►►►求解时代之问②

当算法参与决策，安全如何保障？

本报记者　**孙亚慧**

确保人工智能始终服务于人，而不是替代人的最终判断

记者：当前算法参与社会决策的广度和深度，已经发展到了一个什么样的阶段？它带来的安全风险，是不是已经到了必须系统性应对的临界点？

肖茜（清华大学人工智能国际治理研究院副院长）：随着大模型能力的快速提升，人工智能确实正在逐步从传统意义上的“辅助工具”走向“参与决策”。今天，无论是在医疗诊断、金融风控、交通调度，还是公共治理、司法辅助等领域，算法都越来越多地参与到决策过程中。但总体来看，人工智能成为完全意义上的“决策者”仍然只是少数特定场景，并非社会普遍现象。更准确的表述应该是人工智能正在成为越来越重要的“决策参与者”和“决策影响者”。

从人工智能安全治理的角度来看，确保“人在环上”（Human-in-the-Loop）仍然是一项不可突破的基本原则。人工智能虽然在数据处理和模式识别方面具有明显优势，但当前技术仍然存在诸多局限，例如“黑箱效应”导致决策过程缺乏透明性和可解释性，训练数据可能存在偏差，模型可能产生“幻觉”，在复杂环境下也缺乏人类所具备的价值判断、伦理考量和情境理解能力。因此，如果将涉及生命安全、公共利益或重大社会影响的决策完全交由人工智能作出，不仅可能导致错误决策，也可能出现责任主体不清、问责机制缺失等治理难题。

更重要的是，我们需要关注算法决策不断扩张所带来的系统性风险。当越来越多的社会运行依赖算法时，一旦算法出现系统性偏差或遭受攻击，其影响可能迅速扩散至金融、医疗、交通、能源等多个关键领域，甚至影响社会公平和公共信任。因此，我认为，我们已经到了必须系统性应对人工智能决策风险的重要阶段。未来的治理重点，不应是简单限制人工智能的应用，而是在推动技术创新的同时，建立更加完善的风险评估、算法审计和责任追溯机制，确保人工智能始终服务于人，而不是替代人的最终判断。

魏凯（中国信息通信研究院人工智能研究所所长）：算法伴随着计算机软件的普及，早已经深度嵌入社会运行的各个环节，一切计算机程序都建立在算法之上。早期程序依托的算法，完全由人工预设的明确规则驱动。随着神经网络等人工智能技术成熟，算法的形成逻辑发生重要转变：它不再单纯依靠人工设定规则，模型训练数据同样深刻影响最终输出。

十余年间，消费互联网率先完成迭代，信息搜索、商品推荐、广告分发等场景全面运用 AI 算法。

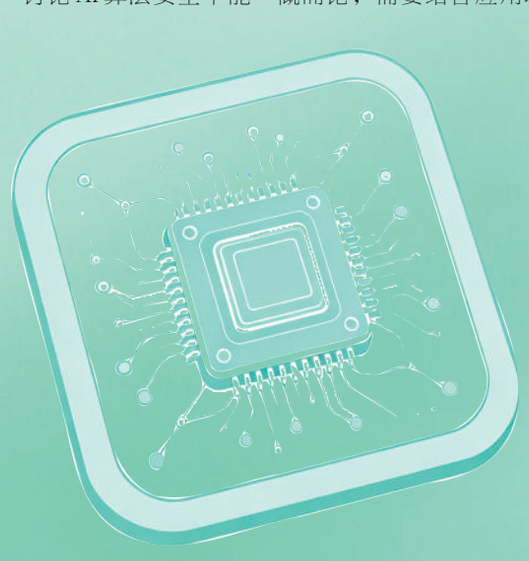
当下新趋势显现：大众信息查询行为逐步转向智能体，依靠大模型整合素材、直接生成应答。需要警惕的是，大模型“幻觉”问题难以彻底消除，使用者应当对智能体输出内容做好真伪辨别。当然，家里的扫地机器人，自动驾驶汽车，也都是AI算法驱动的。总体而言，算法参与社会决策呈现不均衡发展态势：面向普通公众的消费场景渗透更广，面向经营主体的产业场景仍处在深化探索阶段。

确保AI进入决策过程之后，不会削弱决策的可靠性、公正性、可控性和可问责性

记者：在“算法参与决策”这个特定语境下，您认为“安全”的核心要义是什么？它和我们传统说的网络安全、数据安全、信息安全，最本质的区别在哪里？是不是可以理解为，这是一种更深层的“决策安全”或者“规则安全”？

魏凯：安全本身具有相对性，AI算法安全同样不存在绝对安全。放到“算法参与决策”这一特定语境下，我认为安全的核心要义，就是保障AI算法具备充足的透明性与运行可靠性，规避由算法自主推演、模型决策带来的重大风险。

讨论AI算法安全不能一概而论，需要结合应用场景、面向的不同主体分类考量。面向普通消费者提供服务的AI算法，应当匹配大众认知水平，搭建相适配的安全防护机制，守住用户合法权益底线。具体涵盖内容合规、防范算法歧视、强化个人隐私保护、警惕AI算法不当引导公众认知等多个维度；针对未成年人等特殊群体，还要实施更加严格、差异化的算法约束。



维护AI算法安全，不能只要求一方出力，技术服务商与使用者双方都负有责任。对于技术服务商而言，应当严格遵循法律法规与相关技术标准，落实算法安全各项合规要求，从技术源头筑牢风险防线。与此同时，也需要持续呼吁广大使用者不断提升风险辨别与自我保护意识。

肖茜：在算法参与决策的特定语境下，我认为“安全”的核心要义，是确保AI进入决策过程之后，不会削弱决策的可靠性、公正性、可控性和可问责性。它所保护的不只是一个技术系统能否正常运行，更是社会中的决策权能否被正当、审慎地行使。

传统意义上的网络安全、数据安全和信息安全主要解决的是决策所依赖的技术基础和信息基础是否安全。算法决策安全则进一步追问：即使网络没有遭到攻击、数据没有泄露、信息也是真实的，算法作出的判断是否一定合理？如果出现错误，最终由谁承担责任？

因此，我认为可以把它理解为一种更深层次的“决策安全”，同时也涉及“规则安全”。所谓决策安全，是要防止算法误判、自动化偏见和人类过度依赖机器，确保重大决策仍然处于有效的人类监督和控制之下。所谓规则安全，是要审视被写入算法的目标、指标和价值排序是否合理。因为算法并不是完全中立的，它会把开发者设定的规则转化为大规模、持续性的社会决策。

记者：生成式AI爆发之后，大模型参与决策的场景变多了。和传统的、针对特定任务设计的算法相比，大模型的“通用决策能力”更强，但不确定性也更大，这是不是给算法安全保障带来了全新的、更大的挑战？

肖茜：大模型时代的算法安全已经进入了一个新的阶段，它面临的挑战与传统算法有着本质区别。传统算法通常针对某一特定任务设计，功能相对单一，应用边界比较明确，因此风险也相对可预测。而大模型则具有通用能力，可以理解、推理、规划、生成内容，并在逐步向智能体演进，开始参与越来越复杂的决策过程。与此同时，大模型还具有持续学习、不断迭代、多场景应用等特点，能力边界始终处于变化之中。这意味着，技术的发展速度远快于治理规则的更新速度，也使算法安全治理面临前所未有的挑战。

大模型时代与传统算法时代最大的不同，至少体现在三个方面。一是风险更加难以预测。大模型具有一定的涌现能力，一些能力并非开发者预先设计，我们很难像传统软件那样准确预测系统在所有场景下的行为。二是算法开始影响决策，而不仅仅是执行任务。三是风险的外溢性更强。一个通用大模型往往服务数亿用户，可以同时应用于多个领域，一旦出现系统性缺陷、偏见或被恶意利用，其影响可能迅速跨行业、跨地区甚至跨国界传播，很可能造成系统性风险。

敏捷治理更强调边发展、边治理、边完善规则，提高治理的适应性和前瞻性

记者：全球范围内，算法治理有不同的思路，中国的算法安全治理路径，有哪些自己的特点和制度优势？我们在平衡“发展”和“安全”这对关系上，有没有走出一条自己的路？

肖茜：我认为，中国人工智能治理的一大特点，就是始终坚持发展与安全并重，努力在促进创新和防范风险之间寻找平衡，而不是把二者对立起来。人工智能是推动经济社会发展的重要引擎。如果因为担心风险而过度监管，可能会抑制创新；但如果只强调发展而忽视安全，又可能带来新的社会风险。因此，中国提出统筹发展和安全，把安全作为发展的保障，把发展作为安全的基础，在鼓励技术创新的同时，加强风险治理，努力实现二者的动态平衡。

在治理方式上，中国近年来提出了风险导向、敏捷治理的理念。今年发布的《2026世界人工智能大会暨人工智能全球治理高级别会议主席声明》再次强调，坚持“风险导向、敏捷治理”。这意味着，面对快速演进的人工智能技术，治理模式不能一成不变，而应根据技术发展和风险变化不断调整，实现监管与创新同步演进。相比于“先发展、后治理”或“先立法、后创新”的传统路径，敏捷治理更强调边发展、边治理、边完善规则，提高治理的适应性和前瞻性。

此外，中国还高度重视人工智能治理的国际合作。人工智能的发展和应用具有跨境性，其风险也具有全球外溢性，任何国家都无法单独应对。因此，中国不仅积极完善国内治理体系，也积极参与联合国框架下的全球人工智能治理，倡导开放合作、能力建设和风险共治，推动人工智能真正成为促进人类共同发展的公共产品。

技术治理必须与制度治理相结合

记者：这些年学界和产业界都在探索技术解决方案，从实际进展看，这些技术能在多大程度上解决算法决策的安全问题？技术路径有没有它的局限性？会不会出现“道高一尺、魔高一丈”的情况？

肖茜：用技术来解决技术带来的问题，确实是一条非常重要的路径。近年来，无论是算法可解释性、红队测试（模拟真实攻击者对系统进行对抗性测试的方法，旨在发现漏洞并提升安全性）、还是安全评估、模型对齐，都在不断提升人工智能系统的安全性和可靠性。但是，技术路径始终存在其局限性。我认为，这种局限主要体现在两个方面。一方面，人工智能的发展速度远快于治理规则的演进。今天解决的问题，可能很快就会被下一代模型所突破。随着模型能力不断增强，特别是智能体、自主决策、多智能体协同等技术快速发展，未来人工智能甚至可能具备一定程度的自主学习、自主优化和自我迭代能力。在这种背景下，安全治理确实可能面临“道高一尺、魔高一丈”的挑战，单纯依靠技术手段很难做到一劳永逸。另一方面，许多算法决策安全问题本身就不是技术能够独立解决的。例如，算法是否公平、是否符合伦理、是否侵犯个人权利，这些都属于价值判断和制度选择，而不是工程问题。

因此，技术治理必须与制度治理相结合。未来算法决策治理至少应坚持几项基本原则：第一，对于涉及生命安全、基本权利和重大公共利益的决策，必须保留实质性的人类判断，并留出充足的判断时间；第二，建立算法影响评估、模型测试审计和持续监测机制，确保人工智能在部署前、部署中和部署后都处于可测试、可监督的状态；第三，保障必要的透明度、解释权和救济权，使受到算法影响的个人拥有知情和申诉的机会；第四，明确开发者、部署者、使用者和监管者之间的责任边界，形成完整的责任链条。

除此之外，随着人工智能越来越广泛地应用于跨境场景，很多风险已经超越了一国的治理能力。例如，前沿模型可能被全球用户使用，算法漏洞可能迅速扩散，军事、网络、生物等领域的风险更具有全球外溢性。因此，还需要加强国际合作，推动风险评估、安全测试、技术标准和治理规则方面的协调，共同应对人工智能带来的全球性挑战。

只有形成完整的责任链条，才能避免出现“人人参与、人人免责”的治理困境

记者：算法参与决策最棘手的问题之一，是“出了事谁担责”，真出了安全问题，责任该怎么界定？

肖茜：这是目前全球人工智能治理面临的一个核心难题。过去，责任往往比较容易界定，谁作出决定，谁承担责任。但在人工智能时代，一项算法决策往往涉及多个主体，包括模型开发者、数据提供者、系统部署者、平台运营者以及最终使用算法作出决策的机构。任何一个环节出现问题，都可能影响最终结果。因此，很难简单地把责任归咎于某一个主体。

国际社会目前越来越倾向于建立“责任链条”的理念，也就是说，根据各主体在人工智能全生命周期中的职责划分责任，而不是由某一个主体承担全部责任。只有形成完整的责任链条，才能避免出现“人人参与、人人免责”的治理困境。

我国近年来出台《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》等法规，对平台责任、算法备案、安全评估等方面已经进行了积极探索，为人工智能治理提供了重要制度基础。但总体来看，目前我国乃至国际社会，都还没有形成一套十分成熟、可操作的算法决策问责机制。特别是在大模型和智能体快速发展的背景下，责任如何划分、举证责任如何分配、算法“黑箱”如何解释、跨境应用如何监管等问题，仍然需要进一步探索。

记者：对于政府、企业和公众这三方来说，各自应该扮演什么样的角色，才能共同织密算法安全这张网？

肖茜：算法安全不是哪一方能够独立完成的，而是一项系统工程，需要多方共同参与，形成协同治理的格局。对于政府而言，需要不断完善算法治理体系。面对快速发展的人工智能技术，监管不能停留在事后处置，而要坚持风险导向、敏捷治理，建立覆盖人工智能研发、部署和应用全过程的治理机制，完善法律法规、技术标准、风险评估和责任追究制度，为人工智能健康发展提供制度保障。

对于企业来说，应当切实履行安全治理的主体责任。安全不应该只是产品上线后的“补丁”，而应贯穿产品生命周期全过程，加强安全评估、红队测试、风险监测和持续优化，真正把“安全设计”融入技术创新之中。

对于公众来说，要不断提升人工智能素养。我们既要善于利用人工智能提升工作和生活效率，也要保持独立判断，不能盲目迷信算法，更不能把所有决策都交给机器。公众还应积极行使知情权、解释权和申诉权，共同推动人工智能更加透明、可信和负责任地发展。

算法安全不是限制人工智能的发展，而是为了让人工智能更好地服务于人。只有政府加强治理，企业守住责任，公众积极参与，形成共建共治共享的治理格局，才能共同织密算法安全这张网，让人工智能真正做到安全、可信、可控、向善。

本版图片均由AI辅助生成

如果算法判定我有罪……

孙亚慧

电影《极限审判》里有一个让人印象深刻的设定：洛杉矶警探雷文被绑在“审判椅”上，一个名为马多克斯的AI法官用90分钟决定他的生死。AI判定嫌疑人有罪，概率将近98%，嫌疑人需要自证清白、洗脱罪名。

这是科幻电影中的一幕，但仍旧给观众带来了很大的思考空间。

如果有一天算法更深程度、更大范围参与决策，那会是一个怎样的未来？

其实我们每天都在面对类似的“算法决策”。外卖平台用算法推荐餐厅，短视频用算法“猜你喜欢”，银行用算法决定你的信用额度。这些决策渗透在我们生活的细节里——你看到的每一条推送、被推荐的每一家店、被审批的每一笔贷款，背后都有算法在默默工作。大多数时候，它们运转良好，让生活更便捷、更高效。

《2026年国际人工智能安全报告》将人工智能风险概括为三类：一是恶意行为者滥用人工智能所产生的风险，二是人工智能系统出现故障或失控所产生的风险，三是人工智能广泛应用所带来的系统性风险。这一分类为理解算法决策安全提供了一个比较清晰的框架。

专家认为，当算法进入医疗、教育、就业等领域时，一个局部偏差可能通过自动化、规模化和跨系统连接被迅速放大，最终演变为系统性社会风险。例如，一个招聘算法的偏差可能影响数以万计求职者；一个信贷模型的错误可能造成某些群体长期被排除在金融服务之外；军事决策系统中的一次误识别，则可能直接造成人员伤亡甚至引发危机升级。因此，对于涉及生命安全、基本权利和重大公共利益的决策，尤其需要坚持实质性的人类监督，建立算法影响评估、持续监测、审计问责和申诉救济机制，防止局部技术问题演变为更广泛的社会风险。

近年来，我国相继出台了《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》等一系列文件，在全球范围内率先建立了较为完整的人工智能监管框架，这体现了我国对算法安全治理的高度重视。

但从实践来看，算法安全监管难点不少，很大一个原因在于监管对象发生了根本变化。传统监管主要针对固定的产品、明确的行为和相对稳定的规则，而人工智能算法，特别是大模型和智能体系统，则具有持续学习、不断迭代、多场景应用等特点，能力边界始终处于变化之中。技术的发展速度远远快于规则更新的速度，因此容易出现“监管追着技术跑”的情况。

技术在进步，它在帮助人，而不是取代人。真正的答案，是建立一套与技术进步同步的治理框架：让算法知道边界在哪，让开发者知道责任在哪，让用户知道申诉的门在哪。为算法参与决策系上“安全带”，它才能既跑得快，又跑得稳。