2023 年第 5 期（总第 8 期）

人工智能与国际安全研究动态

ARTIFICIAL INTELLIGENCE AND INTERNATIONAL SECURITY STUDIES REVIEW

国际智库及媒体对 ChatGPT 领域动向的评估

清华大学战略与安全研究中心
CENTER FOR INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

战略与安全
SECURITY & STRATEGY
清华大学战略与安全研究中心
CENTER FOR INTERNATIONAL SECURITY AND STRATEGY TSINGHUA UNIVERSITY

国际智库及媒体对 ChatGPT 领域动向的评估

编者按：为推进人工智能与国际安全领域的相关研究，清华大学战略与安全研究中心（CISS）组织研究团队定期跟踪最新国际研究动态，重点关注人工智能应用对国际安全带来的风险挑战，并针对人工智能安全领域国际动态、智库报告、学术论文等资料进行分析。本文是CISS推出的人工智能与国际安全研究动态第8期，主要聚焦国际智库及媒体关于ChatGPT领域动向的评估。

1. 斯坦福大学：政府需在解决公众关切和促进负责任的人工智能开发间取得平衡

斯坦福大学以人为本人工智能研究所（HAI）发布《2023 年度人工智能指数报告》，全面分析了人工智能的现状，考察了研究贡献、模型进步、人工智能滥用激增、私人投资模式、工作机会、行业应用及其他趋势等各个方面，认为政策制定者必须在解决公众关切和促进负责任人工智能开发之间取得平衡，促进经济增长和改善美国人的生活质量。报告要点如下：工业界领先于学术界；人工智能继续发布最先进的结果，但许多基准的同比改进仍微不足道；达到基准饱

和的速度正在加快；人工智能既保护环境，也在损害环境；人工智能模型开始迅速加速科学进步；有关滥用人工智能的事件数量激增；几乎每个美国工业部门对人工智能相关专业技能的需求都在增加；过去十年中，人工智能领域的私人投资首次出现同比下降；虽然采用人工智能的公司比例趋于稳定，但它们实现了显著的降成本增收入；政策制定者对人工智能的兴趣正在上升；中国公民是对人工智能产品和服务态度最积极的人群之一。

https://aiindex.stanford.edu/report/

2. 哈佛商学院刊文分析人工智能监管

5 月 2 日，哈佛商学院网站刊登美国 UPS 基金会商业物流名誉教授詹姆斯 •赫斯克特的文章《人工智能应如何监管》。文章认为，如果我们需要对人工智能的开发和使用进行监管，使用统一的全球标准和实践似乎不太可能。中国、欧盟和巴西等国已经立法规范本国的人工智能发展。作为人工智能发展最先进的国家之一，美国才刚刚开始出现要求某种形式监督的呼声。欧盟的人工智能法案代表了迄今为止最广泛的监管方法，并对监管提出了一系列复杂挑战。例如，人工智能学习所依据的数据库集的范围界定，将决定其输出内容的准确性，以及是否存在偏见。那么，输出的某些用途是否应该得到限制，以及是否需要关于基于人工智能工作准确性的免责声明，目前这些问题都有待解决。

https://hbswk.hbs.edu/item/how-should-artificial-intelligence-be-regulated-if-at-all

3. 哈佛大学贝尔弗中心：构建可信赖的人工智能增强透明度

美国哈佛大学贝尔弗中心刊登其伯克曼互联网与社会研究中心研究员内森·桑德斯、IBM 安全中心特别顾问布鲁斯·施奈尔的文章《我们可以构建值得信赖的人工智能吗》。文章认为，构建可信赖的人工智能需要增强透明度。人工智能无法完全被信任的原因在于人工智能应用主要是由大型科技公司创建和运营的，其目的主要是带来更多的利益和利润。有时我们被允许与聊天机器人互动，但它们从来都不是我们的。这是一种利益冲突，会破坏信任。构建可信赖的人工智能需要进行系统性变革。首先，可信赖的人工智能系统必须是用户可控的；其次，可信赖的人工智能系统应该透明地允许用户控制它使用的数据。第三，用户还应该了解人工智能系统可以代表他们做什么。我们需要在不同的环境中测试人工智能系统，观察它们的行为，并为它们如何响应我们的指令建立心智模型。只有当人工智能系统对它们的能力、使用的输入、何时共享以及它们正在进化以代表谁的利益等问题均透明时，才有可能获得信任。

https://www.belfercenter.org/publication/can-we-build-trustworthy-ai

4. 哈佛大学贝尔弗中心：人工智能也可促进民主

4 月 21 日，美国哈佛大学贝尔弗中心网站刊登 IBM 安全中心特别顾问布鲁斯·施奈尔，约翰·霍普金斯大学政治学家亨利·法瑞尔、伯克曼互联网与社会研究中心研究员内森·桑德斯的文章《人工智能如何帮助民主》。文章认为，人工智能可以促进公共利益，促进民主。这需要人工智能不受大型技术垄断的控制，而是由政府开发并可供所有公民使用。要为民主构建辅助人工智能。例如，主持两极分化的政策讨论、解释法律提案的细微差别，或在更大范围的辩论中阐明一个人的观点。这提供了一条让大型语言模型（LLM）与民主价值观“对齐”的途径：通过让模型生成问题的答案、犯错并从人类用户的反应中学习，而不让这些错误损害用户和公共领域。在怀疑人工智能和技术的政治环境中捕捉这种用户互动和反馈通常是具有挑战性的。很容易想象，那些抱怨像元宇宙这样的公司不值得信任的政客们，对于政府在技术开发中发挥作用的想法会更加愤怒。下一代人工智能实验应该先在州和市进行。商业上可用的和开源的大型语言模型可引导这一过程，并为联邦投资公共人工智能选项建立势头。

https://www.belfercenter.org/publication/how-artificial-intelligence-can-aid-democracy

5. MIT 专家：急于部署生成式人工智能将导致世界变得更糟

5 月 11 日，美国麻省理工学院网站刊登了对麻省理工学院助理教授兼计算机科学与人工智能实验室（CSAIL）首席研究员雅各布·安德烈亚斯的采访文章《3 个问题：雅各布·安德烈亚斯谈大型语言模型》。作者表示，急于部署生成式人工智能工具将导致世界变得更糟。模型正在趋于能够构建事实，但即使是今天最先进的模型也会产生不正确的事实。由于模型输出时显示的统计数据表面看起来是无误的，所以模型会以人类难以察觉的方式生成有误代码。从长远看，解决生成代码的真实性、连贯性和正确性尚有很长的路要走。
https://news.mit.edu/2023/3-questions-jacob-andreas-large-language-models-0511

6. MIT 斯隆管理评论刊文分析生成式人工智能的挑战

5 月 18 日，美国《麻省理工学院斯隆管理评论》网站发布文章《负责任的 AI 程序是否已为生成式人工智能做好准备？专家存疑》，作者为《麻省理工学院斯隆管理评论》和波士顿咨询集团 (BCG) 召集的一个国际人工智能专家小组。文章认为，强调负责任的人工智能（RAI）原则，并不断更新 RAI 计划以应对新风险，会有效减少生成式人工智能工具带来的风险。传统人工智能系统专注于检测模式、做出决策、完善分析、分类数据和检测欺诈。生成式人工智能使用机器学习来处理大量的视觉或文本数据，其中大部分是从互联网上的未知来源抓取的。生成式人工智能工具，如 ChatGPT、

Bing Chat、Bard 和 GPT-4，正在通过使用在大量数据上训练的机器学习算法来创建原始图像、声音和文本。生成式人工智能引入了新的风险，即从数据分类和预测转向内容创建。对此，文章建议如下：巩固 RAI 基础，并致力于保证 RAI 程序不断发展以应对新的及不可预见的风险；对员工进行培训，加强生成式人工智能风险的教育和意识建设；为最大限度地降低风险，应考虑对服务供应商进行初步风险评估、偏差测量和持续风险管理实践。

https://sloanreview.mit.edu/article/are-responsible-ai-programs-ready-for-generative-ai-experts-are-doubtful/

7. MIT 斯隆管理学院：产业界正主导人工智能研究

5 月 18 日，美国麻省理工斯隆管理学院网站刊登文章《产业界正主导人工智能研究》。文章指出，与学术界相比，产业界在主导人工智能研究发展方面拥有不可比拟的优势。企业在访问大型数据集方面拥有天然优势，它们的运营需要与用户和设备交互，会产生大量数据。产业界在人才方面也拥有优势，自 2006 年以来，学术界研究人员的数量基本持平，而产业界的招聘人数增加了 8 倍。在产出方面，每年最大型的人工智能模型中，约 96%来自产业界。人工智能要更新迭代，新功能必须完全建立在现有模型之上。而产业界既拥有人工智能模型，也可负担得起价格高昂的计算能力；而学术界缺乏开展这项工作的资源。

https://mitsloan.mit.edu/ideas-made-to-matter/study-industry-now-dominates-ai-research

8. MIT 斯隆管理学院：人工智能是人类迫在眉睫的生存威胁

5 月 23 日，美国麻省理工斯隆管理学院网站刊登文章《为何杰弗里·辛顿对人工智能敲响警钟》。文章指出，图灵奖获得者、多伦多大学名誉教授杰弗里·辛顿担心，越来越强大的机器以不符合人类最佳利益的方式超越人类，人类可能无法限制人工智能的发展。以 GPT-4 为代表的大型语言模型的知识是人类的一千倍。模型能够持续学习并轻松分享知识。相同人工智能模型的多个副本可以在不同的硬件上运行，但做的事情完全相同。人工智能还可以“通过阅读所有的小说和马基雅维利曾经写过的一切”来操纵人。人工智能还有可能学会自己编写和执行程序。在最坏的情况下，“人类只是智能进化的一个过渡阶段。”生物智能进化到创造数字智能，吸收人类创造的一切，并开始获得对世界的直接体验。辛顿表示没有看到任何明确的解决方案。停止开发人工智能可能是合理的，但由于公司和国家之间的竞争，以及在医学等领域取得的成就，那是天真且不可能的。

https://mitsloan.mit.edu/ideas-made-to-matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai

9.《自然》杂志：化学界的人工智能革命尚未发生

5 月 17 日，英国《自然》杂志网站发表社论《对于化学家来说，人工智能革命尚未发生》。文章称，对于包括化学在内的许多科学领域来说，大规模的人工智能革命尚未发生，因为缺乏可用于训练人工智能系统的数据。如果化学家想要充分利用生成式人工智能工具，他们需要实验数据、模拟数据、历史数据、不成功实验的数据等更多数据。同时，研究人员必须确保由此产生的信息是可访问的。除非人工智能系统拥有全面的化学知识，否则它们最终可能会导致不正确的实验结果。若想人工智能系统创建或访问更多更好的化学数据，一种可能的解决方案是建立从已发表的研究论文和现有数据库中提取数据的系统，加速人工智能在有机化学中的应用；另一种解决方案是使实验室系统自动化。然而，与人类化学家相比，该系统只能进行相对狭窄范围的化学反应。此外，人工智能工具需要开放数据。但即便如此，也可能不足以让人工智能工具发挥其全部潜力。化学应用要求计算机模型比最好的人类科学家更好。只有采取措施收集和共享数据，人工智能才能满足化学方面的期望。

https://www.nature.com/articles/d41586-023-01612-x

10. 美国企业研究所：人工智能会促使个性化学习时代到来

5 月 5 日，美国企业研究所刊登研究员约翰·贝利发表文章《个性化学习的承诺从未兑现，今天的人工智能是不同

的》。作者认为，人工智能会推动个性化学习时代的到来，让所有学生都能充分发挥潜能，并培养更公平、更有效的教育体验。新一代人工智能应用的成功得益于：更智能的能力、推理引擎、能够解释和响应自然语言命令、前所未有的规模。这类人工智能工具可成为辅导助手：根据每个学习者的独特需求和兴趣提供量身定制的解释、指导和实时反馈；人工智能可以通过自动执行帮助教师快速生成教案创意、开发工作表、起草测验以及翻译；基于人工智能的反馈系统能够对学生的写作提出建设性意见与帮助。尽管这些技术前景广阔，但同样重要的是要认识到它们的局限性。它们不应取代教师的专业知识和判断力，而只能提供支持性的帮助。

https://www.aei.org/op-eds/the-promise-of-personalized-learning-never-delivered-todays-ai-is-different/

11. 《泰晤士报》：**教育型人工智能或引发教科书革命**

4 月 22 日，英国《泰晤士报》网站刊登文章《教育型人工智能聊天机器人引发教科书革命》。文章认为，在生成式人工智能可能引发的混乱中，教育首当其冲。某初创企业开发了一款生成式聊天机器人，利用一家世界顶级教科书出版商的某本畅销教材对其进行训练。在被“投喂”了几乎每一页内容后，这款机器人模型高分通过测验，从而提升了开发新一代人工智能“学习伙伴”的可能性，这种“学习伙伴”将成为单一学科的专家。各种生成式人工智能产品的工作原

理相同，推动生成答案的功能实际上是一种经过高度训练的猜测，即根据数十亿个句子猜测接下来最有可能出现的单词。然而，像 ChatGPT 这样的大型语言模型本身并不具备知识，而是在猜测。结果是，它们常常捏造事实，并且十分自信地陈述事实。该初创企业认为可以克服这个问题，方法是让人工智能工具产生“记忆”，这种记忆只包括有限的、人们想要的特定材料，然后把答案限制在这个范围内。这一理念能否站稳脚跟是一个悬而未决的问题，但显而易见的是，教育正在迅速发生变化，从业者正在争相调整。

https://www.thetimes.co.uk/article/educating-ai-chatbots-spells-a-textbook-revolution-08n69cndz

12. 加拿大《新闻报》：人工智能须融入专业人员的培训中

5 月 22 日，加拿大《新闻报》发表社论《我们应该害怕教育中的人工智能吗》。文章认为，对人工智能置之不理的教育机构从长远来看会使自己处于难以为继的境地。人工智能甚至必须融入许多学生、教师、程序员等专业人员的培训中。对于人工智能工具在高等教育中的使用，还有很长的路要走。魁北克省奥陶瓦地区大学最近对七所大学的数百名师生进行调查后发现，51%的受访教师表示他们计划“修改”他们的评估，“以避免人工智能工具的剽窃”。这表明许多教师已意识到人工智能工具的普及所带来的风险。随后，高级教育委员会和科学技术伦理委员会宣布成立联合专家委

员会，在年底前提交报告，重点关注使用人工智能进行学生评估和学习以及教师培训所面临的挑战。社论称，教育部门的各个参与环节都要非常迅速适应当前形势，直面人工智能工具普及带来的挑战。

https://www.lapresse.ca/debats/editoriaux/2023-05-22/faut-il-avoir-peur-de-l-intelligence-artificielle-en-education.php#

编辑：孙成昊、郑乐锋、王家琪

审核：肖茜、董汀