2024年3月 出版
（总第31期）

清华大学产业发展与环境治理研究中心
Center for Industrial Development and Environmental Governance
Tsinghua University

CIDEG 决策参考

政策研究报告

作者：贾 开

厘清算法安全（影响）评估制度逻辑 加快提升人工智能治理水平

《CIDEG 决策参考》

《CIDEG 决策参考》主要关注产业发展、环境治理和制度变迁三个领域的研究议题，希望学者们就这三个议题领域中的热点话题、研究前沿和国际比较等方面撰写政策报告，提供给相关领域决策者和学者们参考、学习和交流。每期推送一篇学者稿件，阅读受众包括CIDEG学术委员、学者网络与公众。其中高质量的稿件将经由CIDEG学术委员会推荐报送给国家决策部门。

欢迎您将相关主题的研究、观点和实践投稿给我们

投稿方式：请将稿件邮件发送至cideg@tsinghua.edu.cn

投稿邮件标题请注明【投稿-决策参考-单位-姓名】

◆ 期待您的赐稿！◆

清华大学产业发展与环境治理研究中心
Center for Industrial Development and Environmental Governance
Tsinghua University

CIDEG

厘清算法安全（影响）评估制度逻辑，加快提升人工智能治理水平

作者：贾开 上海交通大学国际与公共事务学院长聘副教授
清华大学产业发展与环境治理研究中心兼职研究人员

一、《生成式人工智能服务安全基本要求》出台的重要性

2023年8月15日，《生成式人工智能服务管理暂行办法》正式施行，其中明确指出，“提供具有舆论属性或者社会动员能力的生成式人工智能服务的，应当按照国家有关规定开展安全评估”，同时也要求在生成式人工智能技术研发过程中需要“开展数据标注质量评估”。为进一步落实该文件，2024年2月，全国网络安全标准化技术委员会发布了技术文件《生成式人工智能服务安全基本要求》（以下简称《安全基本要求》），从语料安全、模型安全、安全措施等方面提出了安全评估要求，为生成式人工智能服务提供者的安全评估工作、相关主管部门评判生成式人工智能服务安全水平提供了重要参考。

尽管《安全基本要求》作为技术文件并不具有强制性，其治理效果仍然取决于监管部门、一线技术研发与应用主体、市场第三方评估机构等利益相关主体如何理解、使用、执行，但从制度演化视角来看，其是我国国家层面治理机构第一次对人工智能算法安全评估的对象、范畴、标准等实质性内容提出制度要求，对于我国完善人工智能算法安全评估制度建设具有关键性的推动作用。换言之，只有迈出了《安全基本要求》这样一个重要起点，我们才可能在不断的补充、调整、完善过程中积累治理经验，真正开启人工智能算法安全评估的制度探索进程。

从全球视角来看，当前世界各主要国家的人工智能治理正在经历从原则理念向制度落实方向发展的新阶段，人工智能安全（影响）评估作为基础性制度已经被普遍纳入各国制度框架，欧盟、加拿大、美国等也初步形成了一定的制度探索。相比之下，我国虽在不同政策文本中都明确提出了要开展人工智能安全（影响）评估的基本要求，但直到《安全基本要求》的提出之前，我国并未形成较明确的评估框架，也未提出较明确的评估要求。在此背景下，《安全基本要求》的出台，有助于我们在比较不同国家制度差异的基础上，积累中国在算法安全（影响）评估方面的治理经验，进而向人工智能全球治理做出贡献，协力各方共同促进人工智能治理。

不过人工智能安全议题本身的复杂性和不确定性，使得人工智能算法安全（影响）评估的体系构建与机制探索并非单靠一个文件就能完成，其仍然依赖于自上而下对整体性顶层设计逻辑的斟酌权衡，以及自下而上治理经验的汇集与扩散。在《安全基本要求》作为重要制度文件出台的背景下，对人工智能算法安全评估的制度逻辑展开分析，因此也具有了极为重要的理论价值与实践指导意义。

二、人工智能算法安全评估的制度逻辑与目标

从制度特点来讲，安全（影响）评估制度的特点在于将治理重心从结果转向过程，在淡化追溯直接责任因果链条的同时，强调治理信息的记录、共享、监督，并在此过程中积累治理经验，同时为算法设计者、应用者的及时调整提供参考。具体而言，传统监管是直接面向可能出现的风险而展开，旨在对市场失灵现象作出回应以避免出现权益侵害风险，其针对的是风险结果而并不特别在意导致风险产生的过程；但安全（影响）评估制度却“反其道而行之”，更关注安全风险产生过程，并要求作为一线主体的技术创新者、应用者基于评价结果而对其内部的技术生产、应用过程作出调整。如果说传统监管仍然建立在较为明确的“市场-政府”二分边界基础上，那么安全（影响）评估制度则打破了这一“二分法”，而直接切入至技术创新应用的生产管理过程，以求对风险作出敏捷回应。

具体到人工智能算法作为治理对象而言，算法安全风险的环境嵌入性特征、算法技术方案的难解释特征、算法安全风险的概率性而非个案性特征，都是催生算法安全（影响）评估制度的重要原因。在此背景下，人工智能算法安全（影响）评估制度的目标主要围绕两方面而展开：对算法治理过程进行记录以累积治

理经验与知识，基于评估结果而要求算法设计与应用者及时干涉创新应用流程以作出敏捷回应。

然而，作为一种新兴事物，人工智能算法安全（影响）评估制度究竟如何建设，仍然是一个全新命题。综合制度沿袭和具体政策实践来看，算法安全（影响）评估有可能采取环境影响评估、数据保护信息评估、列表式评估这三种基本模式。但每一种模式都不是完美的，都存在其优势和不足，而这也正是我们探索算法安全（影响）评估制度建设过程中所需要关注的取舍选择与权衡之处。

三、人工智能算法安全评估的制度建设模式比较

自上个世纪六七十年代在全球各国逐步建立之后，环境影响评估几乎可被视为安全（影响）评估领域最为典型、最具代表意义的制度典范。从制度内涵来看，环境影响评估的关键在于三点：基于环境影响级别的大小而分别提出不同程度的环境影响评估要求、广泛纳入公众参与以确保环境影响评价的完整性和有效性、要求开展替代方案对比分析以寻找最优方案。在算法安全（影响）评估领域，部分利益相关方提出的制度设计方案即是以环境影响评估为蓝本，较有代表性的例如纽约大学智库AI Now在2018年提出的公共部门算法影响评价框架，以及欧洲议会研究服务中心（European Parliamentary Research Service, EPRS）2019年研究报告提出的包含8个环节的算法影响评价方案。不过相关反思也指出，环境影响评估的冗繁要求，以及对广泛公众参与的强调，并不一定适合人工智能算法治理领域，特别是完全的公开可能侵害商业秘密并带来更大衍生风险。

数据保护影响评估（Data Protection Impact Assessment, DPIA）的制度框架基本上源起于欧盟《一般通用数据规定》第35款。从流程要求来讲，DPIA包括审查基本信息以判断是否需要进行数据保护影响评估、围绕各目标展开影响评估以确定风险、向监管者报告并分析降低风险的可能措施、在全过程咨询利益相关方、复审以在必要时重启影响评估等。与环境影响评估相比，DPIA的差别体现在两点：不特别强调替代方案的比较、公开并非一项法律责任的要求而取决于数据处理者的决定。DPIA与环境影响评价的差别，事实上体现了其特殊的制度逻辑：考虑到数据治理的复杂性，DPIA更强调监管者与被监管者的合作而非对抗，数据处理者作为被监管方的利益（例如商业秘密）因此需要得到承认以激励其寻找数据保护方案的积极性。但对DPIA的反思也正来源于此：在缺少充分公开与监督的环

境下，“监管俘获”的担忧是否会影响DPIA试图构建的监管者与被监管者“合作”愿景的实现？

第三种模式是采取问题列表清单的方式，而代表性实践便是加拿大政府在2019年4月出台施行的“自动决策指令”（Directive on Automated Decision-making），其要求所有被用于行政决策领域的自动决策系统都需要在采购或使用前，按照问题列表清单进行算法影响评估，而该清单每两年都将被重新评估并更新。具体而言，问题列表共包含风险以及风险预防这两个部分，形成了涉及商业流程、所用数据、模型逻辑等多个方面的80余个具体问题，要求被评估主体对这些问题作出回答，然后基于答案来对被评估系统的风险进行评分，并按照评分将之相应归类到上述1-4个风险级别之中。该列表清单基本涵盖了当前算法影响评估利益相关方关心的主要问题，而评估结果的公开也进一步促进并提升了自动化决策系统的可解释性、可被监督性，并因而提升了算法治理水平。但针对该模式的反思在于：列表清单能否涵盖算法治理的所有问题，以及预先设定好的问题及答案是否可能会淡化治理经验积累这一算法安全（影响）评估的制度目标。

上述各有优点和不足的三种模式的对比分析说明，算法安全（影响）评估制度建设还远未到成熟时期，在“基于标准的监管与探索多方合作”、“保护商业秘密与公众参与”、“监管俘获与透明”、“封闭列表与开放治理”等多对治理关键点上，都需要决策者的权衡取舍，并最终找到适应本国技术发展与应用治理需要的平衡点。

四、欧、美开展人工智能算法安全评估的政策实践

不同国家出于差异化的制度文化背景、产业发展阶段与治理需求，在算法影响评价的制度选择方面仍然体现了一定的侧重，并具有差异化特点。除上文提到加拿大采取的列表清单模式之外，欧盟、美国也在一定程度上形成了各具特色的算法安全评估制度。

欧盟《人工智能法案》主要通过两种方式引入了算法影响评估，并奠定了其在该法案中作为制度基础的重要作用。一方面，法案要求高风险人工智能系统在投入使用前应执行“合规性评估（Conformity Assessments）”，即需要按照现行法律规则来评估该系统是否满足各项法律要求，并将结果公开；另一方面，法案要求开展基本权利影响评估（Fundamental Rights Impact Assessments, FRIA），较为

宽泛地提出了9项评估内容要求，且不要求完全公开而主要向监管机构以及利益相关方汇报或分享。FRIA较有启发意义的两个特点，其一在于将环境影响视为基本权利而纳入评估范畴之中，这可被视为针对大模型快速发展的敏捷回应；其二在于提供了“例外豁免”条款，规定在应急状况或对于中小企业等缺少评估资源和能力情况下，可以在未展开FRIA的前提下应用人工智能。

与欧盟相比，算法安全（影响）评估制度在美国人工智能治理框架下并没有扮演特别重要的基础性角色。美国当前主要将人工智能治理置于垂直领域监管框架之下，因而跨领域、一般性的算法算法安全（影响）评估制度建设并非其重点。但即使如此，两个政策文件仍然值得重点关注。一方面，国家标准和技术研究所（National Institute of Standards and Technology, NIST）提出的“人工智能风险管理框架”（AI Risk Management Framework, AI RMF）可被视为专业机构试图影响行业实践的重要工作，该框架同样提出了风险评估的框架、流程乃至指标；另一方面，拜登政府联邦管理与预算办公室在2023年10月发布的“促进联邦机构推动人工智能治理、创新、风险管理的备忘录”作为联邦机构应用人工智能的治理框架，同样明确了算法影响评估的关键地位，其区分了“安全影响”（Safety Impacting）和“权益影响”（Rights Impacting）两种类型，并基于这两类区分而分别提出了不同程度的评估责任要求。

五、《安全基本要求》要点特征评析及中国算法安全评估制度的未来改革

在理解算法安全（影响）评估制度逻辑、模式及主要国家政策实践的基础上，再回过来分析《安全基本要求》的要点特征，以及更一般的我国算法安全（影响）评估制度建设的改革未来，便可能具有更完整的观察视野。

从内容上讲，《安全基本要求》涵盖语料、模型、场景等多个环节的治理思路，同样反映了算法安全（影响）评估作为一种治理制度而与传统规制思路的差异所在。在不同环节提出的具体治理要求，在模式上可以类比于加拿大列表清单模式，即基于预先设定的具体问题或指标要求来引导自评估或第三方评估工作。正如前文分析所指出，考虑到算法安全（影响）风险的不确定性及其产生机制的复杂性，类似于DPIA，算法安全（影响）评估同样需要依赖包括监管部门、一线企业或机构、第三方机构乃至公众在内的多方主体的监管合作，而这其中又特别以一线企业或机构的自评估或第三方评估为要，因它们所拥有的信息、资源优势

能够贡献出有效的治理知识和经验；但另一方面，如果缺少一定的约束，完全放任的自评估或第三方评估也可能因为缺少充分引导而起不到积累治理经验的积极作用，如果进一步考虑到“共同无知”现象的存在则更是如此。在此背景下，以预先设定方式提出具体评估要求，便可被视为启动评估流程、推动后续制度持续演化的一个重要基础，而这同样应被视为《安全基本要求》的积极作用所在。

从未来改革来看，如果要充分发挥《安全基本要求》之于算法安全（影响）评估乃至于人工智能治理的潜在价值，后续的执行以及配套措施的跟上，仍然是关键。

一方面，监管部门、一线主体、第三方机构等多方主体对算法安全（影响）评估的制度逻辑，以及《安全基本要求》在该制度框架下所应扮演的角色，需要清晰、整体且具有共识性的理解与把握。在一定程度上，《安全基本要求》不应被视为对监管部门以及被监管主体套上的一个“硬筐”，其更多应被视为多方主体通过算法安全（影响）评估而积累算法治理知识与经验、在适当时机对算法生产应用流程做出干涉的起点。在此意义上，究竟如何应用《安全基本要求》、发挥其推动算法安全（影响）评估的积极作用，仍然是监管部门需要考虑的重要工作。

另一方面，继《安全基本要求》之后，算法安全（影响）评估的配套制度可能需要加快演进，以减少制度不确定性空间。这既体现为我国算法安全（影响）评估制度顶层设计基本逻辑的讨论与共识，也体现为具体机制层面的丰富与拓展。就当前进程而言，我们还有诸多重要问题尚待回答：“符合、不符合、不适用”作为《安全基本要求》的评估结果应该如何应用，算法安全（影响）评估如何与算法备案、算法分级分类治理等衔接起来？自评估或第三方评估的有效性应该如何界定，应该如何以《安全基本要求》为底本而展开监管部门与一线主体、第三方以及公众的联动？算法安全（影响）评估的基准共识、标杆示范、同行评议等工作应该如何开展，以实现治理知识的扩散和一般治理水平的提升？

改革不是一蹴而就，但改革过程中的每一个节点都是往前推进的重要一步。在此意义上，我们同样应以《安全基本要求》为起点，切实推进算法安全（影响）评估工作，不断提升人工智能（尤其是当前生成式人工智能）的治理水平。

扫码关注

清华大学产业发展与环境治理研究中心

主　编：薛 澜　陈 玲　　责任编辑：赵 静

文字编辑：苗馨竹　潘莎莉

清华大学产业发展与环境治理研究中心 编辑出版

Email：cideg@tsinghua.edu.cn

电　话：010-62772497　62772593