2022 年第 2 期（总第 2 期）

人工智能与国际安全研究动态

ARTIFICIAL INTELLIGENCE AND INTERNATIONAL SECURITY STUDIES REVIEW

美国《人工智能权利法案蓝图》评析

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

美国《人工智能权利法案蓝图》评析

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）组织专门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2期，主要聚焦美国《人工智能权利法案蓝图》提出的人工智能技术应用五项原则，并对相关观点进行整理分析。

2022 年 10 月 4 日，美国白宫科技政策办公室（OSTP）发布《人工智能权利法案蓝图》（Blueprint for an AI Bill of Right），旨在指导自动化系统的设计、使用和部署。本专题将重点梳理该蓝图提出的人工智能技术应用五项原则，并在此基础上归纳总结相关智库观点分析。

一、人工智能技术应用五项原则及其实践保护措施

《蓝图》提出人工智能技术应用的五项原则和相关技术指南，具体内容如下：

1.建立安全和有效的人工智能系统

原则内容：关注不安全或无效自动化系统的意外使用和影响给使用者造成的损害，强调在当前实践基础上建立更积极和更广泛的保护措施，以增强人们对于自动化系统的信心并推动系统创新发展。

具体措施：在自动化系统设计、实施、部署、获取以及维护阶段对可能受不同影响的使用者进行调研；在系统部署前完成与现实操作环境相近的大量测试，持续地主动识别和降低自动化系统的潜在风险；确保自动化系统包含必要且持续的监测和校准程序；负责开发或使用自动化系统的主体应建立明确的治理结构和治理程序；建立相关性的高质量数据，跟踪和审查派生数据源，对敏感领域数据提供额外监督以确保使用的安全有效。

2.建立算法歧视保护措施

原则内容：关注因算法导致的歧视，强调系统设计和使用应当遵照平等的原则。自动化系统的设计者、开发者以及使用者应当采取并一以贯之地落实未雨绸缪的举措，保障系统算法过程不会基于人们的种族、肤色、民族、性别、宗教、年龄、国籍、残疾、退伍军人身份、基因信息及其他任何受法律保护的类别等产生歧视。

具体措施：事前预防和事中监督。事前预防主要包括设计阶段的公平性评估；数据代表性检验；防止人口特征代理；设计、开发、使用三阶段的无障碍性保障；不间断的差异评

估及缓解等。事中监督即在算法系统使用过程中，有关方应提供算法影响评估报告，并明确评估系统的使用方和纠偏措施。

3.保护数据隐私

原则内容：个人数据不应被滥用，且使用者保有使用数据的主导权。通过特定设计默认保护隐私，强化监督和规约监视保护公民自由和权利，完善公众收集、使用、访问、传输和删除数据的权利应当被尊重。改变目前个人数据“通知”、“选择”的做法。健康、就业、教育、刑事司法和金融等敏感领域的数据应当受到额外保护。系统应能够提供报告已证明对使用者数据决定的尊重。

具体措施：在系统设计开发的全生命周期中评估隐私风险，提供技术和政策层面的改善措施；提供机制支持，保证用户对数据的许可、访问和控制；与敏感领域相关的数据收集和使用必须确保有限和适度、接受道德审查和质量审计并提供定期公开报告。并通过独立评估和报告机制证明数据隐私保护的真实性。

4.强调通知和透明度的重要性

原则内容：需要向公众提供清晰、简短且易于理解的通知，以便其了解一个自动化系统是否正在使用，其运作方法和相应影响是什么以防止潜在危害。

具体措施：系统提供清晰、及时、便于理解的操作说明，包括通知系统更新、相关内容查找方法等。为自动化系统决

策和行为提供的解释需要根据目的、解释对象、风险等级以及有效性定制，并提供相应报告。

5.鼓励开发选择退出机制

原则内容：在不损害更广泛公众利益的适当情况下，用户能够选择退出系统或使用人工替代。退出机制不应当增加不合理负担。在教育、卫生、就业和刑事司法等敏感领域应增加有针对性的选项。

具体措施：为人工替代服务提供简洁说明，界定人工介入的各种适当情况，人工服务应当在用户选择退出时立即出现并保障便利程度；考量这一机制下的公平度；对参与该项人工替代服务的有关各方进行相应培训和评估；充分发挥政府监管作用以减少该机制下的歧视现象并完善运作效果。

二、相关观点分析

《人工智能权利法案蓝图》的发布受到美国社会各界广泛关注，尤其是各科技巨头、研究院及智库纷纷发表分析与评论，现将相关观点整理总结如下。

第一，该蓝图有利于保护人工智能时代下的公民网络隐私，并蕴含强化人工智能伦理治理的特有价值。10 月 29 日，人工智能 SaaS 企业 SalesChoice 首席执行官**辛迪·戈登（Cindy Gordon）**在福布斯新闻发表文章《美国人工智能权利法案蓝图将推进人工智能治理》。她认为，人工智能解锁了一个未经检查的、正在破坏人类隐私安全的数据世界。但

如果美国能够成为人工智能治理的全球领导者，就能够促进人工智能技术创造价值、带来更多效益，而这一蓝图就为发展现有政策、法规创造了一个基础框架。Mozilla 基金会执行董事**马克·苏尔曼（Mark Surman）**也表示，该法案能够帮助保护公众隐私。他指出，人工智能已渗透进人类生活的方方面面，但现有模型往往为收集个人数据而建立，具有不透明、有偏见的特征，这与该蓝图所倡导的原则不相符合。此外，数字权利组织 Access Now 的分析师**威尔玛丽·埃斯科托（Willmary Escoto）**特别指出该法案在伦理方面的贡献，认为它将服务于全球非洲裔与拉丁裔的基本公民权利。

第二，该法案蓝图的关注范围明确而广泛，将推动人工智能在更多领域安全采用。世界经济论坛网站于 10 月 14 日刊登其执委会成员**凯·弗斯·巴特菲尔德（Kay Firth-Butterfield）**等共同撰写的文章《了解美国〈人工智能权利法案蓝图〉及其如何帮助人工智能保持责任感》。三位作者首先梳理了该文件内容，并指出该蓝图明确澄清其适用范围仅限于可能对美国公众权利产生影响的自动化系统，而一般不涵盖使用人工智能的其他工业与业务。但同时，该蓝图也将美国对人工智能权利的关注拓展至人工智能技术在借贷、人力、监控等更多领域的应用实例。艾伦人工智能研究院联合创始人、软件自动化供应商 UiPathInc.的人工智能咨询委员会主席**奥伦·埃奇奥尼（Oren Etzini）**表示，如果实施得当，该蓝图能够有效减少人工智能不当使用，并促进人工智能在

医疗、驾驶、企业生产等方面的有益采用。

第三，部分科技企业高管担心该法案的监管“过度”，从而遏制人工智能创新发展步伐。比如，软件公司 CircleCI 首席技术官**罗布·祖伯（Rob Zuber）**称，该蓝图虽能发挥监管作用，但也可能扼杀创新。他认为，技术领导者应营造一种环境，使其团队在承担治理责任的同时，负有推动人工智能发展的责任。Alphabet Inc.’s Google 前任首席执行官**埃里克·施密特（Eric Schimidt）**同样认为，早期监管的介入将阻碍新的发现，因此“不到万不得已，我不会选择监管”。

第四，该法蓝图被认为是一份“框架性”“起点式”文件，具有细化与优化空间，实际效果亦有待实践检验。斯坦福大学以人为本人工智能研究院（HAI）主任**拉塞尔·沃尔德（Russel Wald）**指出，该蓝图缺乏设立相关执法机制的细节，比如由联邦层面协调的监控、审计和审查行动。

人工智能与数字政策中心（Center for AI and Digital Policy）负责人**马克·罗滕伯格（Marc Rotenberg）**则在肯定该蓝图价值的同时表示，他希望看到一些对最具争议的人工智能部署的明确禁令，如利用面部识别进行大规模监控等。

对话智能研究机构 Converseon 首席执行官兼创始人**罗伯·基（Rob Key）**则针对算法模型提出三项改进建议。一是每一分类模型应有明确的性能评估指标以供使用，二是应在模型建立的标记过程中吸纳不同观点，并通过测试以努力达到最高精度，三是应有专人负责在问题出现时及时修改模型

的能力。

科技游说集团 BSA 的人工智能政策总监**沙恩德拉·沃森（Shaundra Watson）**还指出，继该蓝图关注人工智能风险影响后，“更重要的是下一步”，即确保该法案蓝图能在实际应用中产生真正可靠的保护效果。

撰稿：王叶湑、王一诺、汤文君、王星懿
审核：肖茜、董汀、孙成昊、郑乐锋

参考文献：

[1] “White House Issues ‘Blueprint for an AI Bill of Rights’,” *The Wall Street Journal*, October 4, 2022, https://www.wsj.com/articles/white-house-issues-blueprint-for-an-ai-bill-of-rights-11664921544, 访问日期：2022 年 11 月 20 日。

[2] Cindy Gordon, “The USA AI Blueprint Bill of Rights Advances AI Governance,” Forbes, https://www.forbes.com/sites/cindygordon/2022/10/29/the-usa-blueprint-for-an-ai-rights-advances-usa-ai-governance/?sh=33c52f6950f5, 访问日期：2022 年 11 月 20 日。

[3] Kay Firth-Butterfield, Karen Silverman and Benjamin Larsen, “Understanding the US ‘AI Bill of Rights’- and how it can help to keep AI accountable,” World Economic Forum, October 14, 2022, https://www.weforum.org/agenda/2022/10/understanding-the-ai-bill-of-rights-protection/, 访问日期：2022 年 11 月 20 日。

[4] Melissa Heikkiläarchive, “The White House just unveiled a new AI Bill of

Rights," MIT Technology Review, October 4, 2022, https://www.technologyreview.com/2022/10/04/1060600/white-house-ai-bill-of-rights/, 访问日期：2022 年 11 月 20 日。

[5] "Conversecon's Response to the White House Blueprint for an AI Bill of Rights," October 13, 2022, https://converseon.com/resources/blog/response-whitehouse-blueprint-ai/, 访问日期：2022 年 11 月 20 日。