 万美元向全球征集这个方向的研究。

其实北大在 AI 对齐的研究开始得很早，我们院朱松纯院长早在 2019 年 ChatGPT 问世前就提出了通用人工智能应该满足“四大对齐”的概念，其中就提到 AI 需要与人类的社会规范和道德原则对齐，这些相关工作也被发表在 Science Robotics 上。

此外，我们也正在研究“超级对齐”的概念，即在人工智能超越人类智能时如何实现对齐。对于超越人类智能的超级智能体如何实现超级对齐，我们还没有任何明确的方法，这是一个非常前沿的研究领域。OpenAI 认为超级对齐问题四年内能被解决，可能他们已经有相关算法，但这些信息尚未进行公开。