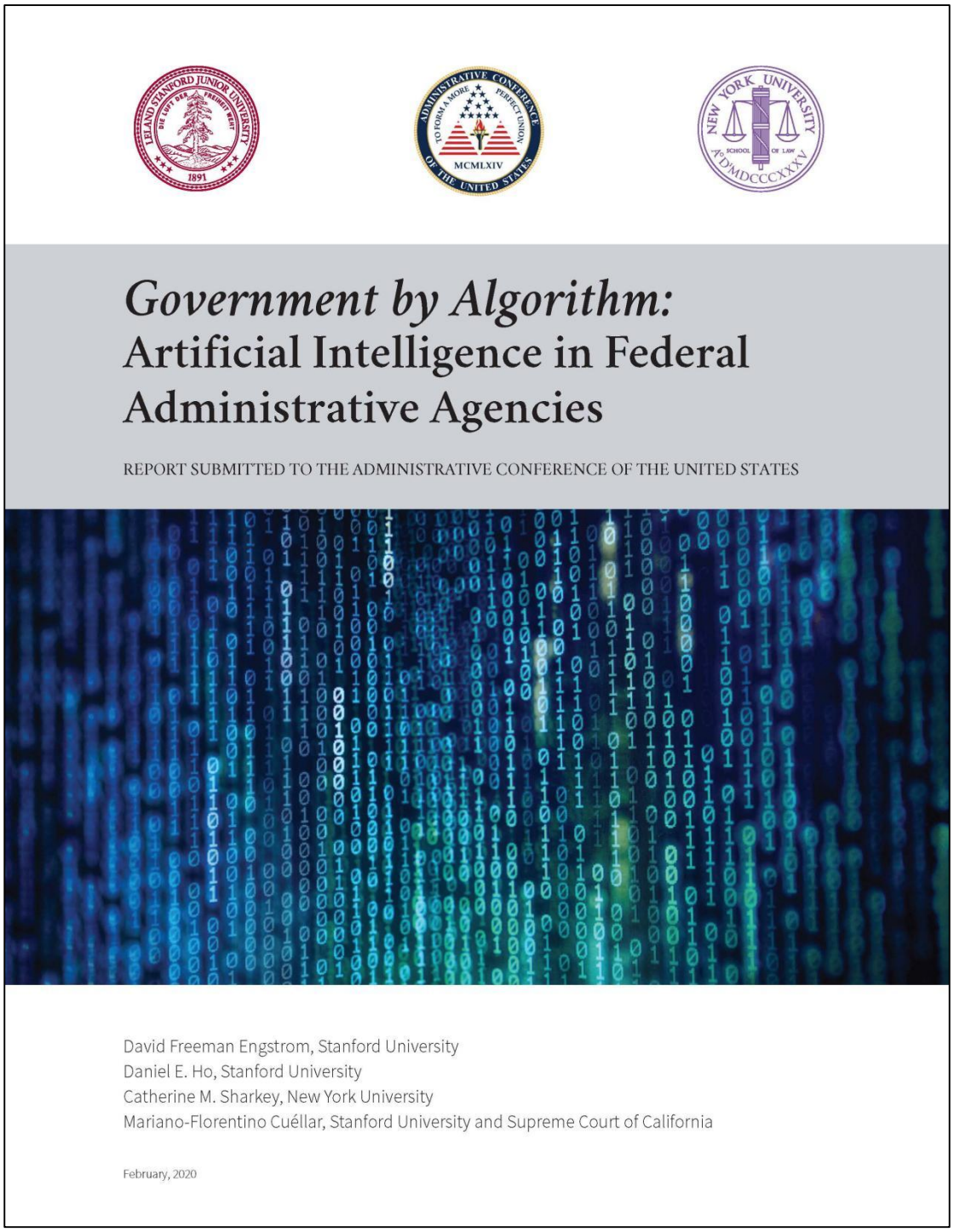
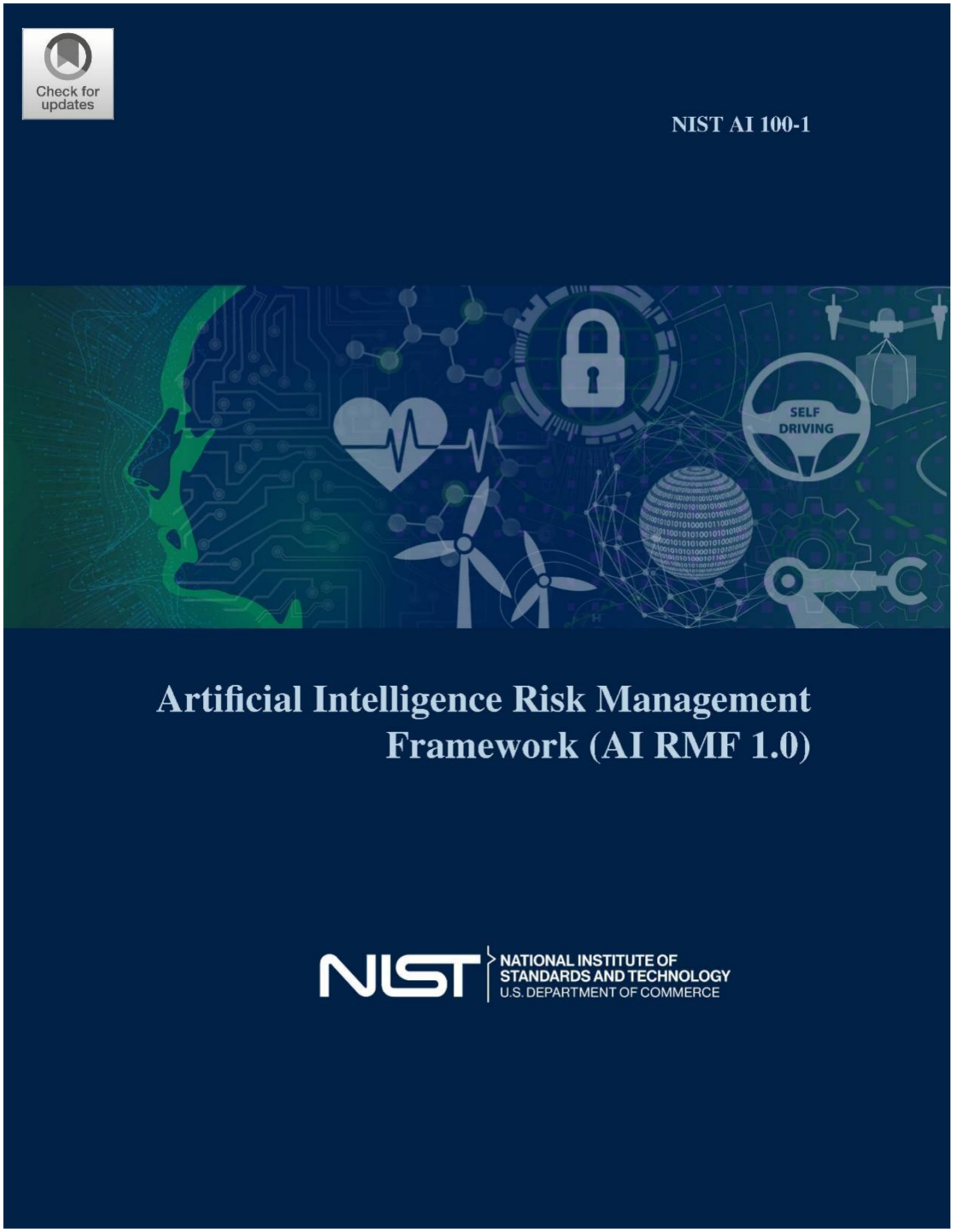
AI 进入监管执法：
美国联邦层面的应用图景与治理清单



核安全监管技术支持工作涉及大量技术审查、数据分析、文件审查、监测评估和公众沟通等通用信息处理场景，了解

其他监管体系如何讨论人工智能（AI）应用与治理，有助于跟踪国际公共部门 AI 治理的最新公开信息。本文基于美国公开报告和联邦政策文件，梳理 AI 在监管执法中的应用场景、风险议题、治理框架和制度安排。



AI 已经从“实验”进入监管执法流程

美国联邦机构 AI 使用已不是个别部门的试点现象，而是逐步进入监管执法、裁决、监管分析、公众参与和内部管理等核心治理环节。

用例数量快速增长

美国行政会议（ACUS）研究报告《人工智能与监管执法（Artificial Intelligence and Regulatory Enforcement）》指出，2020 年 ACUS 报告已识别出 64 个联邦机构中的 157 个 AI 用例；仅两年后，美国政府问责局（GAO）的调查又识别出超过 1200 个当前和计划中的用例。报告认为，公开可见的统计很可能只是“冰山一角”，AI 对政府运行方式的长期影响更为深刻。

斯坦福大学和纽约大学团队为 ACUS 完成的《算法治理：人工智能在联邦行政机构中的应用（Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies）》对 142 个联邦部门、机构和子机构进行了系统调查。结果显示约 45%的受调查机构（64 个）已经规划、试点或实施 AI/机器学习工具（共记录 157 个用例），说明美国联邦机构的 AI 应用面已经较广，但技术成熟度、文档透明度和部署深度差异很大。

监管执法是高风险、高影响场景

《人工智能与监管执法》将“监管执法”界定为发现、调查和起诉当前或潜在的不合规行为，包括证券执法、税务欺诈调查、环境违法检查等。报告强调，监管执法直接面对

被监管对象，涉及程序正义、正当程序、公平对待和机构合法性，因此是 AI 应用中风险较高的场景之一。

《算法治理》进一步将 AI 治理任务分为五类：监管研究、分析和监测；执法；公共服务与公众参与；内部管理；裁决。报告指出，AI 在执法和裁决中的使用，正在接近国家强制力和再分配权力的核心，因而比一般内部管理工具更受关注。

监管执法中的典型用例

公开资料提供了多个联邦机构的具体案例。

美国证券交易委员会（SEC）：用算法“缩小排查范围”

SEC 开发了多类执法辅助工具：

公司发行人风险评估（Corporate Issuer Risk Assessment，CIRA）：面向公司发行人财务报告异常，使用约 200 项指标和随机森林模型，基于历史申报文件、财报重述和既往执法行动等数据，识别可能存在的可疑行为。算法输出由执法人员结合其他材料审查，不直接作出执法决定。

高级关系型交易执法指标调查系统（Advanced Relational Trading Enforcement Metrics Investigation System，ARTEMIS）：分析交易记录，识别多名交易者之间的模式和关系，重点针对重复性内幕交易等行为。由于难以获得完整“真实标签”，系统采用无监督学习进行异常检测。

异常交易与关联分析系统（Abnormal Trading and Link Analysis System，ATLAS）：面向首次内幕交易，使用人工设计的少量特征和监督学习模型，对交易是否可疑进行判断。

ADV 表格欺诈预测模型（Form ADV Fraud Predictor）: 对投资顾问申报文件进行自然语言处理，使用主题模型和随机森林模型，将注册机构分为高、中、低优先级，供执法人员进一步审查。

SEC 一线执法人员不满足于“高风险”分类结果，而要求知道“申报文件的哪一部分触发了分类、为什么”。这种来自一线使用者的解释性需求，推动技术人员关注可解释性。相关资料还指出，执法工具面临训练数据偏差、历史执法反馈循环、违法行为动态变化、被监管对象“博弈”等问题。SEC 已推动机器可读申报格式，并建设合并审计跟踪（CAT，consolidated audit trail）等数据基础设施。

美国海关与边境保护局（CBP）：人脸识别与风险预测

CBP 使用人脸识别和风险预测工具处理旅客和货物。

人脸识别：用于机场、陆路口岸等场景，CBP 通过承包商获取算法和系统，部分技术细节因知识产权保护而不公开;

风险预测：“自动目标系统”及其后续项目整合多个政府和商业数据源，对旅客和货物进行风险评分或筛查。

但使用中还存在以下问题：由于使用专有技术，无法解释人脸识别技术的失败率原因；相关应用引发隐私、数据共享、平等保护和行政程序法上的讨论；人脸识别系统可能因训练数据不均衡而对不同人群产生不同误差，系统也可能受到对抗性攻击。

这个案例说明，执法类 AI 不仅涉及技术性能，还涉及数据来源、承包商透明度、公众同意、跨机构数据共享和法律

责任。

美国社会安全局（SSA）：大规模裁决中的 AI 辅助

SSA 是美国联邦政府中裁决量最大的机构之一。相关资料介绍了三类工具:

案件聚类与“微专业化”: 利用案件管理系统中的年龄、伤残类型、州别等元数据，将相似案件聚类，帮助裁决人员按主题集中处理。SSA 报告称案件处理时间减少 12%，行政上诉法官退回率降低 7.5%，但该工具为自愿使用，效果计算方式不完全公开。

快速伤残认定（QDD，Quick Disability Determination）与获准概率预测: 通过预测模型识别更可能获批的案件，将其送入快速通道; 在听证阶段，也用模型预测案件获准概率，用于调整审理顺序，但不向裁决人员分享概率，不直接用于裁决。

自然语言处理工具 Insight: 分析裁决草案，生成 30 多个“质量标记”，提示可能违反政策或存在内部不一致之处。标记只有在准确率超过 80%时才纳入系统。SSA 监察长办公室审计显示，80%的受访者认为标记准确，20%认为不准确; 仅 30%的受访者认为 Insight 改善了处理时间。

SSA 的经验表明，既懂业务又懂技术的内部人员非常重要; 同时，数据非结构化、标签成本高、旧职业词典过时、自动化偏见和裁决独立性等问题仍需关注。

美国专利商标局（USPTO）：分类、检索与商标审查

USPTO 在专利分类、专利现有技术检索、商标图形分类

和商标近似检索中试点 AI 工具。

相关资料指出：专利分类由承包商提供机器学习分类服务；曾开发“Sigma”检索工具，但因只对具有计算机科学背景的审查员有明显效率提升，最终未部署；商标图形分类使用卷积神经网络等深度学习模型，商标近似检索使用图像嵌入和相似度排序；深度学习有望提升检索召回率、跨语言检索、权利要求与现有技术段落匹配等能力；仍存在承包商利益冲突、申请人可能“博弈”分类和检索系统、审查员对工具的接受度等问题。

Sigma 案例是一个重要信息点：技术上有一定先进性的工具，如果不符合一线用户的实际工作方式和能力结构，也可能无法产生预期效果。

美国食品药品监督管理局（FDA）：不良事件报告系统（Federal Adverse Event Reporting System，FAERS）自然语言处理（NLP，natural language processing）

FDA 试点使用文本挖掘和 NLP 分析 FAERS 数据库中的药品和器械不良事件报告。

一个肝衰竭项目使用文本挖掘、主题建模、决策树和简单神经网络预测严重不良结局。决策树在预测严重结局时真阳性率约 91%，假阳性率约 4.9%。但仅靠 FAERS 数据难以推断药品与肝衰竭之间的因果关系，因为数据库主要记录不良事件，缺少人群用药基线。

另一个与斯坦福合作的项目，用专家标注的报告训练模型，对报告进行风险排序，帮助审评人员优先处理。但 FDA

官员认为部分试点尚未达到可部署的准确度。

相关资料指出，除了用 NLP 从非结构化文本中提取结构化数据，也可以考虑要求被监管方提交更适合用途的结构化数据。

该案例说明，AI 在监管分析中可以辅助排序和分流，但不能替代因果推断和专家判断。

联邦通信委员会（FCC）与消费者金融保护局（CFPB）：评论和投诉处理

FCC 在“网络中立”规则制定中收到超过 2000 万条评论。相关资料指出：只有 8%的评论是独特的，前 10 条和前 100 条最常见评论分别占 66%和 89%；分析采用人工分类、关键词匹配和 NLP 相结合的方式，NLP 仅用于最后约 1%的评论，分类准确率约 95.2%。分析还发现大量重复、虚假或机器生成评论，控制这些因素后，结论可能发生变化。

CFPB 使用 NLP 对消费者投诉进行主题建模、趋势识别和危害预测，并用自动化和人工复核相结合的方式去除个人身份信息。相关资料指出，评论和投诉处理工具可以帮助机构应对“超大规模参与”，但也带来机器人评论、代表性偏差、承包商利益冲突、对特定语言风格人群的系统性忽视等问题。

美国邮政局（USPS）：自动驾驶车辆

USPS 试点农村路线自动驾驶投递车和长途自动驾驶卡车。相关资料指出：自动驾驶车辆涉及侵权责任、数据所有权和隐私、工会和劳动力影响、公众接受度等问题。USPS 监察长报告认为，自动驾驶可能提高安全性、降低事故和燃油

成本、提升投递效率，但也可能带来长期劳动力替代和公众信任问题。

风险与挑战

监管执法中使用 AI 的风险议题主要集中在以下方面。

偏见与歧视

美国国家标准与技术研究院（NIST）《人工智能风险管理框架（Artificial Intelligence Risk Management Framework）》将“公平——有害偏见得到管理”列为可信 AI 特征之一，并将偏见分为系统性偏见、计算与统计偏见、人类认知偏见。《人工智能与监管执法》和《算法治理》都提到，训练数据不具代表性、历史执法偏差、特征选择不当、反馈循环等可能导致歧视性结果。

透明性与可解释性

NIST 报告指出，透明性回答“发生了什么”，可解释性回答“为什么作出决定”和“为什么作出决定及其含义”。但在监管执法中，完全公开算法细节可能导致被监管对象“博弈”系统；承包商又以商业秘密为由限制披露。多份资料都提到透明与效能之间的张力。

人的参与并不自动等于有效监督

《人工智能与监管执法》指出，人对自动化决策的审查并不总是有意义的，人类有服从自动化建议的倾向。SEC 案例中，一线人员要求解释，说明“有意义的人类监督”需要制度、能力和文化支撑。SSA 的 Insight 工具也面临自动化偏见和裁决独立性问题。

数据质量与“真实标签”难题

SEC、SSA、FDA 等案例都提到，监管执法数据往往不是随机样本，历史执法数据只覆盖被调查过的对象，难以代表全部违法行为；违法行为会动态变化，模型需要持续更新；高质量标签需要法律和领域专家，成本很高。

博弈与对抗性学习

被监管对象可能根据算法规则调整申报、交易或申请行为，以规避审查。SEC 的 ADV 表格工具、USPTO 的分类和检索工具都可能面临此类问题。增加模型复杂度、随机化、使用不可变特征、对抗训练等手段可以提高防博弈能力，但也可能降低可解释性。

承包商与采购风险

多份资料指出，承包商提供的 AI 系统可能存在“黑箱”问题。CBP 承认无法解释虹膜扫描失败率；USPTO 的专利分类承包商同时向专利申请人提供服务，存在利益冲突；采购合同可能限制政府获取技术细节和数据进行审计。相关资料建议在合同中明确数据权利、知识产权、性能监测、持续改进、供应商锁定保护等条款。

能力建设不足

《算法治理》强调，AI 工具需要内部技术能力，才能把法律和政策要求转化为系统设计，才能评估承包商工具，才能持续更新和退役系统。数据显示，超过一半的用例由机构内部开发，说明内部能力具有重要价值；但政府面临薪酬、招聘、数据基础设施和旧系统改造等约束。

程序正义与救济

美国行政管理和预算局（OMB）备忘录要求，对高影响AI，机构应确保受影响个人能够获得及时的人工审查和申诉机会；如果现有申诉程序可以扩展，应覆盖 AI 辅助决策。《人工智能与监管执法》也建议，风险评价应配套对特定敏感用途的限制和对个体损害的救济机制。

公共信任与机构合法性

《人工智能与监管执法》指出，公共信任可能代际积累、一夜受损。监管执法中的 AI 应用如果被视为不公平、任意或不透明，可能进一步侵蚀公众对行政机关的信任。相关资料建议，风险评价不仅评价错误率，还要评价系统对机构信任和合法性的影响。

失败退出与沉没成本

《人工智能与监管执法》和 NIST 报告都强调，机构应预先设计系统退役和退出路径，避免因沉没成本而继续使用表现不佳的系统。OMB 备忘录明确，如果高影响 AI 不符合最低风险管理实践，机构必须安全停止使用相关功能。

美国联邦治理框架

NIST 报告和 OMB 备忘录系统描述了治理框架。

NIST 报告：自愿性、全生命周期风险框架

NIST 报告于 2023 年 1 月发布，定位为自愿性、权利保护、非行业特定、用例无关的框架。其核心内容包括:

可信 AI 的七个特征：有效可靠、安全（safe）、安全防护与韧性（secure and resilient）、可问责和透明、可解释与可

解读、隐私增强、公平且有害偏见得到管理。NIST 强调，可信度取决于最薄弱环节，各特征之间可能存在权衡。

四个核心功能：治理：建立风险管理文化、政策、流程、角色和责任，贯穿其他功能；**映射**：建立上下文，识别风险和影响；**测量**：使用定量、定性或混合方法分析、评估、基准测试和监测风险；**管理**：分配风险资源，制定响应、恢复和沟通计划。

全生命周期和持续监测：NIST 报告强调，风险管理应在设计、开发、部署、运行和退役全过程中持续进行；应记录剩余风险，定期评审，支持系统退役。

第三方和供应链风险：治理和管理功能都要求管理第三方软件、数据和模型风险。

人机交互：附录讨论人类角色、认知偏见、自动化偏见、人类对 AI 输出的不同理解等。

NIST 报告不是强制要求，但被多方视为美国联邦 AI 风险管理的基础参考。

OMB 备忘录：联邦机构 AI 使用的最新管理要求

OMB 备忘录适用于行政部门和机构，包括独立监管机构，重点包括：

三大优先事项：创新、治理、公众信任。

AI 战略：CFO 法案机构应在 180 天内制定并公开 AI 战略，说明当前和计划中的重点用例、AI 成熟度目标、基础设施、数据、人才、治理和采购计划。

首席 AI 官（CAIO）：每个机构在 60 天内指定 CAIO；

CFO 法案机构的 CAIO 应达到高级行政服务或同等层级；CAIO 负责推动创新、协调合规、监督高影响 AI 风险管理、跟踪用例等。

AI 治理委员会： CFO 法案机构在 90 天内建立或利用现有治理机构，由副部长级官员任主席、CAIO 任副主席，覆盖 IT、网络安全、数据、预算、统计、法律、隐私、民权等领域。

首席 AI 官理事会：OMB 在 90 天内召集跨机构理事会，协调 AI 发展和使用，共享模板、格式、技术资源和优秀实践，5 年后解散。

AI 用例清单： 机构至少每年编制 AI 用例清单，提交 OMB 并公开。

高影响 AI： 定义为输出作为决策或行动的主要依据，并对个人或实体的公民权利、公民自由、隐私，或教育、住房、保险、信贷、就业等机会，或关键政府资源和服务，或人类健康与安全，或关键基础设施和公共安全，或战略资产和资源产生法律、实质性、约束性或显著影响的 AI。备忘录还列出推定高影响类别，包括关键基础设施安全功能、可能造成显著伤害的机器人/车辆/工业设备、执法中的风险评估和生物识别、关键联邦服务申请行政裁决等。

高影响 AI 最低风险管理实践： 部署前测试和风险缓解计划；AI 影响评估，至少包括预期目的和收益、数据质量和适用性、潜在影响、复评安排、成本分析、独立审查结果、风险接受确认；持续监测性能和不利影响；对操作人员进行

充分培训和评估；适当的人工监督、干预和问责，包括失效保护设计；为受影响个人提供及时人工审查和申诉机会；征求并纳入最终用户和公众反馈。

豁免与公开：CAIO 可在特定条件下豁免部分要求，但须每年认证、集中跟踪、30 天内报告 OMB，并公开摘要。

不合规处理：如果高影响 AI 不符合最低实践，机构必须安全停止使用相关功能。试点项目在限定规模、CAIO 认证、告知参与者和尽量应用最低实践的条件下可豁免。

这套要求体现了“分级治理、全生命周期、公开透明、人工干预、可退出”的制度设计。

联邦层面的组织与制度安排

GAO 报告《人工智能：在要求与咨询小组指导下的联邦行动（Artificial Intelligence: Federal Efforts Guided by Requirements and Advisory Groups）》梳理了截至 2025 年 7 月的联邦 AI 要求和治理组织。报告识别出 94 项联邦政府层面 AI 相关要求，涉及法律、行政命令和指导文件；并列出 10 个行政分支 AI 监督和咨询组织，包括：

- 白宫科技政策办公室（OSTP）；
- 国家科学技术委员会（NSTC）；
- NSTC 机器学习与人工智能小组委员会；
- NSTC 人工智能遴选委员会；
- 科学和技术企业委员会；
- 网络与信息技术研发 AI 研发机构间工作组；
- 美国总务管理局（GSA）AI 卓越中心；

- 国家 AI 倡议办公室;
- 国家 AI 咨询委员会;
- 总统科技顾问委员会。

GAO 报告还指出，2023 年 12 月报告发现多数机构已建立 AI 用例清单，但存在不完整、不准确问题；GAO 向 19 个机构提出 35 项建议，截至 2025 年 7 月仅 3 个机构落实了 4 项建议，多数建议尚未落实。这说明治理要求与执行能力之间仍有差距。

决策与评估方法

决定是否使用机器学习的多因素分析

ACUS 报告《政府机器学习应用框架（A Framework for Governmental Use of Machine Learning）》中提出，政府决定是否用机器学习替代或辅助人类决策时，可采用“多因素政策分析”。其核心不是简单问“AI 是否先进”，而是问“与现有人类决策相比，AI 是否更好”。

框架包括三类因素:

使用前提: 目标清晰且可通过数学表达；有足够数量和质量的相关数据；数据具有外部有效性，能适应变化环境。

改进结果: 目标绩效更准确、更快、更便宜、更一致；对直接受影响用户的影响；对更广泛公众的影响。

法律和公众接受风险: AI 在决策中的角色是“输入”“默认”还是“决定”；事项影响高低；诉讼和公众争议风险。

该报告还提醒，人类决策本身也有记忆限制、疲劳、年龄、冲动、感知偏差、认知偏差和群体决策问题；因此比较

基准应是“现有人类决策”，而不是完美理性。

面向实施机构和联邦治理结构的建议

《人工智能与监管执法》在结论部分提出一组面向实施机构和联邦治理结构的建议，主要包括:

- 理解风险评价的局限，考虑 AI 系统产生不可预测结果和超出初始批准用途的倾向;
- 风险评价不能只看直接损害，还要考虑 AI 是否削弱公众对机构的信心和合法性认知;
- 风险评价应配套对 AI 方案成本和收益的批判性评估，表现不达标的系统应搁置，警惕沉没成本;
- 早期风险缓解应包括退出路径,并在产品全生命周期更新;
- 建立覆盖决策全流程的外部咨询,包括参与式论坛和与公民社会等的共同决策机制;
- 透明是咨询的前提，至少公开风险评价和相关文件，必要时将报告转化为更易理解的格式或为社区伙伴提供培训;
- 信息披露应偏向透明,只有明确损害司法有效运行且损害大于公共利益时才不公开;
- 对特别严重的风险,风险评价应辅以禁止性措施或个体救济;
- 联邦层面应有跨学科、独立、有执行力的 AI 监督结构。

这些建议与 NIST 报告、OMB 备忘录形成互补: 前者偏

治理原则和制度设计，后两者偏操作流程和机构责任。

结语

美国联邦层面关于 AI 与监管执法的讨论，已经从“要不要用 AI”转向“如何用、如何管、如何退”。反复出现的关键词包括：风险分级、全生命周期、高影响 AI、影响评估、人工监督、申诉救济、公众参与、透明度、采购合同、内部能力、失败退出和跨机构协调。

应该特别注意的是：AI 不是天然更优。《政府机器学习应用框架》强调，比较对象应是现有人类决策，而不是理想化的完美决策;《人工智能与监管执法》强调，风险评价有局限，最坏情形、机构合法性和失败退出必须纳入;《算法治理》强调，技术能力、数据基础设施和制度设计决定 AI 能否真正提升治理质量; NIST 报告和 OMB 备忘录则提供了较为系统的流程和机构责任框架。

国际交流合作部　余少青　供稿

编译自美国核监管委员会官网等

文章内容不代表本公众号观点