人工智能的安全风险及治理模式探索

鲁传颖　张璐瑶

内容提要：人工智能技术在21世纪有了新的发展机遇，但不断涌现的安全问题也成为制约其发展的重要因素。要想做好人工智能的安全治理，不仅需要了解技术安全挑战存在于何处，更应该系统了解这些风险的成因，并将破解治理困境作为核心着力点。在相互叠加的人工智能安全风险和治理困境中，以多方参与性、时间敏感性和反馈迭代性为核心特征的敏捷治理模式逐渐被多国采纳。这种有别于传统模式的治理思路不仅为破解人工智能的不确定性安全难题和提升治理能力提供了有效方案，还进一步为创建普遍、公平、合理的人工智能全球治理机制提供了契机，是解决人工智能安全问题的良方。

关键词：人工智能安全 | 敏捷治理 | 中美技术脱钩 | 全球技术治理

作者介绍：鲁传颖，上海国际问题研究院网络空间国际治理研究中心秘书长、研究员，主要研究领域为网络安全与新兴技术治理；张璐瑶，上海国际问题研究院硕士研究生。

人工智能被认为是通用型技术，将会广泛应用于国家、社会和个人的方方面面。但是，人工智能自身技术的复杂性和应用的广泛性相结合，也会带来极为复杂的安全挑战。人工智能面临内生安全、应用安全，以及与其他新兴技术之间存在交叉安全等多重风险。这三重风险的相互叠加，对总体国家安全观所覆盖的政治安全、国土安全、军事安全、经济安全、文化安全、社会安全、科技安全、网络安全等多个领域构成了挑战。与此同时，现有的治理理念和模式已经无法应对人工智能这一类新兴技术所带来的安全挑战，亟需以敏捷治理为

理念来探索新的治理模式。

一、总体国家安全观视角下的人工智能安全风险

作为通用性技术，人工智能在政治、经济、社会和军事等方面的运用既能创造丰厚的价值，也会带来复杂甚至具有颠覆性的安全风险。从总体国家安全观的视角来看，人工智能的安全风险主要体现在以下五个方面。

第一，人工智能军备竞赛已经悄然打响，带来新的军事安全困境。“随着人工智能技术的成熟，它将会被越来越广泛地应用于军事领域，武器系统、军事策略、军事行动，甚至战争的意义可能会发生深刻改变，人类社会也有可能在进入人工智能时代之后迎来一个不同的军事安全环境。”① 在这一最新出现的军事增量面前，各国特别是具备技术能力的大国很难抵御人工智能驱动的新式战争武器的诱惑：他们希望抢占军事力量的发展先机，尽可能地拉开与竞争对手的实力差距，以最大限度地谋取军事安全，甚至收获技术转化所产生的巨大经济利益。在这种心理的驱动下，一场以人工智能技术为核心的新的军备竞赛恐难避免。此外，在人工智能技术的加持下，大量无人作战武器开始从幻想走进现实，但问题也随之产生：2020 年 3 月，一架土耳其 STM 公司生产的“卡古 -2”军用无人机，在被编程为不依靠操作员的情况下，在利比亚战场上跟踪并攻击了正在撤退的利比亚国民军，导致一人死亡。有报道称，这可能创下了致命性自主武器在自主模式下攻击人类的首个案例。② 可以预见，未来致命性自主武器将会越来越多地走向战场。当然，这一趋势背后存在的军事伦理、国际法理和战略稳定问题，值得各国重视和思考。

第二，人工智能被应用于干扰各国政治秩序，存在巨大政治安全风险隐患。人工智能技术及其背后的大数据和算法能够潜移默化地影响公众行为，直接对

① 封帅、鲁传颖：“人工智能时代的国家安全：风险与治理”，《信息安全与通信保密》，2018 年第 10 期，第 36 页。

② 石汉娟、张慧军、高庆龙：“‘卡古 -2’给人类拉响警报”，《解放军报》，2021 年 6 月 17 日。

国内政治行为产生干扰，给选举制度带来深层次挑战。以人脸识别和深度造假技术为例，在美国 2020 年总统选举之前，就有一段众议院议长南希·佩洛西的演讲视频被深度造假技术进行了调整，这被视为对民主党领导人的抹黑。更值得一提的是，当发现这一视频是被伪造后，美国各网络平台对此的反应也不同——YouTube 及时对视频进行了下架，但 Facebook 拒绝对此视频进行删除 。[①] 在这一事件中，对个人的生物信息识别涉及隐私安全问题，在互联网社交媒体上传播捏造的信息则涉及网络安全问题和声誉问题，而针对政府领导人的行动又可能会危害到政治稳定和国家安全，在与互联网平台企业交涉的时候还需要关注合规问题。

第三，人工智能技术与资本深度结合，进一步强化国家、阶层之间的不平等。经济方面，在人工智能技术的影响下，资本与技术在经济活动中的地位获得全面提升，而劳动力要素的价值则受到严重削弱，由此引发机构性失业风险、贫富分化和不平等现象。更进一步，人工智能技术带来的全球经济结构调整，将引导全球资本和人才进一步流向技术领导国，由此留给发展中国家走上现代化道路的机遇期将变得更加有限。

当人工智能技术所推动的社会经济结构变革逐步深入时，资本和技术力量的垄断地位有可能结合在一起，在一定程度上逐渐分散了传统上由民族国家所掌控的金融、信息等重要的权力。科技企业本身甚至就可以通过人工智能的推荐算法为政府和用户塑造“信息茧房”和“回音壁”，从而带来严重的社会问题。同时，在人工智能技术的加持下，经济生活数字化的水平迅速提升，在为用户带来便捷的同时也埋下了重大安全隐患。比如说，在这次乌克兰危机期间，伴随着美国发起的金融制裁，PayPal 也宣布封锁俄罗斯的个人账户。[②] 这意味着科技平台企业甚至可以拿捏一个主权国家的重要经济活动。这种经济安全风险也是未来人工智能加持的国际关系中不容忽视的一个方面。

① “Faked Pelosi Videos, Slowed to Make Her Appear Drunk, Spread Across Social Media,” *The Washington Post,* May 24, 2019.

② 徐鸿波：“PayPal将于18日封锁所有俄罗斯人的电子账户”, https://world.huanqiu.com/article/47AZbJHbBTK。（上网时间：2022年8月5日）

第四，人工智能“算法黑箱”带来了“偏见”“歧视”，以及过度收集公众隐私数据等一系列的社会安全风险。微软 2016 年曾经推出过一个人工智能聊天机器人 Tay，这个机器人通过在 Twitter 上与用户聊天来训练自己的评论能力。但是仅仅一天时间，由于 Twitter 用户使用了不当言论的数据对其“训练”，Tay 已经成为了一个“种族主义者”，导致微软不得不在上线几小时后就将其关停。[①] 而随着人工智能在多个领域中应用的扩展，数据集的安全性也与社会公平和生命财产安全挂钩。

此外，人工智能需要大量数据集作为资源训练机器学习算法，给数据安全带来了风险。人工智能的“数据饥渴”与人类的“隐私警惕”之间存在着不可避免的张力，人工智能技术对数据的“过度采集”是否会危害到数据和隐私安全，也成为社会公众普遍担心的问题。在微软的一项调查中，41% 的受访者表示不信任智能语音助理，并认为其通过实时的被动声音收集损害隐私；约 52% 的人表示他们担心自己的个人信息不安全。[②]

第五，人工智能的潜在利益和发展前景引发了大国之间的竞争，加剧了国际安全的不稳定性。各国不仅加大了在技术标准、国际规则等方面的竞争，同时还为争夺人工智能发展的主导权强化了在芯片、数据等方面的“科技脱钩”。以当前最为激烈的中美人工智能竞争为例，美国出台的《2022 年美国竞争法案》等一系列文件“以竞争之名，行霸权之实”，[③] 试图在人工智能的整个生态链条中排斥和防备中国：美国政府不仅通过出口管制、投资审查等方式插手中美企业之间的互动与合作，而且通过人工智能防务伙伴关系（AI-PfD）、人工智能全球合作伙伴关系（GPAI）等“小圈子”推进排挤中国的人工智能规则，甚至不惜以政治手段插手留学生和科学家的正常交往。美国这种针对性竞争塑

① Elle Hunt, “Tay, Microsoft’s AI Chatbot, Gets a Crash Course in Racism from Twitter,” https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.（上网时间：2022年7月26日）

② Microsoft Advertising, “The 2019 Voice Report,” https://about.ads.microsoft.com/en-us/insights/2019-voice-report.（上网时间：2022年7月26日）

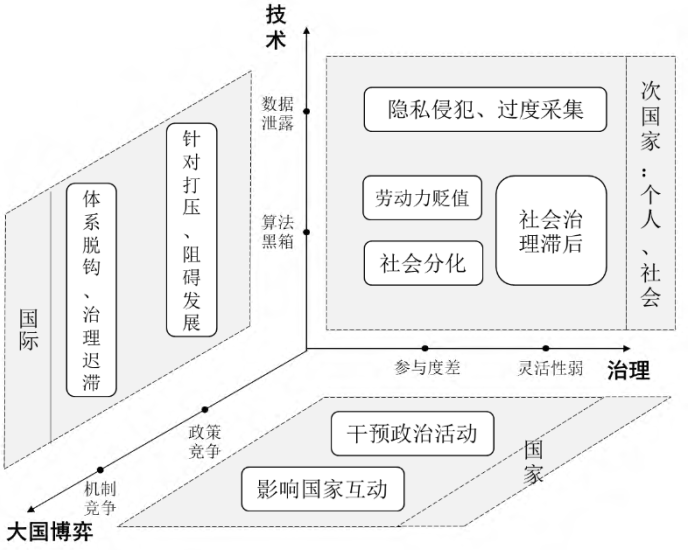
③ “‘2022年美国竞争法案’以竞争之名行霸权之实”, http://www.news.cn/sikepro/20220524/9c99d767fb9c4153a3685ae3833b0680/c.html。（上网时间：2022年8月5日）

造的人工智能攻势不仅可能扰乱中国的人工智能发展步伐，带来直接的技术风险，还会造成经济损失，甚至引发关乎公民人身安全的隐患。而更严重的问题是，美国这种行为会在人工智能版图中塑造对中国的地缘封锁和价值观对立，将会引发全球人工智能技术、产业与标准生态体系的分裂甚至是对抗。

二、人工智能安全问题的成因

人工智能所带来的安全风险具有全局性、广泛性和战略性的特点，既普遍存在于军事、政治、经济等不同的领域，同时又对国际、国家、社会、个人等不同的层次都产生了影响。导致这样复杂的人工智能安全风险体系产生的原因，可以归结为三个方面：一是人工智能的内生安全，即技术本身的安全缺陷；二是治理能力的缺失，即所谓的“科林格里奇”陷阱；三是大国博弈引发的国际治理机制缺失。（见图 1）

图 1　人工智能安全风险及成因示意图



资料来源：作者根据相关资料整理。

（一）人工智能内生安全风险

与其他领域相比，人工智能技术自身存在的安全缺陷是产生很多安全问题的根本来源，主要包括算法风险和数据安全风险两个维度。

人工智能算法风险具体指算法的设计或实施有误，可能产生与预期不符甚至伤害性结果。其中最为典型的“算法黑箱”问题，是指由于人工智能算法的复杂性而导致存在技术门槛，使得一些主动设置或被动生成的“算法黑箱”横亘于设计者和用户之间。这一方面导致监督审查缺失，可能会产生严重的安全后果；另一方面也给设计者带来了后果不可预测的挑战，由机器学习所产生的后果往往超出了设计者开发算法的初衷。这引起了人们对人类能动性和自主性影响的担忧。[①] 此外，“算法黑箱”还牵涉伦理和社会规范问题。如果不能将算法决策合理有效地嵌入到社会规范和法律道德中，将会给现代社会的稳定带来冲击。

数据安全风险主要是有偏见的数据集和决策规则导致人工智能的训练结果存在偏误，并进一步影响系统决策，扩大偏见甚至使偏见永久化。[②] 麻省理工大学研究员与微软科学家曾对微软、IBM 和旷世科技三家公司的人脸识别系统进行测试，发现其针对白人男性的错误率低于 1%，而针对黑人女性的错误率则高达 21% ~ 35%。[③] 造成这一结果的主要原因，是互联网上白人男性和黑人女性在数据样本的数量和质量上存在较大的差距。还有人会利用数据来进行“数据投毒”，恶意地将机器学习所使用的数据集中混入伪装样本，从而导致人工智能输出有偏差的结果。

（二）治理滞后引发的“科林格里奇困境”

“科林格里奇困境”（Collingridge dilemma）由英国学者大卫·科林格

① Mireille Hildebrandt, “The New Imbroglio – Living with Machine Algorithms,” https://doi.org/10.25969/mediarep/13395.（上网时间：2022年8月6日）

② Klein Aaron, “Reducing Bias in AI-based Financial Services,” https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/.（上网时间：2022年8月6日）

③ 中国信息通信研究院安全研究所：《人工智能安全白皮书 2018》，http://www.caict.ac.cn/kxyj/qwfb/bps/201809/t20180918_185339.htm。（上网时间：2022年8月5日）

里奇提出，本质上是要解决技术控制的困境。科林格里奇注意到技术发展与社会认知并不一致，甚至相互背离，当技术危害严重到广为人知之时再想要去控制技术，则变得异常困难且代价高昂。人工智能的安全风险可能伴随着其研发过程而产生，并且当技术产生不良后果时，它往往已经成为整个经济和社会结构中难以抽离的一部分，以至于难以对它进行控制。①

“科林格里奇困境”所指出的治理滞后也广泛存在于人工智能领域。2021 年 Twitter 识别出其推送算法会在无意中扩散右翼团体的言论，而通过识别后才发现，这最早可以追溯到 2020 年 4 月，并已经造成了传播性的影响。② 导致这种情况的主要原因在于技术发展超出了以往治理的经验，由于人工智能极大的影响力，这种滞后往往会带来难以承受的后果。

产生“科林格里奇困境”的一个主要原因在于，政府和社会习惯于用“管理”的旧思维来应对人工智能新安全风险。人工智能技术作为新兴技术，具有极强的不确定性，而其独有的新特征——“涌现性”和“自主性”③ 又进一步加强了这种不确定性，这就意味着政府通过出台明确条款而进行的“强制性”治理手段往往难以奏效。人工智能安全衍生的范围非常广泛，涉及多重利益相关的主体，民用人工智能技术远远走在了军事人工智能技术之前，企业在技术、人才方面所掌握的资源超出了政府。这就决定了政府要想对人工智能安全风险进行有效的识别和治理，就需要与企业乃至社会个体进行更加充分的沟通和合作，建立包容、开放和多方参与的治理模式。

① D. Collingridge , *The Social Control of Technology,* Milton Keynes, UK: Open University Press, 1980, pp.16-17.转引自文成伟、汪姿君：“预知性技术伦理消解AI科林格里奇困境的路径分析”，《自然辩证法通讯》，第43卷，2021年第4期，第10页。

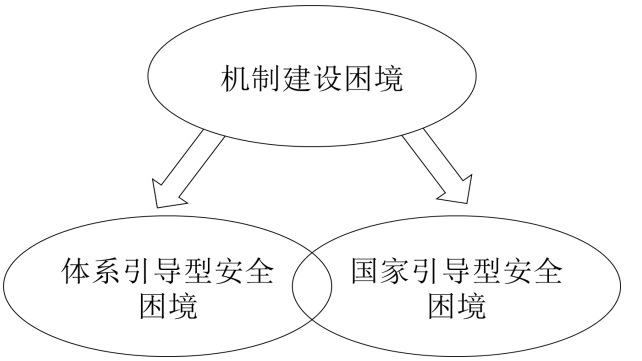
② Katyanna Quach, “Twitter’s Algorithms Favour Right-Wing Political Content,” https://www.theregister.com/2021/10/25/twitters_political_bias/.（上网时间：2022年8月6日）

③ “涌现性”具体指算法底层的简单规则能够生成的复杂行为，智能并不由预定的算法前提所决定，因此，仅从前期特征开展的算法治理可能会失效；“自主性”则指当下人工智能深度学习算法可以从海量无标注的大数据中自我学习、自我进化，未来技术可能会突破工具属性，具有自主意识。参阅刘劲杨：“人工智能算法的复杂性特质及伦理挑战”，《光明日报》，2017年9月4日。

（三）大国竞争引发的国际治理规则缺失

人工智能的发展和应用具有非常强烈的国际化属性，持续涌现的安全问题也需要国际社会进一步扩大在国际规则方面的共识。大国之间围绕争夺人工智能发展主导权所展开的博弈，阻碍了国际社会在人工智能国际规则方面的努力，形成了相互关联的三重安全困境。（见图 2）

图 2 人工智能的多重安全困境



资料来源：作者根据相关资料整理。

一是人工智能技术本身的安全复杂性导致国家间信任不足，从而引发“超规格”的治理手段和“高警惕”的安全互动，并进一步阻碍全球范围内形成成熟有效的人工智能治理模式。这一逻辑实际上是传统国际关系在新兴技术领域的映射：由国际无政府状态所引发的“体系引导型安全困境”（system-induced security dilemma）经过人工智能技术安全的复杂性被进一步扩大，形成了更加难以破解的安全困境。[①] 比如，在人工智能的数据要素上，各国都保持着较高的敏感度，难以达成统一的标准，推动数据安全、便捷地跨境流

① Alan Collins, *The Security Dilemmas of Southeast Asia,* London:Macmillan Press, 2000, p.21.转引自尹树强：“‘安全困境’概念辨析”，《现代国际关系》，2003年第1期，第57 ~ 61页。

动；在人工智能伦理标准上，则由于东西方政治文化和历史认同上的不一致，彼此都对对方提出的技术标准报以怀疑和审慎的态度。

这一类安全困境从根源上讲不是由国家政策驱动的，而是根植于客观复杂的技术环境。一个典型的案例就是尽管美欧在战略方向和技术政策上都保持着较高的默契度和一致性，但是仍然在“数字税”、技术标准等方面产生了结构性的合作障碍。这一重安全困境说表明，如果不从人工智能的技术特征上看待人工智能治理，就难以破解由技术特征驱动的安全困境。

二是霸权国的针对性竞争战略塑造了人工智能领域的“不相容”状态，以政治性目的绑架了技术的发展与合作，阻碍了人工智能全球治理的操作空间。如果说上一重困境的来源是客观的，这一重安全困境则更多归咎于国家政策的主观驱动，基本符合了柯林斯和施耐德提出的“国家引导型安全困境”（state-induced security dilemma）的逻辑链条。[①] 这一重安全困境是横亘在中美两个技术大国之间的主要矛盾。在人工智能领域，美国通过反复强调“中国技术威胁论”“中国技术偷窃论”“数字威权主义”“中国企业不正当竞争”等话语，毫不掩饰地将中国塑造成为损害其安全的行为主体，并进一步采取了脱钩、断供、技术审查、建设排他性机制等更加咄咄逼人的泛安全化手段进行抵制和打压。

这种强势进攻型的人工智能政策不仅无法给美国提供真正的安全——美中双方都会以最坏的情势揣摩对方的意图并采取极具冲突性的政策；而且会进一步妨害全球范围内人工智能治理机制的形成，并把其他国家也卷入到针对性竞争的围场内，生成了一系列“集团化”和“排他性”的人工智能治理机制。这一方面妨害了技术要素的流动，降低了技术研发的效率；另一方面也提高了人工智能技术的风险性和治理成本。类似的情况也出现在美欧之间，虽然美欧之间的冲突远没有中美之间的激烈，但是欧盟在提升自主性、规范美国人工智能相关企业方面也是动作频频。2020 年以来，欧盟先后提出了技术主权、数字主

① Alan Collins, *The Security Dilemmas of Southeast Asia*, London:Macmillan Press, 2000, p.21. 转引自尹树强：“‘安全困境’概念辨析”，《现代国际关系》，2003年第1期，第57 ~ 61页。

权、数据主权三个为提升欧盟在人工智能等新兴技术领域独立自主程度的发展权，同时也出台了《通用数据条例》《数字市场法案》等，意在强化对美国互联网企业的监管。

三是人工智能全球性治理机制的缺失削弱了安全困境缓解的可能性，甚至进一步深化了上面的两重困境。在安全困境理论中，一个经典的解释是，如果国家能够对对方的意图和行为模式有充分的了解，就会避免悲剧性的战略误判和困境升级。从国际关系的现实来看，国际机制的存在不仅可以通过共同承认的规则形成对国家行为的约束，更能够提供一个稳定而长期的信息交流平台，有效阻碍安全困境的形成和升级。①

而不幸的是，在当前的人工智能治理实践中，虽然初步形成了一些行之有效的治理机制，但这些机制大多数还是落入了网络空间治理中“观念一致国家联盟”（Like Minded Countries）的窠臼。② 比如说现在推进较快的美日印澳“四边机制”关键和新兴技术工作组（The Quad Critical and Emorging Technology Working Group）、美欧贸易和技术委员会（TTC）等机制，都是以美国为主导、美国传统盟友和技术伙伴参与、以意识形态为纽带的排他性和竞争性的技术联盟。而全球性的人工智能机制却进展缓慢，亟需建设凝聚中美共识的人工智能国际机制。这些排他性技术联盟的形成，非但无法缓解因中美战略竞争而引发的“国家引导型安全困境”，无法为两国之间的信息交流、技术合作、政策互谅提供平台，甚至会进一步加深人工智能国际治理的复杂性和困难度，割裂了原本就复杂和高风险的技术生态，进一步加深了“体系引导型安全困境”的治理难度。但反过来说，如果能在全球范围内建立起普遍有效的国际治理机制，则是拆解上述人工智能多重安全困境的治本之策。

① Robert Jervis, “The Security Regimes,” *International Organization,* 1982, p. 178.

② 鲁传颖、约翰·马勒里：“体制复合体理论视角下的人工智能全球治理进程”，《国际观察》，2018年第4期，第67 ~ 83页。

三、人工智能安全治理机制

作为新安全议题，人工智能的一些新特征应该得到更多的重视，这也意味着不仅应该走出传统治理模式的思维定势进行人工智能安全治理，同时还应该对人工智能的一些有别于其他新技术的特征，采取更加具有针对性的手段进行治理。

根据人工智能的风险表现和治理困境不难总结出，一个理想的人工智能治理机制应该具备以下特征：第一，既要有技术专业力量的支持以解决人工智能内生的技术安全风险，又要有政府的宏观把控以化解多领域扩散的衍生安全风险；第二，需要重视治理的时效性、灵活性和稳定性，以适应人工智能技术与发展应用的新特征；第三，国际治理机制应该保证成员的“非排他”和规范的“非歧视”，合理管控人工智能的主客观安全困境。

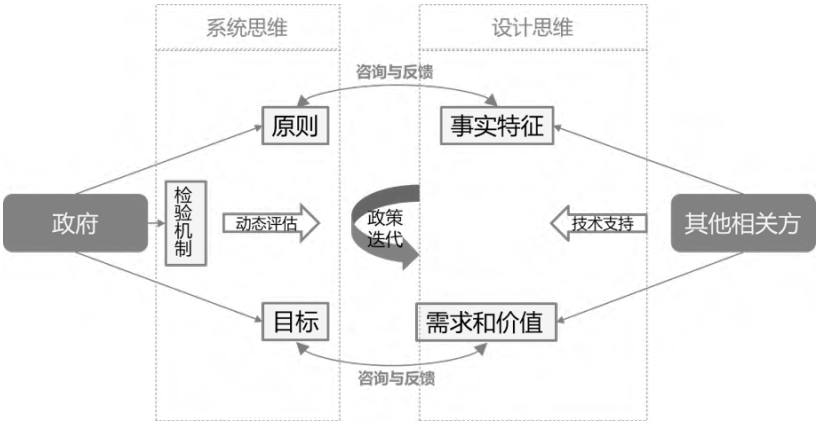
在此基础上，2018 年《世界经济论坛白皮书》在多国的联合倡议下，针对人工智能的治理需求提出的“敏捷治理”新概念，逐渐在各国的人工智能治理实践中得到有效运用。在实践中，敏捷治理催生了一系列人工智能的国内治理机制，并有望在共识的基础上进一步推动人工智能的国际治理机制建设。

（一）敏捷治理的理论内涵

敏捷治理模式的重要特点就是通过政府的系统性整合，与人工智能技术的各利益攸关方共同组成一体化的人工智能治理生态，并通过灵敏、及时、持续的“咨询—反馈”机制，促进治理政策的迭代升级，以弥补政府治理中信息的滞后性，形成对人工智能风险的前瞻性评估与治理。[①] 在整个敏捷治理的思路中，包括三个要素——两条“咨询—反馈”路径和一个“动态评估”机制，一起致力于治理政策更新。（见图 3）

① 薛澜、赵静：“走向敏捷治理：新兴产业发展与监管模式探究”，《中国行政管理》，2019 年第 8 期，第 28 ~ 34 页。

图 3　敏捷治理理论的作用机制



资料来源：作者根据相关资料整理。①

其一，第一条“咨询—反馈”路径。建立在政府所规划的人工智能“治理原则”与各利益攸关方掌握的人工智能技术“事实特征”之间。在敏捷治理模式下，政府并不需要制定明确严格的人工智能治理规范，而是需要设定一种原则框架，并通过这条“咨询—反馈”路径与掌握市场和技术信息的多利益攸关方进行配合，保证治理原则设置的科学性。

其二，第二条“咨询—反馈”路径。建立在政府所规划的人工智能“治理目标”与各利益攸关方掌握的“市场需求信息、价值观与公正性”之间。与治理原则类似，政府在敏捷治理模式下所需要设计的第二个内容就是人工智能的治理目标，以规划整个治理活动的道路和方向。但是这条道路要想行稳致远，就需要满足各利益攸关方切身的市场需求和精神价值需求。

其三，动态的评估机制。在上述两条路径之外，政府还需要设置一套动态的评估机制，用于评估人工智能技术活动前是否符合原则、技术活动后是否满

① “Agile Governance Reimagining Policy-Making in the Fourth Industrial Revolution,” https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf.（上网时间：2022年8月6日）

足目标并联通整个过程。此外，敏捷治理模式还通过企业设计试点技术与政府审查机构的动态评估相结合，塑造出一种政策的迭代机制，从而在兼顾战略目标导向的同时回应技术与环境的变化。这为人工智能技术本身的“不确定性”风险应对提供了有力指引。

从上述敏捷治理的特征来看，它基本对应了本文前两条所提出的治理要求，已经能为国家应对人工智能的多重安全风险、推动人工智能技术健康向善发展提供对策良方，因此已经在多国的人工智能治理机制建设中得到了有效的应用。

（二）敏捷治理的多国实践

由于敏捷治理的思路贴合了人工智能的治理需求，目前已经在多国得到了不同程度的实践。

在中国，国家新一代人工智能治理专业委员会于 2019 年发布了《新一代人工智能治理原则——发展负责任的人工智能》[①] 这一框架性的战略文件。根据该文件，敏捷治理与其他七项治理原则一起成为指导中国人工智能治理的重要标度。而在治理主体层面，中国通过成立新一代人工智能治理专业委员会，吸纳来自高校、科研院所和企业的相关专家共同参与治理。目前，中国的人工智能治理在治理标准和多方整合上已经取得了巨大的进展，但政府在如何进行动态监管和角色调整等方面仍处于进行时，尚未形成敏捷治理的完整生态。

美国则不仅落实了敏捷治理的思路，而且将其与美国的联邦体系进行创造性结合，并在此基础上进一步落实人工智能治理机制创新。2020 年《美国人工智能计划法案》（NAII）的出台拉开了美国人工智能机制建设的序幕。在 NAII 的整体机制中，白宫和国会是主要的设计者，通过人工智能计划法案确定了人工智能的治理原则：优先考虑研发，加强基础设施建设，推进技术标准，培养人工智能人才，促进国际参与合作，重视国家安全建设，等等。与此同时，法案也设计了人工智能的发展目标：维持领先地位，值得信赖的人工智能技术，

① 中华人民共和国科学技术部：《发展负责任的人工智能：新一代人工智能治理原则发布》，https://www.most.gov.cn/kjbgz/201906/t20190617_147107.html。（上网时间：2022 年 7 月 25 日）

面向未来的人才培养，整合人工智能系统，协调开发应用。[①] 这些原则和目标符合敏捷治理的思路，只是一些比较宽泛的设计和规则，并不对人工智能的治理内容进行明确严格的限制。

NAII 在其他非政府主体的“设计思维”方面则在敏捷治理思路上进行了创新。首先将分散的各攸关方进行整合：NAII 设置了国家人工智能咨询委员会（NAIAC）和国家人工智能研究资源特别工作组（NAIIRRTF），这两个是多利益攸关方深度参与的机构。在国家人工智能咨询委员会中，各方主要根据其在市场前沿所掌握的事实特征和市场需求提供人工智能治理和发展的相关建议，是整个 NAII 系统中最重要的咨询机构；国家人工智能咨询委员会则进一步平衡了各方与政府主体之间的地位，通过平级的机构设置，让学界、产业界、非政府机构进一步参与到决策咨询中，使敏捷治理的两条“咨询—反馈”路径更加稳健有效。此外，国家人工智能研究资源特别工作组还致力于为人工智能的长远发展设置一个公共基础设施平台——各方主要提供技术的前沿发展特征及数据等技术支持，并通过多轮会议推进人工智能特别研究资源的设置规则。人工智能特别研究资源作为公共基础设施的设置将为整个 NAII 系统提供养分，而这也是敏捷治理进一步与人工智能技术特征进行结合的一个创新表现。

NAII 依据敏捷治理的思路，建立了实时、高效的审查机制。在 NAII 的框架中，总体上设置了一个三层审查机制：自我审查、人工智能计划办公室的机制内审查以及人工智能专责委员会的审查。这一系列审查机制既兼顾了联邦传统的分权与制衡的审查原则，充分思考治理方案是否满足白宫和国会设置的从原则到目标的宏观框架，又通过 NAII 系统内的自我和平级审查保证时效性，是敏捷治理思路与联邦设计思路结合的一个特色方面。

这一特色还体现在人工智能计划办公室（NAIIO）的职能设计中，这一白宫下设的专业人工智能职能部门，具备了较强的行政特色，并兼顾机构之间的

① 116th US Congress, “H.R.6216 - National Artificial Intelligence Initiative Act of 2020,” https://www. congress.gov/bill/116th-congress/house-bill/6216?__cf_chl_tk=aq2aBo0eMPPRwKVkU0EWnlBuhMiHEul5V90n1R.mpY8-1653477655-0-gaNycGzNBz0.（上网时间：2022年7月25日）

协调。更重要的是，人工智能计划办公室代表了政府的高度渗透，因为它可以直接对多利益攸关方提供指导和规范。通过这一功能的设计，NAII 机制进一步巩固了敏捷治理中所要求的政府核心地位。

除了整体设计之外，敏捷治理的三个要素——多方参与性、时间敏感性和反馈迭代性，也被运用于美国新兴技术相关机制的创新设计中。比如说，就多方参与而言，在网络和信息技术研究与开发计划（NITRD）项目的设计中和在国家科学基金会（NSF）的科技中心建设中，都进一步地将企业、研究院所、高校等纳入其中。而国防高级研究计划局（DARPA）针对新兴技术的竞赛和项目奖励机遇如“AIE 项目”等，则将时间敏感度进一步提升，并以此设计整个机制的运行模式。因此，美国人工智能治理机制的创新，实际是在整体上对敏捷治理思路的改造运用，并根据其核心特征推动机制创新。

而在中美两国以外，敏捷治理在新加坡、加拿大、英国等国家和欧盟也有了不同程度的实践，呈现出“遍地开花”的发展姿态，为进一步建设全球性的人工智能治理机制提供了良好的契机。但是，也要注意到敏捷治理在很大程度上是对传统政府管理模式的颠覆，无论是理念还是实践层面都存在着流于形式的风险。例如，虽然建立了相应的多利益攸关方咨询机构，但却不真正授予机构权限，最后还是无法发挥咨询机构的作用。

（三）推动敏捷治理的国际机制建设

如前所述，除了人工智能技术自身的复杂性和安全风险所带来的治理困境以外，技术的跨国性特征和大国战略竞争的态势也塑造了多重的安全困境，并随着公平性机制的缺失而进一步加深了人工智能安全发展的沟壑。因此，需要找到敏捷治理基础上的国际机制建设途径，以助推全球人工智能治理进程。对此，本文在敏捷治理环节的基础上提出以下三条建议。

第一，以共识性原则和非对抗性目标的系统思维构建全球人工智能治理机制。首先，人工智能治理需要技术原则作为指引，而这也是目前全球治理进程推进较快的一个环节。2019 年，经济合作与发展组织（OECD）通过了《人工

智能原则》，提出了包容、可持续、以人为本、公平透明、可解释等人工智能发展原则。[①] 这也是全球第一个包含中国和美国的跨政府间人工智能治理原则，指导了 OECD 人工智能政策观察站的搭建，以促进治理原则的落实。在这一良好开端下，世界各国应该进一步协商细化人工智能的治理原则，形成共识性的规范约束，以避免安全困境的产生。此外，在各国人工智能战略设置中，也应避免出现超越普通竞争边界的对抗性目标，并配合以适当的解释，避免形成战略误判，从而导致不必要的安全升级。

第二，积极吸纳非政府主体参与人工智能全球治理。充分发挥专家团体、技术社群、跨国企业所具备的技术能力和掌握的市场信息，让各利益攸关方的设计思维在人工智能全球治理中发挥作用。比如说，微软等人工智能前沿企业针对其所掌握的技术前沿特征信息，为人工智能发展的方向和红线设计不断提出更新建议，微软首席执行官萨蒂亚·纳德拉就在阿西莫夫和霍金的理论基础上进一步细化了人工智能开发的十条原则。[②] 斯坦福大学的人工智能“百年项目”也在出台人工智能全球发展的年度报告，为治理活动提供可信的参考资料。[③] 同时，谷歌等科技企业不仅在企业内部出台了人工智能原则，组建了相关的自我审查委员会，还提供了技术工具和管理工具，助力破解人工智能全球治理中的技术客观性问题。因此，从全球人工智能治理的维度看，多利益攸关方主体的参与是至关重要的。

第三，通过公共资源建设助推全球人工智能治理的敏捷高效。根据人工智能的发展特征，各国都着手对本国的人工智能资源进行整合，以推动更高效的人工智能发展，并在发展中谋求安全。事实上，在人工智能的特定领域中，还

① “The OECD Artificial Intelligence (AI) Principles,” 2019, https://oecd.ai/en/ai-principles.（上网时间：2022年7月25日）

② Matt Clinch , Turak Natasha, “Microsoft CEO Satya Nadella on the Rise of A.I.: ‘The Future We Will Invent Is a Choice We Make’,” https://www.cnbc.com/2018/05/24/microsoft-ceo-satya-nadella-on-the-rise-of-a-i-the-future-we-will-invent-is-a-choice-we-make.html.（上网时间：2022年7月25日）

③ “One Hundred Year Study on Artificial Intelligence (AI100),” https://ai100.stanford.edu/.（上网时间：2022年7月25日）

需要更为广泛的技术资源，比如说数据和算力资源等。因此，如果各国能够协力搭建资源公共平台，在合理的开放原则和脱敏程序后进一步共享技术资源，则不仅能够在研发环节中减少成本和风险，还能够进一步发挥人工智能技术在监测、预测、应急响应等方面的应用价值，甚至可以进一步推动各国之间的良性相互依赖，减弱安全困境的烈度。

（责任编辑：黄丽梅）

Keywords: emerging technologies, forward-looking governance, national security

AI-Related Security Risks and Governance Models

Lu Chuanying and Zhang Luyao

Abstract: Artificial Intelligence (AI) technologies face new opportunities in the 21st century, but security risks that continue to emerge have become important factors retarding their development. To improve AI security governance, it is not only necessary to identify where the security challenges exist, but also to systematically understand the causes of these risks and take resolving governance dilemmas as the core focus. Given the context of overlapping AI-related security risks and governance dilemmas, agile governance, characterized by multi-stakeholder participation, time sensitivity, and iterative feedback, has gradually been taken up by many countries. This governance approach, different from traditional ones, provides an effective solution to the security challenges caused by AI-related uncertainties; offers an effective way to ameliorate governance capabilities; and affords opportunities for building a universal, fair, and reasonable mechanism for global AI governance. In sum, it is a good remedy against AI-related security challenges.

Keywords: AI security, agile governance, China-US technological decoupling, global technology governance

National Security Exception Clause Governing Cross-Border Data Flows: Check and Balance, Boundaries, and Construction

Wei Qiuyue and Hong Yanqing

Abstract: Cross-border data flows are an important link and factor in the development of digital free trade. Regulating such flows has received close attention in recent digital trade agreements. At the global level, cross-border data flows are faced with a theoretical dispute between data sovereignty and data freedom, which centers on the potential impacts cross-border data flows could have on a country's national security. Such a dispute is reflected in the attitudes of different countries toward the national security exception clause during the negotiations of international agreements. The dispute on whether the security exception clause should be treated as self-judging matters has given birth to three legislative models: the Comprehensive and Progressive Agreement for Trans-Pacific Partnership negotiations represented by the United States that do not advocate a security exception clause; the digital trade negotiations represented by the Regional Comprehensive Economic Partnership that propose limitations on the self-judging nature of the national security exception clause; and suggestions by other countries that the traditional form of WTO security exception clause should be maintained. With a view to safeguarding national security and promoting digital free trade, China should adopt a cautious attitude toward the self-judging nature of the national security exception while maintaining the existence of such a clause, and promote a paradigm shift in a gradual manner.

Keywords: national security exceptional clause, self-judging clause, cross-border data flow, good faith principle