人工智能法治中“系统测评”的应用限度

——与苏宇教授商榷

刘瑞强

【内容摘要】 从“算法解释”到“系统测评”，标志着人工智能法治基础信息工具的变革。“系统测评”在弥补算法解释机制局限、构建可信人工智能生态方面具有显著优势。然而，“系统测评”在实际应用中仍面临公平性失真、性能误导及技术垄断等风险，在技术尚未成熟、标准尚未统一的背景下，不宜贸然将其全面法治化。当下的妥善方案在于，审慎界定“系统测评”在人工智能法治框架中的“底线保障”角色，通过分类分级实施、构建权威技术标准、确立独立透明的认证机制等路径，以“软法”先行、渐进规制的策略推动其与法治体系良性融合，实现治理效能与创新保护的平衡。

【关 键 词】 系统测评　人工智能治理　法治限度　技术标准

【作　　者】 刘瑞强，中国政法大学习近平法治思想研究与国际传播中心研究员。（北京 100088）

大模型的兴起，标志着人工智能领域正经历一场范式革命。同时，其不可预测的“涌现”能力及其复杂风险对传统治理手段构成重大挑战，我们迫切需要建立贯穿研发全生命周期的嵌入式治理模式，以技术工具赋能监管，并构建多方协同的敏捷治理新范式。相比于通过立法设定价值目标、伦理红线和基本原则的监管方式，技术工具赋能监管的方式代表了“工具理性”的延伸。其核心思想是利用隐私计算、联邦学习、可解释、算法审计、系统测评等人工智能自身的技术来规制其风险，实现“以子之矛，攻子之盾”。

苏宇教授在《探索与争鸣》2025 年第 3 期发表的《从算法解释到系统测评——人工智能法治的信息工具变革》一文中，主张将“系统测评”从自发的市场行为塑造为一种制度化的治理工具，进而纳入算法治理的制度工具箱，并使之在法治轨道内运行，这可以在一定程度上缓解算法

解释存在的计算量、随机性与理解力等方面的局限。[①]在人工智能治理的法律框架与具体方法方面，斯坦福大学《2024 年人工智能指数报告》提出了以“算法解释”为核心的治理理念。该观点主张，通过对作为技术基础架构之一的算法进行规制，实现“算法透明”与“算法可信赖”，从而防范和管控人工智能技术可能带来的风险。这一路径目前已成为众多学者关注的重点与研究的方向，并获得了相当数量支持者的认同。[②]“系统测评”的观点认为，应将人工智能和算法治理的焦点从“解释”转向“测评”，评价人工智能模型或应用的状态与性能。“系统测评”作为一种结构化的评估机制，凭借其全面性、适应性和制度兼容性，从局部解释走向系统评估，从静态透明转向动态测评，在应对复杂风险、适配多元性能需求以及融入法治框架方面展现出显著优势，填补了算法解释在大模型时代的能力局限。

“系统测评”的本质是人类对模型进行监督的技术方案，[③]在计算机科学领域常被视为产品检验的工序，是评估人工智能系统功能和性能的关键流程。当其成为人工智能治理的核心规范后，便不再只是计算机学科的技术概念，而是直接输出治理制度。系统测评依赖技术手段进行基本运作、风险规避并降低合规成本，监管者也必须跳脱传统的人工监管模式，利用技术手段进行同步的监管和监测。然而，在当前人工智能立法尚未形成闭环的情形下，必然存在法治的刚性要求与“系统测评”技术工具不成熟之间的根本性矛盾——前者要求即时、明确和可诉，后者仍处于可变的理论阶段。因此，本文在肯定“系统测评”为我国人工智能治理提供新方法论的同时，主张应回归法治的逻辑链路，审慎辨析“系统测评”作为技术治理手段在人工智能法治中的角色，寻求“系统测评”与“法治”之间的平衡，既避免对人工智能技术环节施加过度的法律规制，又防范其陷入无序发展状态。

人工智能“系统测评”的治理优势

面对大模型时代算法治理范式的根本性转变，尽管算法解释在司法举证、个案审查等局部场景中仍具不可替代的价值，但其高昂的计算成本、内在随机性与高理解门槛严重限制了其在大模型时代作为核心治理工具的可行性。在此背景下，“系统测评”凭借对模型性能、风险与合规性的全面评估能力，有望成为人工智能治理体系中的新型基础设施与信息工具。

（一）对算法解释机制的突破

“系统测评”采用标准化与科学化的方法，将原本模糊的“系统表现”转化为可量化、可比较、可验证的客观数据与技术结论。它把诸如“是否安全”“是否公平”“是否合规”等定性问题，转变为“在特定测评集上得分多少”“偏差率是否低于某一阈值”等具体定量事实。[④]在大模型时代的算法治理中，“系统测评”凭借全面且动态的评估机制展现出重要价值，能够有效应对算法风险的扩散与模型安全需求的提升。

随着生成式人工智能的迅猛发展，算法风险已从传统的歧视、共谋、“大数据杀熟”等扩展至国家安全、社会秩序、信息安全等关键领域。面对深度伪造内容生成、篡改数据集等新型风险，传统的算法解释机制已显得力不从心。而“系统测评”通过多维度、多场景的测试方式，能够全面评估模型在抵御恶意攻击、防范数据污染和处理对抗样本时的能力，系统性识别其潜在漏洞与风险点。与算法解释仅提供静态、局部决策逻辑的特点不同，“系统测评”能够模拟真实环境中的异常与攻击行为，动态检验模型的鲁棒性与安全性，从而为风险识别与防范奠定更加全面和

坚实的信息基础。值得注意的是，“系统测评”并不排斥算法解释，而是可将其纳入测评框架中，如借助模型的自我解释机制以增强测评深度，通过归因分析方法检验测试项目本身的合理性等。

（二）实现“可信任”的治理转变

人工智能市场的健康发展高度依赖信任，人工智能信任的建构需要以法律来调整人类与人工智能之间的交流与互动行为。用户需要信任服务提供商，企业需要信任所采购的模型，投资者也需要信任其所投入的技术。在技术层面，由于大模型参数量庞大、结构复杂、涌现能力强，实现完全的解释透明度存在困难，加之专业解释与公众理解之间存在差距，单纯依靠算法解释难以构建可靠信任。[⑤]“系统测评”并不单纯依赖算法透明度或决策过程解释，而是能够精准契合不同应用场景对模型性能的关键需求，并在多元治理目标之间实现有效协调与平衡。公正和权威的“系统测评”机制，类似于市场的“信用评级”和“质量认证”，借助信号传递的作用，帮助用户识别优质产品与服务，同时将不合规、高风险的系统排除出市场。它不仅降低了市场中的信息不对称和交易成本，还为构建可信、健康的人工智能生态提供了基础设施支撑。

在具体应用中，“系统测评”通过建立针对性测试基准与科学的评价体系，覆盖准确率、响应速度、逻辑一致性、伦理对齐度等多类指标，对模型在特定任务中的表现进行客观、量化评估，从而为模型选型、政府采购、标准制定及法律归责提供依据。该系统强调结果的可信性优于过程的透明性，更契合当前复杂人工智能系统的实际运行特点。此外，“系统测评”还能够对模型在伦理符合性、社会责任及文化适配性等多重治理目标方面的表现开展综合评估，并检验模型抵御成员推理攻击、模型窃取等安全威胁的能力，有助于在实践中贯彻“统筹发展与安全”的治理原则。同时，依托标准化、可复现的测试流程，“系统测评”输出客观、可量化的结果，具有良好的可操作性与可比性，展现出比算法解释更强的适应性与制度整合能力。它不仅与传统治理工具形成互补，也更容易融入现有制度体系。

人工智能“系统测评”引发的治理风险

苏宇教授自己也意识到，“系统测评”的原理与实践尚未成熟，在制度化过程中可能面临训练针对性、基准科学性、利益关联性等挑战。基于此，本文认为，即便是可以通过合理界定测评的应用场景，科学建立测评基准筛选机制、测评基准质量管理机制、测评过程公正保障机制来一定程度上克服这些挑战，“系统测评”引发的风险依旧不容忽视。这是因为，“系统测评”基准的不明晰、不统一或不完善，可能会使“系统测评”这项原本旨在降低不确定性的“信息工具”本身产生新的风险和不确定性，甚至可能影响整个人工智能治理体系。

（一）公平性失真风险

2017 年，Galhotra 等人首次定义了软件公平和歧视，开发了一种基于测试的方法来衡量软件是否涉嫌歧视以及歧视的程度，并重点关注歧视行为中的因果关系。[⑥]在人工智能时代，公平性是模型重要的非功能性需求之一，涉及数据集公平、算法公平、模型公平等关键问题，这需要迁移正确性测试、鲁棒性测试等伦理技术作为支撑。

OpenCompass、LMBench、SuperCLUE 等测评体系可以检测偏差，而这些检测模型公平性的工具包，其研究开发的主体不同，背后的利益关系可能会影响测评结果的公平性，这种公平性更多表现为测评结果的可靠性和公信力。当存在利益关系时，“系统测评”的公平性是非常值得关

注的，利益因素的影响会渗透到“系统测评”的各个环节，甚至“系统测评”结果会产生“寻租空间”，测评机构和人员拥有过大的、未被有效约束的自由裁量权。某些企业可能通过非技术手段影响测评结果，人为参与到“系统测评”的过程中，或者测评双方建立合作关系，最终影响“系统测评”的结果，导致劣币驱逐良币，破坏市场公平竞争。在测评市场质量参差不齐时，“商测一体”乱象常表现为测评方与开发商达成某种利益关联。⑦

这一点，苏宇教授在文中有所提及，然而我们还需进一步意识到，“系统测评”潜在的公平性失真风险，更深刻地表现为其对现实中结构性不公的固化甚至强化，公平性测评本身成为固化不公的制度性工具。若模型所使用的训练数据集未能充分反映真实世界的复杂性与多样性，或所采用的公平性指标本身存在设计缺陷，则测评结果很可能在表面上呈现“公平”，实则掩盖甚至加剧了对特定弱势群体的系统性排斥。例如，使用定制化数据集训练招聘大模型，反而可能在实践中持续削弱某一性别、种族或年龄群体候选人的机会。这种技术悖论不仅复制和深化了既有的社会偏见，更因其携带“科学测评”的权威证明而具备更强的隐蔽性与正当性。其结果是将歧视嵌入决策系统的底层，并以自动化、标准化之名获得豁免。实际上，无论是在立法层面构建规范体系，还是在技术层面设定相关标准，都应融入伦理治理的思维。人工智能技术本身关联诸多伦理与社会问题，如何将这类价值诉求嵌入具体的技术标准之中，并确保其在实践应用中有效落实，仍然是一个亟待深入探索和解决的问题。这也是为应对安全与创新之间日益复杂的平衡挑战，立法与监管机构正面临的一种新型治理情境。⑧

（二）性能与能力误导风险

实践中，针对生成式人工智能基础模型与领域模型的测评活动已广泛开展。然而，可观察到多数测评报告对大模型的评估结果呈现出显著多样性。常见的情况是，同一数据集可能因为评测策略的不同而导致模型得分出现巨大差异。例如，prompt 构建时多一个回车或者冒号的轻微区别，都会导致不同的测评结果。这种现象也与模型本身的动态特性密切相关，现行“系统测评”基准通常仅覆盖已知、常见的攻击模式，而难以应对更复杂、新颖的攻击手段，从而导致测评结果出现差异甚至失真。此时，“系统测评”报告所呈现的模型性能便值得质疑。若性能测评基准被模型开发方“过度拟合”，抑或通过提示过程来引导模型给出结果，则模型可在测评中表现出优异的性能，⑨但在真实应用场景中难以灵活响应动态多变的需求。例如，某自动驾驶系统若在测评中未考虑某种极端天气与传感器干扰的复合效应，在实际道路上可能引发严重事故。这不仅会导致开发企业蒙受巨大的资源损失，更使其陷入合规困境。企业投入大量成本通过某次测评，并不能保证下一次仍能通过，致使企业无法形成稳定的规范预期，显著增加合规成本与法律风险。该现象背离了法律为市场提供稳定预期的基本功能。⑩更有甚者，一旦发生造成人身或财产损害的事故，“系统测评”报告还可能被用作免责的借口，企业可能凭借“不可预见的意外”来推卸责任。

“系统测评”作为人工智能法治治理的信息工具被提出，具有一定的前瞻性，但其发展在一定程度上脱离了当前人工智能法治的实践。若进一步作为正式治理工具运用于法治实践，则更显勉强，难以符合规范性与权威性的要求。苏宇教授在文中指出，“系统测评”结果可制度化的应用场景主要包括：作为执法和司法程序中的证据或参考材料、作为合同义务履行与否的判断标准、作为违约或侵权行为的佐证等。这些应用实际会直接影响当事人的权利义务配置。然而，由于“系统测评”本身难以完全规避在性能与能力表征上的误导风险，其能否承担此类重要职能，究竟应定位为“辅助性”“过程性”还是“决定性”工具，将深刻影响法治的公信力与合法性基础。甚至，

被处罚方完全可能对测评结果的可靠性提出质疑，进而引发新的法律争议。这不仅会增加执法成本，也可能最终损害监管机构的权威性。法治以“定分止争”为目标，而不具稳定性的测评机制本身却可能成为争议之源，与法治的目的相冲突。此外，测评可解释性可以增强模型的可信度和可接受性，[11]但“系统测评”高度依赖专业人员进行操作与解读，即便是专业人士，也往往难以完全揭示其内在推理过程。因此，对缺乏计算机常识和相应知识背景的普通用户而言，此类测评仍难以助其理解人工智能系统的运行逻辑。

（三）技术垄断风险

在人工智能快速发展的时代背景下，人工智能法治建设不应仅着眼于如何治理人工智能本身，更应在制度构建过程中统筹多元目标。苏宇教授在文中指出，“系统测评”因其能够基于多重目标进行综合考量，更加契合不同主体对大模型技术发展的特殊期待，有望成为人工智能法治体系中的一项基础性制度工具。作为基础性制度工具，“系统测评”必须能够客观、有针对性地评估人工智能模型或应用的实际状态与性能，其基准的公正性至关重要。尽管苏宇教授提出了测评基准的筛选形成机制及过程公正保障机制，但仍存在潜在风险。例如，随着“系统测评”走向市场化并逐渐集中化，若某一测评基准被确立为“官方”或“唯一”标准，所有研发力量将倾向于在该特定基准上优化模型性能。这种导向可能导致技术路径趋同，使那些未被纳入当前基准却具有潜力的创新技术逐渐被边缘化。大型科技公司凭借其雄厚的资源，能够针对主流测评基准进行定向优化，从而持续占据评测榜单前列，进一步巩固其市场垄断地位，甚至使某些测评基准成为实际上的“权威标准”。如 Google、Stanford、Microsoft 等科技巨头与高校发布的 MMLU、HELM、BIG-Bench、C-Eval 等测评基准，已成为国际市场的主流评价基准。一旦测评基准被少数主体垄断，其筛选机制与公正性保障机制可能面临失效风险，甚至出现测评标准被反向“定制”以服务于既有垄断利益的情形。

此外，苏宇教授在文中提到可通过监管部门、第三方专家等外部力量对“系统测评”基准进行监督与优化，该机制在实际操作中也可能产生新的人为干预空间，反而为某些企业实施垄断协议、滥用市场支配地位提供制度缝隙。例如，通过定制化指标、封闭数据集或非透明评估流程，边缘化新兴技术路线，初创公司及科研机构即便具备创新实力，也难以在“被操纵”的评测框架中获得公正评价。

人工智能“系统测评”标准若被少数巨头或国家垄断，将带来深远和系统性的危害。这种垄断不仅会扼杀技术创新与市场竞争，导致技术路线单一化、研发资源高度集中，使广大中小企业和后发国家难以参与竞争，更会形成“技术霸权”，垄断方通过制定不公平的标准许可政策、收取高额专利费用或实施技术封锁，极大增加全球其他主体的应用与发展成本。[12]在伦理与安全层面，垄断方可能将自身价值观嵌入标准体系，引发算法偏见加剧、数据滥用等风险，威胁全球数字治理的公平性与安全性。更重要的是，人工智能作为未来经济社会的基础性技术，其“系统测评”标准的垄断将深度绑定全球产业链，强化技术依赖格局，削弱其他国家特别是发展中国家的技术主权与产业安全，最终阻碍全球人工智能生态的健康发展与包容性进步。

人工智能“系统测评”嵌入法治的路径

上述与苏宇教授提出的观点进行商榷，并非意在否定“系统测评”作为人工智能治理方式的

重要价值。相反，我们认可“系统测评”的良好治理潜力。因此，需要进一步明确的是，一旦将其从市场自发行为提升为具有强制力的法律行为，就必须通过体制与机制的审慎安排，最大限度地降低因其内在风险可能引发的规范冲突与制度成本，确保“系统测评”这一技术治理工具能够在人工智能法治进程中发挥积极、稳健的作用。

（一）厘清“系统测评”在法治中的角色

制度变迁通常是渐进式的，而非断裂的。正式规则可以通过立法等方式迅速发生改变，但文化、习俗、惯例等非正式约束的变化则相对缓慢，这常常导致“正式规则与非正式约束之间的紧张关系”，从而影响制度变迁的实际效果。[13]“系统测评”由市场自发行为上升为法律治理工具，是人工智能治理体系化进程中的关键一步。在这一转型过程中，如何尽可能减少其带来的负面影响与适应性成本，成为制度设计者必须面对的现实问题。“系统测评”最初源于市场内生的检验机制，企业通过自愿性测评获取信誉背书、提升产品竞争力，其优势在于灵活、高效和紧跟技术迭代。一旦“系统测评”作为人工智能治理的信息工具，单纯依赖市场自律将逐渐显露出局限性，评测结果反而致使公信力受损，甚至引发新的市场失灵与技术风险。不同行业对法律规制的需求各不相同。统一立法模式往往难以精准适配这些差异，反而会加剧规则与场景错配的法律适用风险。[14]

基于此，人工智能法治的信息工具变革的关键第一步是明确“系统测评”在人工智能法治框架中的定位。法律治理在这一进程中不应寻求对市场机制的全面替代，而应侧重于“底线保障”的角色，为人工智能的关键应用场景设定最低限度的安全、伦理与合规标准。一方面，要为高风险领域提供明确的行为预期和稳定的制度环境，另一方面，要为技术试错与市场自我调节留出充分空间。“系统测评”的法治化本质上是将一部分经过实践检验、具有共识基础的市场实践，通过立法程序转化为具有普遍约束力的规则体系。其目标并非扼杀创新，而是通过建立可信统一的测评规范，引导技术向上向善。基于这一认识，“系统测评”的法治化进程不宜一刀切，也不能毕其功于一役，而更应采取渐进式、分类别、有梯度的推进策略。人工智能治理体系必须尊重该系统原生的市场灵活性与创新弹性，不能以牺牲行业活力为代价换取形式上的规范。

具体而言，可以依据风险等级对人工智能应用场景进行科学划分，实行差异化的测评引入机制。在自动驾驶、医疗诊断、公共安全、金融风控等高风险领域，应将“系统测评”设置为强制性法律义务，明确未通过测评不得上市或部署，并配套建立严格的监督与惩罚机制，以切实防范重大社会风险。在这些领域中，测评要求应侧重可解释性、公平性、安全性与可靠性，如通过对抗测试、可解释性分析、偏差检测等技术手段确保系统行为符合伦理与法律要求。在娱乐推荐、智能写作、辅助设计等低风险场景，则可继续保持测评的自愿性，积极鼓励通过行业标准、第三方认证、企业自律承诺等市场化机制推动治理优化。政府与监管机构可在这一领域中发挥引导与支持作用，如发布测评指南、推荐最佳实践、推动建立跨行业的测评标准共同体等，而非直接施加强制性约束。

此外，还可探索“监管沙盒”机制，允许企业在限定范围内对新兴技术进行测试验证，在控制潜在风险的同时积累有效的测评经验。分类推进的策略不仅有助于降低企业的规范成本，也能够提高监管资源的配置效率，使有限的执法力量集中于真正具有社会风险的领域。[15]同时，这一渐进的法治化路径还有助于培育健康、开放的测评生态，推动市场自我优化与法律外部约束的良性互动。最终，“系统测评”作为连接技术与治理的重要桥梁，应在法治与市场之间找到动态平衡，为人工智能的负责任发展提供持续且适配的制度保障。

（二）构建具有“软法”效果的技术标准

从“市场自律”向“法治工具”的转型是当前人工智能治理的一个重要趋势。我国在相关领域已存在明确的法律实践，“系统测评”并非首例。例如，《人工智能生成合成内容标识办法》规定，人工智能生成内容必须通过合规性标识测评后方可公开传播；同样，网络安全等级保护制度也明确将“系统测评”作为法定义务，对未通过测评的系统施以高额罚款甚至业务关停等处罚。这类制度设计的核心目标，是通过法律规范确立权威性与规范性，提升结果的可信度与可比性，为人工智能的治理提供制度性保障。然而，“系统测评”作为人工智能治理的信息工具，一个亟待商榷的问题是，“系统测评”是否已积累足够的实践经验，从而能够直接上升为具有强制约束力的管理办法或规范性指南。

目前，人工智能技术仍处于“过高期望峰值期”，[16]测评方法、指标体系和评估框架本身也在不断演进。若在实践尚未成熟、共识尚未形成的情况下，急于将现有测评体系全面制度化，可能会带来一系列风险。从技术哲学视角审视，人工智能法治必须深植于技术发展的客观规律之中，尊重技术发展的内在规律，防止出现制度过早介入导致的负外部性问题。基于此，人工智能治理应更加侧重“立法论”层面的审慎研究，而非急于寻求形式上的制度化。技术标准作为具有实际规范效力的“软法”工具，[17]因其灵活性高、修正机制灵敏，往往比刚性立法更能适应快速变迁的技术环境。因此，当前的迫切任务并非将测评体系直接纳入法律强制范畴，而是优先推动建立科学、透明且行业广泛认可的权威性技术标准，以实现促进技术发展与防范风险之间的最佳平衡。从国家治理的角度来看，技术标准正成为一种重要的人工智能规制工具。[18]2020 年 7 月，国家标准化管理委员会等五部门联合发布《国家新一代人工智能标准体系建设指南》，该文件对人工智能标准化工作进行了顶层设计，明确提出要“形成标准引领人工智能产业全面规范化发展的新格局”。

一方面，标准必须具备清晰且可量化的指标维度。在伦理方面，应明确公平性、可解释性等核心指标，并尽可能采用数学定义和统计度量，避免主观判断。公平性可针对不同人群的算法偏差进行度量，可解释性需评估决策过程的可追溯程度。在性能方面，需涵盖鲁棒性、准确性、效率等关键参数，并依据应用场景设定阈值。鲁棒性指系统抗干扰能力，准确性需在多样本集下综合衡量，效率则指有效算力和生成的结果与预期偏差。在社会影响方面，需引入隐私保护强度、系统适应性等评估要素，如数据匿名化与泄露风险指标，以及多环境下的运行稳定性指标等。

另一方面，标准应具备良好的可实施性与可重复性。测评方法须附有详细的操作指引，包括测试环境配置、数据采样要求、测评程序流程和结果解释准则，确保不同测评机构执行同一标准能够得出可比结论。尤其在高风险场景中，测评应具备唯一性，不同的数据集在相同场景下生成结果一致。鉴于人工智能技术迭代迅速，标准制定需设立定期修订机制，融入新的技术洞察和风险发现技术。同时，在医疗、化工、文化、金融与自动化控制领域，根据应用场景对测评要求和等级进行划分，避免因一刀切导致适用困难。

（三）确立独立透明的测评认证机制

为确保人工智能“系统测评”的公正性与权威性，必须建立一套独立、透明且可审计的认证机制。该机制应覆盖机构准入、利益监管、过程记录及结果复核等多个环节，形成贯穿测评全流程的约束与保障体系。

一是实施严格的第三方测评机构认证机制。可参照会计师事务所、信用评级机构等专业服务

领域的监管经验，对从事关键领域的人工智能进行测评，其结论可能用于执法或司法程序的机构，设立行政许可或强制性认证要求。认证标准应涵盖多个维度，包括机构的技术能力，如专用检测工具、基准数据集和专业人员配置；独立性保障，如股权结构、治理架构和业务来源的审查，财务透明度与稳定性，内部质量控制流程的完备性，以及机构及其核心技术人员的专业声誉和历史表现。通过设置明确的准入门槛，从源头上筛选出具备相应资质和公信力的测评主体。

二是构建严密的利益冲突防范与审查机制。强制要求所有测评机构全面披露其与受测模型开发方、产品供应商、投资方以及其他利益相关方之间的任何现有或潜在利益关联。建立统一的测评机构名录库，实行动态管理，并设立针对重大或高风险测评项目的随机指派机制，最大限度避免选择与受测方存在利益关联的机构执行测评。每一份正式出具的测评报告必须附带一份具结式的利益冲突声明，由机构法定代表人及项目负责人签署，明确声明已履行必要的审查程序并承诺不存在可能影响测评结论公正性的任何利益关系。

三是推行测评过程与结果的有限公开及可审计制度。测评机构必须完整记录并安全保存关键的测评操作流程及所使用的测试数据集概要、参数配置、判定规则与依据，形成不可篡改的、可追溯的测评日志。此举并非要求公开模型本身的核心知识产权，而是确保测评行为本身的规范性和可重现性。须明确授权监管部门和司法机关在必要时，如应对争议、投诉或进行例行抽查，有权调取并审计这些后台日志，对测评活动的真实性与规范性进行事后复核。这种可审计性是对测评机构独立性和专业性的持续监督与有力威慑。

四是建立通畅、高效且权威的申诉与复核机制。为确保人工智能测评体系的公正性和公信力，必须为企业提供明确和制度化的异议救济途径。任何对测评结论有异议的企业，均可在规定时限内，向指定的监管机构或行业自律组织提交书面申诉，阐明申诉理由并附相关证据。受理机构应在接到申诉后及时启动复核程序，委托另一家具备资质、无利益冲突的已认证测评机构，对原测评过程及结论进行独立复核。复核工作应基于原测试环境与数据备份进行复现验证，重点审查测评流程的合规性、数据使用的适当性、判定依据的科学性及结论的可靠性。特别是在涉及生成式人工智能模型的测评中，除常规复核项目外，还应将模型对问题的思考与生成过程作为专项评估内容。可设计特定任务要求模型展示其推理链或生成依据，由复核机构对其逻辑一致性、合理性、可解释性以及生成过程的稳健性进行审查，以此更全面、深入地评估模型能力，确保复核结论的科学与严谨。复核机构出具的复核报告应详细载明复核方法与结果，并作为裁定原测评结论是否予以维持、修正或撤销的核心依据。该裁定具有终局效力，双方应予执行。

结语

“系统测评”作为人工智能治理体系中的新型信息工具，其出现与发展响应了大模型时代敏捷与协同治理的制度需求。本文与苏宇教授商榷，意图并非否定“系统测评”本身的重要价值，而是试图在其被推向法治前台的背景下，冷静审视其作为治理手段的内在限度和制度风险。“系统测评”虽具备从“解释”到“测评”、从“透明”到“信任”的范式突破潜力，能够系统、量化地评估模型性能与合规状态，但当下仍处于技术快速演进、标准尚未统一、实践尚未成熟的阶段，若在缺乏充分论证和制度配套的情况下，仓促将其确立为法律治理工具，不仅难以实现有效规制，还可能因公平性失真、性能结果误导和技术垄断等问题，损害法治的权威性和稳

定性。

基于这一判断，本文主张回归法治的基本逻辑，将“系统测评”的法治化进程理解为一个结构性、分阶段、多层次的制度构建问题。首先，必须明确“系统测评”在治理体系中的功能定位，其角色应是“补充而非替代”“保障而非主导”。法律治理应侧重于为高风险场景设置底线要求，而在低风险领域要充分尊重市场自律与技术弹性，通过实施分类分级机制避免“一刀切”带来的制度僵化。其次，在当前技术条件下，不宜急于将“系统测评”全面法治化，相反应优先发挥技术标准作为“软法”的引导作用，通过制定科学、开放、可操作的标准体系，为测评实践提供可靠依据，并为未来立法积累经验。最后，要建立独立、透明、可审计的测评认证机制，包括严格的主体准入、严密的利益冲突防范、过程可审计性及申诉复核渠道，从程序上保障“系统测评”的公正性与权威性，使其在法治框架内稳健运行。

总之，“系统测评”的真正价值不在于其是否成为制度规范，而在于能否以可靠、公平和开放的方式，为人工智能治理提供可持续的信息基础设施与制度信任。只有将技术工具与法治原则系统结合，在创新与规范、市场与监管、效率与公平之间寻求动态平衡，才能使其真正赋能人工智能的向善发展。未来的研究可进一步聚焦测评标准的具体构建、跨场景差异化治理策略的设计以及国际规则互认等问题，推动形成更具包容性和适应性的治理范式，为人工智能的未来治理提供扎实的理论根基与可行的实践路径。

注释：

① 苏宇：《从算法解释到系统测评——人工智能法治的信息工具变革》，《探索与争鸣》2025 年第 3 期。

② 安晋城：《算法透明层次论》，《法学研究》2023 年第 2 期；苏宇：《算法解释制度的体系化构建》，《东方法学》2024 年第 1 期；周翔：《算法可解释性：一个技术概念的规范研究价值》，《比较法研究》2023 年第 3 期；张欣：《算法解释权与算法治理路径研究》，《中外法学》2019 年第 6 期。

③ Yupeng Chang，Xu Wang，et al.，“A Survey on Evaluation of Large Language Models，” *ACM Transactions on Intelligent Systems and Technology*，vol.15，no.3，2024.

④ 倪俊杰、刘宗凡等：《测评系统揭秘——从发展到智能应用》，《中国信息技术教育》2022 年第 21 期。

⑤ 郭小东：《从“可解释”到“可信任”：人工智能治理的逻辑重构》，《北京工业大学学报》（社会科学版）2025 年第 6 期。

⑥ Sainyam Galhotra，Yuriy Brun，Alexandra Meliou，“Fairness Testing: Testing Software for Discrimination，” Meeting of the European Software Engineering Conference，2017.

⑦ 杨淑馨、尹一如：《貌似公允的“测评”可能是“定制”》，《新华每日电讯》2024 年 12 月 6 日，第 05 版。

⑧ 张龑：《中国人工智能立法的价值基础与伦理治理模式》，《探索与争鸣》2024 年第 10 期。

⑨ Rylan Schaeffer，“Pretraining on the Test Set Is All You Need，” https://doi.org/10.48550/arXiv.2309.08632.

⑩ 张守文：《论在法治轨道上推动经济发展》，《法学论坛》2024 年第 3 期。

⑪ 许志伟、李海龙等：《AIGC 大模型测评综述：使能技术、安全隐患和应对》，《计算机科学与探索》2024 年第 9 期。

⑫ 曾繁华、陈建军：《技术垄断竞争问题研究回顾及评析》，《财务与金融》2015 年第 1 期。

⑬ 道格拉斯·C. 诺思：《制度、制度变迁与经济绩效》，杭行译，上海：格致出版社，2008 年。

⑭ 付新华：《人工智能统一立法宜缓行》，《东方法学》2025 年第 3 期。

⑮ 丁元竹：《人工智能技术的潜在社会风险及治理体制机制研究》，《行政管理改革》2025 年第 7 期。

⑯ Stamford，“Gartner 2024 Hype Cycle for Emerging Technologies Highlights Developer Productivity，Total Experience，AI and Security，” https://www.gartner.com/en/newsroom/press-releases/2024-08-21-gartner-2024-hype-cycle-for-emerging-technologies-highlights-developer-productivity-total-experience-ai-and-security?utm_source=chatgpt.com.

⑰ 张欣：《我国人工智能技术标准的治理效能、路径反思与因应之道》，《中国法律评论》2021 年第 5 期。

⑱ 张涛：《通过技术标准规制人工智能：基于合作规制的法理》，《比较法研究》2025 年第 4 期。

编辑　孙冠豪