



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

清华大学CIDEG研究项目

研究报告（结题）

人工智能算法影响评价 体系构建与应用实践

2024 年 3 月

委托方：清华大学产业发展与环境治理研究中心

课题负责人：

贾开 上海交通大学国际与公共事务学院长聘副教授

课题组成员：

赵静、傅宏宇、陈少威、元圣平、杨欣宇、朱韵涵

清华大学产业发展与环境治理研究中心

地址：北京市海淀区清华大学公共管理学院615

邮编：100084

电话：(86-10) 6277 2497

传真：(86-10) 6277 2497

邮箱：cideg@tsinghua.edu.cn

主页：<http://www.cideg.tsinghua.edu.cn>

目录

1. 人工智能算法影响评价的背景：界定治理问题	4
1.1. 算法歧视治理风险	5
1.2. 算法公平治理风险	8
1.3. 算法安全治理风险	10
1.4. 本章总结	12
2. 人工智能算法影响评价的内涵：描述治理实践	13
2.1. 人工智能算法影响评价的制度内涵及其治理定位	13
2.1.1. 算法影响评价的制度内涵	13
2.1.2. 算法审计的制度内涵	15
2.1.3. 算法影响评价与算法审计的共性与差异	16
2.2. 人工智能算法影响评价的三种模式：环境影响评价、数据保护影响评估、列表清单	17
2.2.1. 环境影响评价视角下的算法影响评价	17
2.2.2. 数据保护影响评估视角下的算法影响评价	20
2.2.3. 问题列表清单评估视角下的算法影响评估	22
2.2.4. 不同模式的对比分析	26
2.3. 人工智能算法影响评价的国别特征：欧盟、美国、英国、加拿大与中国	26
2.3.1. 欧盟	26
2.3.2. 加拿大	31
2.3.3. 美国	35
2.3.4. 英国	41
2.3.5. 中国	47
3. 人工智能算法影响评价的治理逻辑：监管者与被监管者的合作	51
3.1. 协同治理（Collaborative Governance）视角下的人工智能算法影响评价	52
3.2. 管理主义（Managerialism）视角下的人工智能算法影响评价	55
3.3. 制度同化（Institutional Isomorphism）视角下的人工智能算法影响评价	60
4. 人工智能算法影响评价的未来改革：形成政策建议	61
4.1. 人工智能算法影响评价的制度要求：开放性、合作性、责任性	62
4.1.1. 开放性	62
4.1.2. 合作性	63
4.1.3. 责任性	64
4.2. 通用人工智能算法影响评价的体系框架：理论基础、层次结构、流程指标	65
4.2.1. 理论基础：心理计量学与机器行为学的理论融合创新	65
4.2.2. 层次结构：技术、组织、社会	71
4.3. 通用人工智能算法影响评价的政策建议	错误!未定义书签。

人工智能算法影响评价（algorithm impact assessment, AIA）作为一种治理机制已经被普遍性地纳入各国算法治理的框架体系之下，但其本身的制度定位与制度逻辑却尚未得到充分讨论。较为典型的反思性问题例如：人工智能算法治理的何种关键特征，使得我们需要纳入人工智能算法影响评价（即人工智能算法影响评价的必要性问题）？与往往被视为其前身的环境影响评估、数据隐私影响评估相比，人工智能算法影响评价体现了更多的继承和连续性，还是差异性（即人工智能算法影响评价的特殊性问题）？人工智能算法影响评价治理绩效的发挥，受到何种因素的影响（即人工智能算法影响评价的局限性问题）？

对于这些问题的探索性回答，构成了本报告的主要内容。

1. 人工智能算法影响评价的背景：界定治理问题

人工智能算法治理的特殊性，是开启人工智能算法影响评价制度建设的起点。从基本内涵来讲，人工智能算法影响评价是指技术研发者、部署者、应用者需要预期、测试、评估技术创新和应用过程中所可能带来的多维度影响，并对结果进行记录、公开或向监管者汇报，进而在积累算法治理经验与知识的同时，对其研发应用过程进行调整以避免负面伤害。作为一种监管手段，人工智能算法影响评价事实上将监管环节延伸至企业内部，要求企业基于评价结果以改变其技术创新和应用过程，而这一思路在很大程度上与传统的监管思路存在本质不同。

在经典监管理论看来，监管的目的旨在对市场失灵现象作出回应，以避免出现权益侵害风险。换言之，监管是直接面向可能出现的风险而展开，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。但影响评价制度却“反其道而行之”，即其更关注过程，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”边界基础上，那么影响评价制度则打破了这一“二分法”，而直接切入进企业内部生产管理过程，以求对风险作出回应。由此带来的关键问题便在于，**为什么我们需要跳出经典监管理论范畴而将监管“触手”延伸至企业内部，为什么需要施行影响评价制度？**

对以上问题的一个简洁回答是，在多类型的市场行为中，存在大量的“未知

风险”：这既包括因监管者与被监管者之间存在信息不对称而导致前者“未知”的风险，也包括监管者、被监管者都不知道的“未知”风险（即“共同无知”）。环境影响评估、隐私影响评估都属典型案例。接下来本报告通过三个算法治理的典型案例对此问题作出更详细的回答。

1.1. 算法歧视治理风险

算法歧视风险可能是最被频繁提及的算法治理风险之一，这不仅是因为其可能是最容易被观察到（或感知到），也源于算法（特别是人工智能算法）的基本技术与应用逻辑所致。换言之，算法正在被大规模应用于各个领域的决策场景（例如招聘简历筛选、信用评分评级、公共服务审核批准、犯罪风险评估、警力资源部署等等），而这些决策场景往往都跟人们的切身利益紧密相关，并尤其与一般意义的社会公平以及具体意义上的少数或弱势群体的特殊属性紧密相关，后者又表现为性别、种族、肤色、身体、收入、地域、行为偏好等具体指标。在此背景下，当人们普遍关心决策歧视问题时，算法作为决策功能的应用便自然与此紧密关联了起来。

一般而言，从算法作为决策功能的基本逻辑和流程来看，引致算法歧视风险的原因可被主要总结为三点。第一，因算法决策目标选择的主观性而带来的偏差与歧视。考虑到决策目标往往都较为宽泛，因而在具体算法设计时往往需要将其转化为更具体指标，而这一转化过程依赖算法设计者、部署应用者的主观性，由此便可能因人为主观意图而导致歧视结果。例如招聘决策的目标是寻找到“优秀”人才，但“优秀”这一决策目标过于宽泛，其不得不转化为学历、资历、外表等维度的具体指标，而究竟选择何种指标来测度、衡量“优秀”（或者是不同指标相对权重大小的设置）便将体现算法设计者或部署应用者的主观性。即使算法设计者或部署应用者并没有“人为、故意”的设置歧视性变量，主观性选择本身始终都会面临是否“公平”的质疑与拷问。第二，在具体指标既定的前提下，采纳何种数据集以对该指标进行测度将是引入歧视的第二个关键环节。数据集往往需要人为标注、整理、清洗以最终服务于算法模型的训练与优化，但在每一个环节都将可能引入主观性偏差或歧视。即使基于“大数定律”可能削弱或消解个别数据标注、整理、清洗时的人为偏差，但数据集作为对现实复杂性的“抽样”，其

是否能够完全“表征”现实仍然存疑，而数据集相对于现实的任何偏差都将自然导致“歧视”结果。第三，在指标既定、数据集既定的前提下，究竟采用何种模型架构也同样存在引入“歧视”的可能性。尽管可以通过特定技术指标来比较不同模型的公平性程度，但算法永远都可能存在“漏洞”的本质属性仍然使得我们不可能在绝对意义上精确比较不同算法架构的公平性强弱水平，这也使得任何选择都仍然是具有主观性的，并因此可能是“歧视”的。

值得注意的是，上述三点原因的总结还只是聚焦在技术层面，即指出人工智能算法技术实现过程中可能引入的歧视因素，但这明显还是不完整的，上述分析尚没有对人工智能算法歧视问题的特殊性作出概括，而后者还需要结合算法应用的制度环境、社会环境才能得到更充分的解释。也正是在此基础上，我们才能更清楚地理解算法影响评估制度的必要性。具体而言，这又体现在以下两个方面。

一方面，以“过错原则”为基础而建立的现行“反歧视”法律框架，不仅要求明确权益侵害的现实，还要求明确导致权益侵害“过错”的因果链条，而这二者在算法决策的应用环境下都将面临新的不确定挑战。

就前者而言，与非算法决策环境相比，基于概率计算的人工智能算法决策结果具有更高不确定性，导致人们事实上很难提前预知每一次特定算法决策结果是否会产生歧视性结果，并因此增加了确认权益侵害现实的难度。在非算法决策环境下，导致歧视结果的原因往往是可解释的、制度性的，并因此可在合理范畴内预期每一次以及一般性的决策结果；相比之下，算法决策逻辑是反过来，尽管决策结果的一般性特征是可预期的，但每一次决策结果却是不确定的，由此使得歧视结果的追踪、确认必须建立在不断且大量的测试、记录、统计基础上，这也自然增加了歧视性权益侵害确认的成本和难度。

不过这还不是算法决策带来的唯一挑战，歧视性权益侵害“过错”因果链条的追寻还将面临更大的不确定性困境。¹歧视问题的分析视角大致可被划分为两种：基于受害者视角以确认导致歧视的客观（例如肤色、种族、收入、出生地等）或主观条件（例如机会、认知、偏好等）并加以解决，基于过错者视角以分析决策者行为的中立性程度从而确保其不会加剧不平等结构。考虑到导致歧视性结果（或不平等结构）的原因是复杂而综合的，单靠法律规则本身难以对此作出全面

¹ Freeman, A. D. (1977). Legitimizing racial discrimination through antidiscrimination law: A critical review of Supreme Court doctrine. *Minn. L. Rev.*, 62, 1049.

应对，因此现行的反歧视法律框架主要采取了过错者视角而非受害者视角，其制度关键在于确保决策者行为的“中立性”而非直接面对、解决导致歧视问题的具体因素。在执行逻辑上，反歧视法律框架要求将决策者行为与被决策对象所处社会环境相独立，并通过比较可能的不同决策方案以判断决策者实际行为在影响受保护权益属性方面是否“中立”。当面对算法决策的新现象时，这一制度设计逻辑将至少面临算法决策的嵌入性与不可穷尽性这两种不确定性挑战：就嵌入性而言，与非算法决策相比，算法决策与社会环境的相互影响关系更为复杂，其基于数据训练的技术逻辑导致其很难被独立于社会环境而展开“中立性”分析；就不可穷尽性而言，算法决策技术方案的多样性使得我们很难证明某一特定决策算法的歧视性是最低的（或“中立性”是最强的），这也使得在算法歧视问题的治理中我们往往只能比较算法决策与非算法决策，而不能比较不同算法决策方案之间的结果差异，由此司法责任的界定标准也只能演化为判断算法决策者是否尽到了比较义务而非是否采用了“最中立”算法²。

另一方面，除了制度环境的讨论之外，产生歧视的社会环境在算法决策普及的情况下也将体现出新的治理复杂性。事实上，诸多研究已经揭示，歧视更多表现为社会建构问题而非生命特征属性问题，即歧视性结果并非源于不同个体生命特征属性的差异，而是社会生产生活进程的演化结果。³举例而言，对于黑色族裔人群而言，导致歧视性结果（或分化性结构）的真正原因是其成长、学习、工作的整个社会生产过程，而非种族或肤色这些表面的生命特征因素；换言之，在反事实假设情况下，即使将白色族裔人群置于当前黑色族裔人群的生存环境，其同样可能面临类似的歧视性结果。在将歧视视为社会建构问题的视角下，算法决策普及带来的新复杂性在于其将陷入或者简化，或者复杂化的“悖论”：对于不是建立在机器学习基础上的非人工智能算法而言，其要求将复杂现实抽象为具体变量的基本逻辑，极大简化了歧视作为社会建构问题的复杂性，将产生“缘木求鱼”的治理错位结果，即针对具体变量的歧视风险治理措施可能错误定位问题内涵而起不到救济作用；与之相对，建立在海量数据学习、训练基础上的人工智能算法，其可能把握住了导致歧视结果的社会过程的复合性、综合性影响，但却受限于可

² Oppenheimer, D. B. (1992). Negligent discrimination. *U. Pa. L. Rev.*, 141, 899.

³ Kohler-Hausmann, I. (2018). Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113, 1163.

解释问题而不能以简单变量的形式表达出来，从而也将影响歧视问题的追责与救济。换言之，歧视作为社会建构问题的治理复杂性在算法决策普及的背景下才真正表现了出来。

以上从制度、社会两个视角对歧视现象的分析解释了该问题的真正复杂性，而这也意味着仅聚焦结果并试图厘清过错因果链条的传统治理逻辑并不能对此复杂性作出充分回应。在此背景下，向决策过程的治理延伸必不可少：算法决策要解决的什么功能性问题，而衡量是否实现该功能的具体指标是什么，为什么选择该指标而非其他指标？测度该指标的具体数据集是按照何种标准选取的，其收集、维护、清洗过程是什么，其是否体现了现实决策环境的完整性？该算法决策方案与其他方案相比，在涉及利益相关方歧视性影响方面是否是“中立”的，这一结论是否经过测试和验证？诸如此类的问题事实上就构成了算法风险评估的基本制度内涵，而这也反过来说明了算法风险评估作为一种制度创新，在回应算法歧视治理问题方面的必要性和重要性。

1.2. 算法公平治理风险

算法公平与算法歧视治理问题紧密关联，但也存在一定差别。算法歧视治理侧重问题的发现与权益侵害的救济，而算法公平治理则更重视如何利用算法决策优势以改变社会分化结构。换言之，算法歧视治理是针对问题的回应，其目标在于实现算法决策影响的“中立”性；相比之下，算法公平治理则聚焦赋能受保护群体的积极行动，其目标是在差异化对待的基础上，利用算法来推动社会结构朝着更公平的方向演化。基于算法公平治理目标的这一界定，算法公平治理风险便可定义为对该目标的偏离，其既可能表现为算法决策应用没有达到优化社会结构的目的，也可能表现为算法决策的应用甚至恶化了当前已分化的社会结构。

与算法歧视治理类似，算法公平治理同样具有内在复杂性，而这又尤其体现在“公平”作为治理目标的内涵多样性，以及不同内涵之间的竞争性乃至冲突性。具体而言，这又表现为以下三点。

第一，算法公平作为治理目标在不同算法技术和应用环境下具有异质性，且伴随技术和业态演化而具有不断发展的动态性。在不同行业、不同领域，对于不

同利益相关方而言，算法公平的治理内涵都有所不同（例如就业信息推荐算法与新闻推荐算法的“公平”性要求是不同的）；同时，在技术发展应用的不同阶段，算法公平的治理内涵可能也有所不同（例如对于不同规模大小的电商平台，对其商品推荐算法的“公平”性要求也是不同的）。**第二**，算法公平作为治理目标与不同群体发展历史及其所处环境紧密相关，但算法设计者和部署者往往难以覆盖全过程并因此存在结构性缺失。“公平”概念在政治上可区分为“形式公平”和“实质公平”两类。相对于前者而言，后者更关心因历史及环境因素而造成的不同群体在事实上的差异地位，并因此要求“公平”体现为对这种差异地位的校准，此时“公平”治理要求也同时发展为“正义”要求；但对于算法设计者和部署者而言，其往往难以将历史情况、环境因素考虑在内，并因此不能达到“实质公平”（或“正义”）的治理要求。**第三**，算法公平作为治理目标既需要平衡“公平”与其他价值目标的内在张力，也需要平衡不同类型“公平”目标之间的内在张力。一方面，算法公平并非绝对意义上的一致性，否则便会丧失包括个性化服务在内的其他技术进步价值；另一方面，在算法技术实现层面，算法公平至少包含三层内涵：针对不同群体的算法计算结果应是一致（一致性要求）、不同群体被正确识别的概率应是一致的（正平衡性要求）、不同群体被错误识别的概率是一致的（负平衡性要求）。技术层面的不同内涵往往反映了差异化的价值目标，而它们在大多数情况下互斥并因此难以同时实现的困境也在事实上体现了不同“公平”目标之间的内在张力。

面对算法公平作为治理目标在上述三点所表现出的复杂性，一个有效的算法公平治理方案需要首先对治理目标本身的内涵进行两个层面的解构：第一层次需要讨论“公平”的来源，即“公平”来源于实质意义的要求（例如法律界定的规则，或者社会所公认、接受的准则），还是来源于结果效用的评估（例如让所有人的福利都不下降，或者让弱势群体得到改善），前者体现了“本体论”的公平，而后者是“结果论”公平；第二层次则需要讨论实现“公平”的过程，即“公平”是通过特定组织过程、流程来实现，还是通过外部干涉来实现，前者倾向于程序公平，而后者侧重分配公平。在这两个层次解构的基础上，算法公平治理在具体执行时便要求与算法设计、应用实践紧密结合，这又大致包括三个步骤：选择属性特征对对象分类、优化该特征以实现拟合、基于优化结果来作出决策或配置资

源。在这三个环节中，“公平”内涵的争议、选择都将在不同层面涌现，而算法设计、应用者也需要在类型化细分中进行选择，并对自己的选择作出辩护性解释。

无论是两个层面的解构，还是在算法应用三个步骤环节贯穿始终的“公平性”分析，其事实上都体现了算法影响评估的具体制度内涵，而这也证明了算法影响评估在回应算法公平治理要求方面的必要性和重要性。

1.3. 算法安全治理风险

此处要讨论的算法安全治理应被理解为“Safety”而非“Security”，其主要是指算法系统能否按照预期机制正常运转，因未意料到情况而影响系统并导致权益损害的风险治理问题。算法安全治理同样是算法治理领域被普遍讨论的重要议题，其甚至是先于算法歧视、算法公平等其他治理议题而需要考虑的首要问题，因其可被定义为算法技术服务或产品投入应用的先决条件与前提，典型案例例如自动驾驶算法、医疗识别与诊断算法等。围绕算法安全治理议题而展开讨论的关键有两点：算法安全治理是否具有不同于传统服务或产品安全治理的特殊性与新颖性？如果算法安全治理具有本质意义上的变革性特征，这种特征是否可以被纳入现有法律框架之下得到回应？从算法技术逻辑及其应用于具体环境下的制度逻辑、社会逻辑来看，这两个问题的答案可能都是具有挑战性的。具体而言，传统法律制度框架对于服务或产品安全治理的回应主要体现在侵权法和产品责任法这两大体系，但它们都很可能难以对算法安全治理的新变化、新需求作出有效回应。

侵权法的制度逻辑与反歧视法律框架有近似之处，其都建立在“过错原则”基础上，并通过寻找导致过错的因果链条而实现责任救济。以医疗诊断为例，诊断错误的医生（或医院）将承担救济责任。但这一“过错原则”在应对算法安全治理风险方面却不得不面临新的挑战。基于概率计算的算法技术逻辑，使得算法决策结果并不能被提前预测，但算法本质上一定会存在“漏洞（Bug）”的属性又使其不可能绝对正确，这些技术规律都决定了算法决策结果的正确与否只能体现为整体上的统计水平，而非针对具体案例的精准预测。仍然以医疗诊断场景为例，假设某类癌症影响的正确诊断率为 80%，那么在医生诊断场景下，20%的诊断错

误将仍然可以通过医生或医院得到责任救济；与之相比，基于海量历史数据训练的医疗诊断算法往往能提高正确诊断率（例如提升到 90%），因而医院采纳该算法有利于提升疾病治疗的整体福利水平，并因此不能被视为侵权法意义上的“玩忽职守（Negligence）”，但对于仍然存在的 10% 错误诊断，却由于算法的不可解释“黑箱”问题而可能难以得到有效的责任救济。在此背景下，算法安全治理在侵权法框架下便不能得到很好回应：一方面算法诊断相比于人为决策正确率的提升满足了勤勉要求（并因此不能被视为医生或医院的“失职”），另一方面算法诊断必然存在的错误率又因算法“黑箱”问题而使权益受损者难以确定因果链条并因此不能得到有效的责任救济。那么我们是否可以将 10% 的错误率定义为产品缺陷，将其归于产品责任法的管辖范围以追究算法设计者、部署应用者的责任呢？从算法实现的技术逻辑来看，产品责任法框架恐也难以对此做出有效应对。

产品责任法的制度设计是围绕产品缺陷问题的发现与救济而展开，但这一逻辑却并不适用于算法产品或服务。正如前文分析所指出，算法产品或服务质量的判断标准是基于统计意义而非针对具体个案。换言之，人工智能算法（例如人脸识别、智能推送等各种类型算法）一般以一定范围内的错误率为优化目标，而并不追求绝对准确。前文提及的 10% 错误率（例如识别错误或推送错误）不仅不能证明算法产品或服务缺陷，反而是算法安全的实际证明，因其说明该系统的确按照预期错误率在运行。事实上，如果导致 10% 错误率的原因被发现，这一错误本身也将被及时修正，但之所以该错误率始终都将存在，还是因为算法设计和应用过程中永远都不可能根除“未知的未知风险”（Unknown unknown risks）。

基于上述分析不难发现，算法安全治理风险难以在侵权法和产品责任法框架下得到很好回应的根本原因，仍然可归结于仅聚焦结果并试图基于结果来回溯责任的治理逻辑的失效。正是在此意义上，算法风险评估再次体现出其必要性和重要性：一方面，尽管我们并不能完全杜绝风险，但算法风险评估作为算法产品或服务投入应用的前置条件，仍然有助于作为事后责任追溯的参考标准；另一方面，更为关键的是，通过记录算法研发、应用过程信息可以积累治理经验，从而能以实验主义的方式敏捷地、渐进地对算法安全治理风险作出回应。这一逻辑便不同于传统的侵权法或产品责任法逻辑，而将主要的治理重心从算法应用的结果转移到了算法设计、开发、应用的过程。

1.4. 本章总结

基于算法歧视治理、算法公平治理、算法安全治理三个案例的分析我们不难发现，算法技术研发与应用的特殊逻辑使之难以被完全纳入到现行法律制度框架之下，这也意味着以“市场-政府”的严格划分为边界、聚焦风险结果而忽视风险产生过程的传统监管逻辑难以回应算法治理所提出的新挑战和新要求。具体而言，这又可以总结为三点。

第一，算法不能被视为独立的产品或服务而被引入人类生产生活的各个场景，其功能效用的发挥及其治理影响的释放都根植于所处环境，正是在与环境中的其他要素的相互关联、相互影响过程中算法才体现其价值。这一技术应用逻辑带来了两方面的治理后果：一方面，算法在其应用过程中需要简化现实复杂性，以单一变量代替现实中多重因素的综合影响，而这便会导致算法模型与现实世界的不一致；另一方面，基于海量数据训练的算法模型也可能体现出现实复杂性，将原本隐藏在表面现象背后的社会分化结构外化出来，从而暴露出不得不作出回应的治理风险。算法与环境的这两种关系都导致建立在“过错原则”基础上的现有侵权法或责任法框架可能失效，因试图追溯因果链条的基本原则在算法与环境的相互混杂中难以被充分满足。

第二，算法技术方案难以被解释、难以被穷尽的基本逻辑使得算法治理将始终处于动态演化过程之中，而不能被静态的绝对原则所束缚或框定。一方面，人工智能算法模型在技术层面的“黑箱性”，以及算法应用过程中与环境要素相互影响的复杂性，都决定了试图追溯从结果倒推责任因果链条的可解释努力面临重重挑战；另一方面，试图实现相同功能的算法技术方案存在多重可能性，以及算法优化标准可能具有多重内涵的模糊性，都使得试图追求“最优算法”的技术努力幻化为“乌托邦”。这两方面因素都要求算法治理呈现出敏捷性与实验性，即任何试图提前界定而不做调适的治理思路都难以真正回应算法治理需求，治理与技术开发、应用伴随而生、伴随而变需要成为新的主导性理念。

第三，算法应用风险往往以统计意义而非具体个案形式出现的基本规律，使得算法治理不得不陷入集体效用改善与个体权益侵害同时发生的公平性“悖论”

之中，而这一治理风险并非传统治理框架的关注重点。以概率计算为基本科学基础的人工智能算法，其效用提升主要体现为统计意义上的指标优化，但算法本身将始终存在“漏洞”的技术特点又决定了个体权益侵害现象必然发生，而前述算法难解释、难穷尽的基本逻辑又使得此种个体权益侵害不能再明确的责任因果链条下得到救济。如何破解这一公平性“悖论”便成为隐藏在算法治理不同案例下的共性要求。

正是基于上述三方面的新特征、新要求，算法影响评价制度的引入才不仅是必要的，也是重要的，其价值主要体现在两个方面：**对算法治理的过程进行记录以累积治理经验与知识（由此回应实验性的治理要求），在风险评价的基础上要求算法设计与应用者及时修改生产流程以作出敏捷回应（由此回应动态性的治理要求）**。接下来本报告的第二部分将基于对人工智能算法影响评价的制度实践进行梳理，并在此基础上对其制度内涵作出界定，以具体回答“算法影响评价是什么”这一基本问题。

2. 人工智能算法影响评价的内涵：描述治理实践

2.1. 人工智能算法影响评价的制度内涵及其治理定位

本报告第一部分的论述指出，就制度缘起而言，人工智能算法影响评价的必要性和重要性在于其将治理重心从结果转向了过程，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，以此积累治理经验并为算法设计者、应用者的及时调整提供参考。就现有的治理实践来看，能够完成此功能的制度工具并不只有算法影响评价，与此相关并在实践中被广泛应用的制度工具还包括算法审计（algorithm audit）。本小节将在比较二者异同点的基础上，进一步明确人工智能算法影响评价的制度内涵。

2.1.1. 算法影响评价的制度内涵

从制度沿袭视角来看，算法影响评价与财政影响评价（Fiscal Impact

Assessment)⁴、环境影响评价 (Environmental Impact Assessment)⁵等具有紧密关联,其本质上都可被视为在政府决策过程中引入专业知识以辅助决策,提供决策的科学化和民主化水平。从作用上讲,影响评价制度可能贯穿决策始终,既可能在决策事前作为项目立项或应用的准入条件,也可以作为决策事后作为项目改进或评价的结果参考。但无论其作用和方式如何,影响评价从制度内涵上讲,都可被视为对项目不同方案的社会影响(广泛涉及个人、群体、环境等不同对象)进行比较,以一定原则(例如最优原则、满意原则、底线原则)为标准选择特定方案或提出改革建议的过程。

就算法影响评价而言,其分析单位是算法系统的设计与应用过程,而分析对象则是该过程对社会各个维度(个体、集体或环境)的多类型影响,从而反过来促使算法设计、应用者调整其行为方式以避免/最小化有害影响。考虑到社会影响的宽泛性(包含不同主体、不同维度)、关联性(算法影响与其他社会因素的影响可能互相交织)、动态性(算法影响将伴随社会环境的变迁而变化)等多重复杂因素特征,算法影响评价的过程往往要求是开放式的、连续性的,即算法影响评价并不是在一个预先限定好边界的范畴中展开评价,而是需要充分考虑相关主体、具体场景、环境变迁等本地化知识以形成一个良性互动、持续推进的评价过程。

在此意义上,算法影响评价是需要被嵌入到算法设计、应用的全生命周期过程,而这也意味着传统视角下往往被视为私主体自主决策、封闭决策权力范畴的技术开发、应用过程也需要向利益相关方开放,以使得影响评价的结果能够被及时吸纳并服务于算法设计、应用决策的调整与修正。此时的利益相关方可以是政府监管部门、社会第三方乃至是全体公众,而根据开放程度的差异,算法影响评价在制度实践过程中也将形成不同模式。

⁴ 主要基于成本-效益分析方法对公共资金的投入进行影响评价,参见 Kotval, Zenia, and John Mullin. 2006. "Fiscal Impact Analysis: Methods, Cases, and Intellectual Debate." Lincoln Institute of Land Policy Working Paper. Lincoln Institute of Land Policy.

⁵ 主要对建筑工程的环境影响进行评价,参见 Taylor, Serge. 1984. Making Bureaucracies Think: The Environmental Impact Statement Strategy of Administrative Reform. Stanford, CA: Stanford University Press.

2.1.2. 算法审计的制度内涵

从制度沿革来看，算法审计与财务审计、劳动保护审计等已经被各个领域广泛应用的审计监管实践紧密相关，其目的在于通过技术方法对算法系统是否按照预期既定要求运行而进行观察、监督，其关键内容包括两部分：确定一个明确的审计标准或要求，判断算法系统的运行过程与结果是否与这个审计标准或要求保持一致。简而言之，算法审计的主要目的在于判断算法系统是否按照其所“声称”的方式在运行。

从这一定义来看，包括算法审计在内的各类型审计制度，其都包含“名义值（nominal information）”和“实际值（actual information）”这两个部分，而审计的目的就在于判断实际值是否与名义值保持一致，并在不一致的时候为执行者的修改与矫正提供参考。因此，算法审计的对象是算法系统本身，而算法审计的主体同样可以是政府监管部门或者第三方主体，乃至是具有技术能力的社会公众。算法审计从定义上讲，是一个封闭系统，因其需要在确定的“名义值”指导下来展开审计，而考虑到“实际值”需要在应用后才能被收集和对比，因此算法审计往往又体现为事后监督行为而非事前准入条件。

从算法审计的基本内涵来讲，其同样打破了私主体的决策范畴和决策权力边界，而要求外部主体（监管者或其他相关方）进入至算法设计、应用的全流程。根据外部主体参与程度的不同，算法审计可以进一步细化为内部审计和外部审计。从当前的政策实践来看，考虑到私主体可以以商业秘密保护、私有产权保护等名义拒绝开放，内部审计仍然是更主要的审计方式；但另一方面，在各国算法治理法律制度建构过程中，我们也越来越多的看到包括监管者审计在内的外部审计影响力的逐渐提升，各国正在通过制度性立法要求算法设计、应用者开放算法生命周期过程的相关信息，以使得外部审计成为可能。

同时值得注意的是，尽管算法审计是近年来才成为各国制度建设的重要组成部分，但在学术界，算法审计已经存在了很长时间。针对搜索引擎算法的偏差歧视问题、针对社交媒体算法的隐私保护问题、针对犯罪风险评估算法的种族歧视问题等，已经涌现出了非常丰富的研究文献。从研究方法来看，这些文献大都是外部审计，其都是在算法设计者、应用者没有开放全过程，而仅将算法视为“黑

箱”并通过输入、输出测试来判断、比较其是否按照既定要求运行的外部监管过程，它们主要扮演了社会公众监督的积极作用。但另一方面，考虑到算法设计、应用过程的封闭性，这种算法审计的信度、效度始终存疑，而近年来算法审计制度的改革也是针对此问题的制度性调整，试图将分散的、自下而上的算法审计过程，上升为体系性的制度改革。

2.1.3. 算法影响评价与算法审计的共性与差异

基于以上两小节的分析不难发现，算法影响评价与算法审计既具有共性，也存在较大差异。

从共性来讲，二者都要求打开私主体的决策权力边界，并将公共性考量融入到算法设计、应用过程之中。在此意义上，二者都不同于传统的建立在“公-私”边界划分基础上的监管逻辑，其目的都在于通过对算法设计、应用过程的观察与监督以发现消极或负面的治理影响，从而反过来要求算法设计、应用过程的调整与修正，以最小化或规避治理风险。

然而另一方面，二者的区别则更为明显。首先，二者的分析对象存在差异：算法影响评价的分析对象是受影响者（因变量），而算法系统是产生影响的自变量；相比之下，算法审计的对象就是算法系统本身，其要评估的是系统“名义值”与“实际值”的一致性程度。其次，二者的分析过程存在差异：算法影响评价要求开放性，即影响评价需要在不断对话的持续过程中迭代演化，具有动态性；相比之下，算法审计需要在明确的“名义值”引导下展开，并因此是封闭性的。第三，二者的制度局限存在差异：算法影响评价在很大程度上受限于其宽泛而模糊的分析对象范畴（即因为算法应用范围的广泛性和应用种类的多样性，算法影响评价所要覆盖的治理目标往往较为多元，且其内涵也较为丰富），并因此可能难以有效执行；相比之下，算法审计的“名义值”往往是确定的，并因此可以得到有效执行，但算法设计、应用者也可能按照“名义值”来优化系统进而规避真正的治理要求，这也构成了“监管规避”的不良后果。

也正是因为二者差异的存在，算法影响评价和算法审计并非相互替代关系，而是相互补充以最终回应算法治理的新要求和新挑战。本报告主要聚焦算法影响评价，接下来的实践梳理与理论分析也主要围绕此而展开。

2.2. 人工智能算法影响评价的三种模式：环境影响评价、数据保护影响评估、列表清单

人工智能算法影响评价作为一种新兴制度，其必然从相关制度中吸取经验，并在此基础上作出改革完善。也正因为此，在当前制度实践中，根据其与不同过往制度的关联度，可对人工智能算法影响评价制度的实践模式进行分类，而这又大致包括环境影响评价模式、数据风险评价模式、列表清单模式这三类。

2.2.1. 环境影响评价视角下的算法影响评价

自上个世纪六七十年代在全球各国逐步建立之后，环境影响评价几乎可被视为影响评价领域最为典型、最具代表意义的制度典范。以 1969 年美国通过《国家环境政策法》（National Environmental Policy Act, NEPA）为标志，环境影响评价已被大多数国家所采纳，并被认为在保护、改善环境方面起到了重要作用。出于不同的治理环境和治理需要，各国形成的环境影响评价制度体系存在一定差异，但基于跨国制度对比以及制度演化历史的梳理分析，我们仍然可以认为环境影响评价的制度框架大致需要包含以下三个重要组成部分，而这也可被视为环境影响评价作为一种影响评价模式的关键特征。

第一，环境影响评价往往建立在环境影响级别分类基础上，并根据不同级别的环境影响而在程序上对应提出差异化的治理要求。以美国为例，尽管 NEPA 宽泛地将对人类环境质量具有重大影响的“所有立法建议、主要联邦行动，以及这些建议或行动的合理的替代方案”都纳入要求进行环境影响评价的适用范畴，但根据对环境质量影响大小的不同，它们所要满足的评价要求却呈现出较大差异性。NEPA 在程序上规定了四个判断节点⁶，每一个节点所对应的评价要求逐次升高，

⁶ NEPA 规定的四个判断节点包括以下内容：第一，首先判断该项目是否需要环境影响评价，特别对于在其他“功能近似（Functional Equivalent）”规制框架下已经被要求进行环境分析的项目可被免于 NEPA 的评价要求；第二，其次判断该项目是否可被归类进免于环境影响评价的范畴之中，该审批主要由联邦监管机构作出以说明该项目的环境影响有限，并通过公开程序征求公众意见；第三，如果不能被归类于豁免范畴，则接下来需进一步判断该项目的环境影响是否“重大”，而联邦监管机构将出具环境评估报告以对此作出判断，并公开征求公众意见，绝大部分项目都将止步于此；第四，如果联邦监管机构认为环境影响重大，则该项目便需要准备详细的环境影响陈述，其包括可能影响、关联人群、替代方案等诸多内容。相关分析参见 Froomkin, A. M. (2015). Regulating mass surveillance as privacy pollution: Learning from environmental impact statements. *U. Ill. L. Rev.*, 1713.

直到最后一个节点才被要求出具完整的环境影响陈述（Environmental Impact Statement, EIS），而 EIS 的内容极为详细、全面并往往包含成百上千页的评估报告。在具体执行中，事实上只有较小部分的被评价对象会进入第四节点，大部分都在前三个节点完成，并因此只承担较轻的评价成本和负担。

第二，环境影响评价需要广泛纳入公众参与，并在公开、通知、反馈、修改、诉讼的迭代过程中充分反映公众意见，以确保环境影响评价的完整性与有效性。环境影响评价的公众参与固然体现了决策民主性要求，但更为重要的制度设计逻辑还在于应对环境影响评价本身的不确定性。因为人类行为的环境影响往往需要很长时间才能表现出来，而导致这种环境影响的因素又往往多元且复杂，公众参与由此扮演了缓解信息困境的重要功能。前述 NEPA 框架下的四个判断节点都被要求向公众公开征求意见，评价主体在对意见进行回复后才能进入下一阶段，而若公众仍然有不同意见则还可以提起诉讼并通过司法机构来判断环境影响评价结论的适当与否。

第三，环境影响评价要求进行替代方案的对比分析，从而“倒逼”一线主体在决策过程中考虑环境影响，从而寻找最小化影响的“最优”方案。环境影响评价的目的在于督促被评价对象改变行为方式以降低对环境的不利影响，但评价者又难以预知解决方案内容，因此在环境影响评价过程中纳入替代方案比较的要求才可能实质性推动方案的改进和优化。

需要指出的是，以上三个特征并非同等程度地反映在各国环境影响评价的制度体系之中，各国根据社会发展阶段、治理需求的差异，而可能在这三项特征上存在差异。⁷但无论差异如何，环境影响评价作为一项影响评价制度范式的基本逻辑是一致的，而这三点特征也自然影响了算法影响评价制度的探索与建构，并反映在已有研究与政策实践之中。

举例而言，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即是以环境影响评价为蓝本，要求公共部门在采购或使用算法时应展开影响

⁷ 例如中国在 2003 年施行《环境影响评价法》之前对公众参与的要求并不高，且在该法施行之后公众参与的角色也主要体现为评议而非审批或行政诉讼。参见王曦、郭祥，《中美环评制度功能比较：督政抑或规制？》，《北京理工大学学报（社会科学版）》，2023 年 1 月

评价。⁸该框架在流程上包括五个环节：采购前评估（Pre-Acquisition Review）、初步公开（Initial Agency Disclosure Requirement）、评议（Comment Period）、正当程序诉讼（Due Process Challenge Period）、影响评价方案更新（Renewing AIAs），而它们都明显体现出了环境影响评价制度所包含的公众评议特征。同时，该框架在评价内容上还提出了五项主要要求：基于影响大小合理界定评价对象范畴（Establishing Scope and Define Appropriate Boundary Around Automated Decision System）、告知公众现有及提出的算法系统方案（Public Notice of Existing and Proposed Automated Decision System）、内部自评估各个维度的算法影响（Internal Agency Self-Assessments on Fairness and other Dimensions）、为利益相关方参与提供实质意义的准入帮助（Meaningful Access），而它们同样体现了环境影响评价中的影响级别分类、公众参与、替代方案比较这三项关键特征。

在 AI Now 方案的基础上，欧洲议会研究服务中心（European Parliamentary Research Service, EPRS）在 2019 年的研究报告中进一步将算法影响评价方案细化为 8 个环节，具体包括界定算法系统范畴、初步公开算法系统信息或关于不进行算法影响评价的决定、自评估、公开系统部署后的信息获取方案、公众参与、公开算法影响评价报告、更新算法影响评价报告、公众参与阶段的质询与起诉等。由此也不难看出，EPRS 方案更是全面沿袭了环境影响评估的制度框架与主要内容。⁹

尽管环境影响评价作为一种制度创新已经被广泛使用，但这并不代表该项制度就已经“完美无缺”。事实上，在环境政策领域围绕环境影响评价制度的争议和质疑始终都存在：环境影响评价报告过于冗长以致于模糊了重点、公众深度参与会牺牲效率并严重影响项目进程、环境影响评价对私主体决策过程的干涉将侵害商业秘密等其他权益，都是其中的代表性观点。这些反思也自然会延伸到算法

⁸ <https://openresearch.amsterdam/image/2018/6/12/aiareport2018.pdf>

⁹ [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU\(2019\)624262_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf)

影响评价领域，例如 AI Now 和 EPRS 所提出的算法影响评价框架针对的主要是公共机构的算法应用，这也是为了规避评价结果公开可能侵害私主体机构商业秘密权益的争议。但也正因为此，算法影响评价制度体系并不能被等同于环境影响评价，其他模式的探索同样必不可少。

2.2.2. 数据保护影响评估视角下的算法影响评价

数据保护影响评估的制度框架基本上来源于欧盟《一般通用数据规定》（General Data Protection Regulation, GDPR）第 35 款，其要求“可能对自然人权益与自由造成高风险的（数据处理者）”都需要执行数据保护影响评估（Data Protection Impact Assessment, DPIA）。DPIA 与算法影响评估（AIA）的关联性在于，无论是欧盟第 29 工作组关于 GDPR 执行指南的说明¹⁰，还是欧盟各成员国在落地 GDPR 时的具体实践¹¹，以及理论界的法理分析¹²，都普遍对第 35 款做了宽泛性解释，即认为任何数据驱动技术事实上都会被纳入管辖范围内并因此要求执行 DPIA，而这便涵盖了所有基于大数据训练的机器学习算法。由此，为避免法律制度的重复叠加，算法影响评估完全可能被纳入并等同于 DPIA 而并不一定成为独立的制度性要求。那么接下来的问题便在于，DPIA 的制度特征是什么，其是否可以满足算法影响评估的制度要求并回应算法治理的诸多挑战？

从流程要求来讲，DPIA 包括审查基本信息以判断是否需要数据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。¹³从表面

¹⁰ Working Party on the Protection of Personal Data 95/46/EC, Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing is “Likely to Result in a High Risk” for the Purposes of Regulation 2016/679, art. 29, WP 248 (Apr. 4, 2017)

¹¹ Kaminski, M. E., & Malgieri, G. (2020). Algorithmic impact assessments under the GDPR: producing multi-layered explanations. *International data privacy law*, 19-28.

¹² Casey, B., Farhangi, A., & Vogl, R. Rethinking Explainable Machines: The GDPR's “Right to Explanation” Debate and the Rise of Algorithmic Audits in Enterprise’ (2019). *Berkeley Technology Law Journal*, 34, 143.

¹³ 肖冬梅、谭礼格，《欧盟数据保护影响评估制度及其启示》，《中国图书馆学报》，2018 年 9 月

上看，DPIA 流程与环境影响评估流程类似，但其在实质内容上存在两点重大差异。

一方面，与环境影响评估要求寻找替代方案并进行对比分析不同，GDPR 并没有对 DPIA 提出类似要求。尽管第 35 款第 7 条指出，数据处理者应“展望”

（measures envisaged）能够控制数据权益风险的可能措施，但这并不完全等同于环境影响评估中的替代方案比较。另一方面，更重要的差异还在于影响评估过程的公开程度以及公众参与程度方面。尽管 DPIA 仍然建议将评估过程及结果公开，但正如第 29 工作组关于 GDPR 执行指南的说明中所指出，“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”，即使 DPIA 显示出可能存在高风险时并要求数据处理者寻求监管方建议时，关于评估结果是否公开的决定依然取决于各成员国。¹⁴

这一制度逻辑的关键在于，DPIA 是建立在“回应性监管（Responsive Regulation）”¹⁵、“协同性监管（Collaborative Regulation）”¹⁶等新型监管理论基础上，其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）需要得到承认以激励其寻找数据保护方案的积极性。此意义上，完全的公开不仅会与数据处理者权益相冲突，也会使数据处理者直接面临社会公众的监督压力，而 DPIA 放松对公开的要求事实上在公众、监管者、被监管者（数据处理者）之间构成了一种“委托-代理”关系，即监管者作为公众的委托者来监督被监管者行为，而考虑到数据保护风险的不确定性，监管者又进一步与被监管者形成合作关系以共同探索应对风险的积极措施。

将 DPIA 延伸以作为算法影响评估的制度基础，在 GDPR 框架下看似是可能的，但其有效性却同样受到了诸多质疑。部分研究者认为 DPIA 并不能有效回应

¹⁴ <https://ec.europa.eu/newsroom/article29/items/611236/en>, P.18

¹⁵ McGeveran, W. (2016). Friending the privacy regulators. *Ariz. L. Rev.*, 58, 959.

¹⁶ Kaminski, M. E. (2018). Binary governance: Lessons from the GDPR's approach to algorithmic accountability. *S. Cal. L. Rev.*, 92, 1529.

人工智能算法引起的伦理和社会风险¹⁷，而另外一些研究则更细致的指出了 DPIA 与 AIA 的差别，即前者考虑的仍然是“知识交流（Knowledge Communication）”问题，而后者的治理重点已经演变为“知识生产（Knowledge Production）”问题，二者的差异使得 DPIA 并不能完全覆盖 AIA 的治理需求。¹⁸当然，对于 DPIA 更本质的质疑仍然可从“监管俘获”理论视角得到解释，在缺少充分公开与监督的环境下，DPIA 试图构建的监管者与被监管者“合作”愿景能否真的实现，始终是利益相关方担心的重要问题；而相比于 DPIA，AIA 将覆盖更多治理主体、涉及更多治理目标，在此框架下监管者与被监管者的“合作”究竟是更为容易还是更为困难实现，同样是值得进一步探究的问题，而非已经形成确定答案的结果。

2.2.3. 问题列表清单评估视角下的算法影响评估

开展算法影响评估的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。

加拿大“自动决策指令”制度框架的主要内容是以算法影响评估为基础，将评估结果作为风险分级的判断标准并对不同级别风险匹配以差异化的监管要求。从指令内容来看，风险分级主要围绕五个维度（具体包括个体或集体权益，个体平等、尊严、隐私与自主性，个体健康或集体福利，个体、实体或集体的经济利益，系统的可持续发展）展开，而这五个维度也构成了算法影响评估的分析对象。按照无影响（little to no impact）、有限影响（moderate impact）、高影响（high impact）、超高影响（very high impact）这四个级别，指令从同行评议（Peer review）、

¹⁷ Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, 34(4), 754-772.

¹⁸ Gellert, R. (2022). Comparing definitions of data and information in data protection law and machine learning: A useful way forward to meaningfully regulate algorithms?. *Regulation & governance*, 16(1), 156-176.

性别等个体特征分析（Gender-based Analysis Plus）、通知（Notice）、人在环路（Human-in-the-loop for decisions）、可解释（Explanation）、记录与测试（Documentation and Training）、连续性管理（Continuity Management）、准入要求（Approval for the system to operate）等八个方面提出了不同程度的治理要求。例如对于最低级别的“无影响”风险，指令仅在可解释这一点上对自动决策系统提出监管约束，要求其对共性决策结果（common decision results）提供有意义的解释；而对于最高级别的“超高影响”风险，则包含前述所有八个方面的治理要求。

在明确算法影响评估的对象及作用之后，接下来的问题就在于加拿大政府将如何展开算法影响评估？指令对此采取了列表清单问题的方式作出回应。具体而言，指令形成了涉及商业流程、所用数据、模型逻辑等多个方面的 60 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。

问题列表共包含风险以及风险预防这两个部分，而针对两个部分问题的答案也将分别形成风险评分（raw impact score）、风险预防评分（mitigation score）。该系统的影响评估总得分是按照以下两条原则来计算：如果风险预防评分小于 80% 的最高可能得分，那么总得分就等于风险评分；如果风险预防评分大于或等于 80% 的最高可能得分，那么总得分就等于风险评分的 85%。评分越高，则该系统的风险评级也将越高。¹⁹

风险部分共包含 48 个问题，最高可能得分是 107。在要求对系统作出一般性描述后，问卷又进一步细分为“影响（impact）”和“数据（data）”两个子部分。一般性描述主要涉及该项目性质（属于哪个领域、为什么需要自动化决策、决策对象是否高敏感人群、自动化决策扮演的作用和功能是什么等）、所用算法性质（是否受到商业秘密的保护、是否可解释等）等问题。在此之后，问卷将进

¹⁹ <https://aia.guide/>

一步对该自动决策系统的影响提出问题，具体问题例如该自动化决策是否存在自由裁量空间、该决策及其影响是否可矫正、该决策影响的持续时间有多长、该决策对于个体权益或自由的影响有多大、该决策对于个体健康安全的影响有多大、该决策对于个体经济利益的影响有多大等等；在数据部分，问卷的具体问题包括该自动决策系统是否使用个人数据作为输入、谁控制所使用数据、该系统是否会使用来自不同来源的数据、该决策系统是否会与其他 IT 系统相交互、谁在收集该系统所使用的输入数据等等。值得注意的是，上述问题绝大部分都是选择题，即该评估系统已经预设好相关答案，被评估人只需要在答案中选择而不需要自主回答。

风险预防部分共包含 33 个问题，最高可能得分是 45，并进一步细分为“咨询(consultation)”和“去风险与预防措施(de-risking and mitigation measures)”两个子部分，其中绝大部分问题(31 个)都在后一个子部分中。在咨询部分，问卷问题主要包括该系统设计和应用过程中主要征求意见的内部利益相关方是谁、主要征集意见的外部利益相关方是谁等。在去风险与预防措施部分，又进一步细分为数据质量、流程公平、隐私保护这三个子部分，并分别关注各个子部分的相关措施。举例而言，数据质量部分的具体问题例如有无测试记录、该记录是否可公开、是否对数据集进行了性别等敏感属性测试(Gender Based Analysis Plus)、组织内部是否有相关责任分配等；流程公平部分的具体问题例如该决策算法系统是否对应法理意义上权力主体、算法审计(audit trail)是否可以确认所有关键决策点、所有关键决策点是否都可以被关联至相关的法律规定、是否保存有系统调整的全部过程数据、该系统能否听取使用者反馈、该系统是否允许人为替代等；隐私保护部分的具体问题例如该系统是否开展了隐私影响评估、该系统在设计初始阶段是否考虑安全和隐私问题、该系统使用的信息是否闭环、该系统在与其他系统交互共享个人信息时是否取得了用户同意、该系统是否对数据做了匿名化处理等。²⁰

不难发现，该列表清单基本上涵盖了当前算法影响评估利益相关方关心的主

²⁰ <https://open.canada.ca/data/en/dataset/aafdfbcd-1cdb-4913-84d5-a03df727680c>

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。但同样值得反思的是，列表清单模式究竟能否以及在多大程度上可被视为算法影响评估的理想模式？针对此问题的回答可大致总结为以下两方面。

一方面，算法影响评估的理想模式应更倾向于开放模式而非封闭模式，自上而下试图框定评估过程的制度逻辑始终都将面临算法治理风险信息不对称以及不确定性的深刻挑战。在第一章“界定问题”的分析中，本报告已经指出积累治理过程经验（而非仅仅是解决治理问题）作为算法影响评估核心目标之一的重要性，而列表清单预先设定好的治理问题及答案都可能淡化这一目标。虽然两年一度的更新要求仍然可能使列表清单动态演化，但问题结构及答案本身的固定性，仍然约束了算法系统开发者、应用者作为一线主体的自主性，特别是其寻找可能解决方案的积极性。特别考虑到算法应用场景的多样性和复杂性，既定问题与答案究竟能在多大程度上反映现实，仍然是值得反思的深刻问题。²¹

另一方面，列表清单模式的制度逻辑仍然可能是不完整的，以致于主要体现了问卷设计者的主观判断而可能导致“南辕北辙”的治理错位问题。列表清单最终是以评分结果来界定治理风险级别，而这可能是对算法治理这一复杂问题的过度简化。这种简化又具体体现在两个层面：对于每一个算法治理问题而言，其答案之间是否可比（每个答案究竟应该被匹配以多少分数）；对于不同不同算法治理问题而言，其答案是否可比（针对不同问题的答案都被以相同分数加以对比）。举例而言，上述列表清单问题中包含“该决策系统是使用在哪个行政决策领域？”这一问题，而答案选项中，“健康相关决策”与“公共服务辅助决策（例如就业保险、残疾认证）”的得分都是1分，但我们是否可以接受这一问卷设计逻辑，健康决策与公共服务辅助决策是否等同，健康决策内部的所有场景是否都可以被视为与公共服务辅助决策等同？同时，该部分的另一问题是“该自动化决策的对象主体是否是敏感群体（vulnerable）？”，而如果选择“是”，那该答案的风险

²¹ 一个典型案例例如加拿大列表清单在数据部分的问题是“谁收集了训练该系统所需要的输入数据？”，该问题的可选答案列表包括“机构自身、另外的联邦机构、另外层级的政府、国外政府或第三方”。对此问题及答案的质疑在于，这四个答案选项能否概括所有数据收集主体，且国外政府或第三方在考虑数据收集治理风险时是可以等同的吗？而在一个开放模式下，评估问题就会被开放性地转化为“请列举训练该系统时所使用输入数据的来源？”

得分是 3 分，由此引发的反思在于：是否敏感群体的风险得分三倍于该决策所应用的场景领域得分，这一假设是否成立、其又是建立在何种合理性论证基础上，都是列表清单模式所不能回答的关键问题。

2.2.4. 不同模式的对比分析

2.3. 人工智能算法影响评价的国别特征：欧盟、美国、英国、加拿大与中国

2.2 节对于人工智能算法影响评价模式的分析是从理论上对治理实践现状的总结，但这并不意味着每个国家的政策实践都可以或仅可以归类于一种，现实的复杂性总是使得不同国家的政策实践往往都体现了对于不同模式的混合采纳。但即使如此，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。因此，在 2.2 节模式对比分析基础上，进一步对代表性国家的算法影响评价制度进行梳理，对于我们理解算法影响评价的治理实践，仍然具有重要意义。

考虑到代表性问题，本节选择欧盟、美国、英国、加拿大与中国作为分析对象，并对其具有示范意义的人工智能算法影响评价制度、政策、方案进行描述，以体现其治理特征以及差异化的治理逻辑。同时，考虑到算法影响评估是作为人工智能治理框架的重要组成部分而出现，因而本节分析也将关注各国人工智能治理框架的主要特征，并在此基础上概括各国在算法影响评估方面采取的差异化治理模式。

2.3.1. 欧盟

在当前围绕新一代人工智能治理而形成的政策浪潮中，欧盟无疑是影响力最大的区域之一，其试图沿袭在数据治理领域的成功经验，通过先发效应以将欧盟政策范式向他国输出，在规则维度体现话语权和影响力。算法影响评估作为人工智能治理政策体系中的关键环节，同样受到欧盟决策者的高度重视，并成为了欧

盟人工智能治理框架的基石。本小节将聚焦欧盟人工智能治理框架下算法影响评估的制度定位及政策内涵，以勾勒出其治理特征与治理逻辑。

（1）欧盟人工智能治理的基本框架

与数据治理框架（General Data Protection Act, GDPR）类似，欧盟人工智能治理同样采取了整体性法律规制的基本思路，试图通过制定具有约束力的一般性法律规则，以为人工智能创新和应用活动提供制度确定性，从而兼顾防控风险和促进创新的双重目标。在此思路指引下，欧盟形成了以《人工智能法案（Artificial Intelligence Act）》为主干的治理框架，该法案一方面将在此之前所提出的不具有约束力的人工智能伦理治理原则转化为了具有约束力的法律规则²²，另一方面也调和了垂直领域的人工智能法律规则从而在制度层面实现了协同统一²³。

从立法思路来讲，《人工智能法案》采取了基于风险分类的规制框架，将人工智能系统区分为不可接受风险、高风险和低风险三个大类别，针对不同风险匹配以差异化的规制措施并尤其对高风险人工智能系统应用提出了详细而明确的规制要求。²⁴正因为此，该法案的关键要素之一在于对人工智能系统进行风险分类，而法案本身采取了一般性特征和垂直领域特征相结合的方式来进行风险类别划分。高风险人工智能主要涉及该系统执行的功能或扮演的作用（例如生物识别、关键基础设施、教育或职业培训、司法等），以及该系统被使用的方式或特征（例如算法系统的使用频率，潜在危害性，结果的可逆转性等）。

与风险分类紧密相关的是针对不同类别风险而提出差异化的治理要求，而《人工智能法案》的关键在于针对高风险人工智能的规制要求，这又具体涵盖风险管理、数据治理、过程记录、透明度和信息公开、人为监管、鲁棒性、准确性、

²² 这主要体现为对欧盟高级别专家委员会（High Level Expert Group, HLEG）提出的“可信赖人工智能伦理原则（Ethics Guidelines for Trustworthy AI）”，以及欧盟发布的“人工智能白皮书（White Paper on AI）”、“欧盟人工智能战略（European AI Strategy）”等伦理治理原则、主张、战略的法律转化，将其中本不具有约束力、主要体现倡议功能的治理条款转化为需要严格执行的法律条款。

²³ 这主要体现为对分布在不同垂直领域法律规范中涉及人工智能法律条款的调和与统一，涉及民事责任领域的《产品和人工智能责任框架（Product and AI Liability Rules）》，安全规制领域的《机器规制（Machinery Regulation）》、《通用产品安全指令（General Product Safety Directive）》，数据保护领域的《通用数据保护指令（General Data Protection Directive）》，以及新近制定的《数字市场法案（Digital Markets Act）》和《数字服务法案（Digital Services Act）》等。

²⁴ 被归类为不可接受风险的人工智能系统将被禁止使用，而低风险人工智能系统往往被免于规制要求，因而主要的规制措施都是针对高风险系统及其应用。

安全性等多个维度。例如风险管理要求是指高风险人工智能系统应持续性开展风险评估措施并不断迭代，过程记录是指高风险人工智能系统的开发、应用过程需要被记录下来，而透明度和信息公开要求则涉及针对用户的可解释性等。在该法案框架下，上述规制要求被分别置于不同治理主体身上，其中又尤其针对人工智能系统的提供者，要求其在市场准入（或系统上线提供服务）之前开展“合规性评估（conformity assessment）”，并发布“合规性报告（declaration of conformity）”。

为监督治理规则的执行和落地情况，《人工智能法案》规定在欧盟层面将建立“人工智能委员会或办公室（AI Board or AI Office）”以为欧盟及成员国的国内监管机构提供规则落地的政策建议，并通过下设的两个子委员会来促进各成员国治理经验的交流与评议。在成员国层面，各国需指定特定监管机构来落实法案内容。而在涉及人工智能治理责任救济方面，与法案同步推进的《人工智能责任指令（AI Liability Directive）》引入了“过错原则”，即当被发现存在权益侵害时，各成员国司法机构有权要求高风险人工智能系统开发或部署者提供合规证据。

（2）欧盟人工智能算法影响评估的制度逻辑

《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。同时，合规性评估结果要求向公众公开，而欧盟也将成立一个专门的数据库来支撑这一公开要求。根据这一规定不难发现，合规性评估在很大程度上类似于“事前”算法审计工作，即在明确界定标准的前提下判断算法系统的设计与应用是否符合该标准要求。²⁵

另一方面，《人工智能法案》更全面体现算法影响评估的机制设计是在其2023年欧盟委员会发布的修正案中，²⁶这便主要体现为在法案中明确加入了基本

²⁵ <https://www.project-sherpa.eu/conformity-assessment-or-impact-assessment-what-do-we-need-for-ai/>

²⁶ <https://competitionlawblog.kluwercompetitionlaw.com/2023/11/13/a-deeper-look-into-the-eu-ai-act-trilogues-fundamental-rights-impact-assessments-generative-ai-and-a-european-ai-office/>

权利影响评估（Fundamental Rights Impact Assessments, FRIA）的新机制。根据修订后的第 29 款，基本权利影响评估是指高风险人工智能系统在申请之前同时还需要围绕其对“基本权利”的影响进行评估，具体内容包括以下 9 项：（a）针对该系统使用目的的明确说明；（b）该系统使用空间和时间范围的明确说明；（c）因该系统使用而受到影响的自然人或群体范围；（d）该系统与欧盟或成员国规定的基本权利相一致的证明；（e）该系统使用后对于基本权利的可预见影响；（f）对边缘人群、敏感人群可能出现的特定风险；（g）对环境产生的可预见的负面风险；（h）削弱或消除对基本权利风险的详细方案；（j）该系统部署者所采取的治理方案，包括人为监管、责任救济等。尽管不要求完全公开，但基本权利影响评估的结果要求向监管机构以及利益相关方汇报或分享。

考虑到新一代人工智能系统与数据的紧密关联性，基本权利影响评估也可能与《通用数据保护指令（GDPR）》中规定的“数据保护影响评估（DPIA）”产生交叠。为避免增加合规负担，修正后的《人工智能法案》同样在 29 款规定了 FRIA 和 DPIA 的关系问题。法案指出，如果 GDPR 要求人工智能系统提供者开展 DPIA，那么在 FRIA 中也应包括 DPIA 的摘要说明（特别是关于在该系统特定的使用目标和使用背景下）。

基于上述梳理不难发现欧盟《人工智能法案》在设计算法影响评估制度方面的一些特征。首先，欧盟明确将人工智能算法影响评估作为单独的监管内容独立于其他影响评估（特别是 DPIA），这体现了欧盟已经认识到人工智能治理的特殊性以致于不能仅仅以 DPIA 来代替算法影响评估。其次，欧盟范畴下的人工智能算法影响评估涵盖范围较广，不仅涉及个体权益或集体权益，同时还将环境权益纳入其中，明确要求将人工智能系统的环境影响评估（包括但不限于人工智能系统的能源消耗问题）纳入其中，而这也明显反映了法案针对当前大模型快速发展这一时代背景的敏捷反应。最后，欧盟框架下的算法影响评估在很大程度上类似于算法审计，即其试图以较为明确的标准（具体是指法律规则）来指引算法影

响评估，在此意义上影响评估也与合规评估紧密联系起来。基于这种制度设计逻辑，我们便不难发现欧盟试图兼顾风险治理和保护创新这双重目标的基本思路。

（3）欧盟人工智能算法影响评估的总结与反思

尽管欧盟《人工智能法案》高度重视算法影响评估，并在极为精巧的法律设计中建立了算法影响评估的主要框架，但这并不代表其已基本完善。无论是合规性评估（Conformity Assessment）还是基本权利影响评估（FRIA），都存在不可忽视的缺陷与短板，并因此引起了诸多反思与批评。

对于合规性评估而言，一个主要批评在于其可能只能解决形式合规问题而难以促进实质治理目标的实现。²⁷一方面，考虑到人工智能治理法律制度尚在发展进程中，现行法律规则事实上难以为合规性评估提供明确的审计标准和目标；另一方面，即使存在明确的法律规则，考虑到法律本身作为风险底线的制度定位，且其要求提供明确权益侵害证明以及责任因果链条的制度逻辑，使得大量因人工智能系统应用而产生负面影响的治理风险难以被框定在内。在此意义上，合规性评估很可能只能关注是否建立了数据管理制度、是否存在报告流程、是否存在风险预警和解决措施等程序上的形式要求，而难以切中环境影响、基本权利影响等实质性目标的实现与否。在此意义上，基本权利影响评估能够作出补充，但其仍然存在不足之处。

对于基本权利影响评估而言，普遍性的质疑和反思是认为其可能加重被规制对象的合规成本，同时也可能因为存在“例外豁免”情况而沦为摆设。一方面，批评者认为基本权利影响评估存在执行层面的困难与挑战，其与其他垂直领域已经提出的法律要求相重合并可能导致多重合规成本，而且将《人工智能法案》管辖范畴延伸至司法、环境等领域也可能模糊其聚焦隐私、安全等基本权益保护的立法初衷。正因为此，批评者往往提出建议，认为应删除基本影响评估条款。²⁸另一方面，批评者同样指出《人工智能法案》的制度设计存在漏洞，以致于大型数字平台企业或高风险人工智能应用可能通过“例外豁免”条款来有效规避基本

²⁷ <https://www.project-sherpa.eu/conformity-assessment-or-impact-assessment-what-do-we-need-for-ai/>

²⁸ <https://www.digitaleurope.org/resources/ai-act-trilogues-a-vision-for-future-proofing-governance-and-innovation-in-europe/>

权利影响评估的规制要求。法案中规定，在应急状况或对于中小企业等缺少进行评估的资源和能力情况下，可以在未展开基本权利影响评估的前提下应用人工智能。²⁹在批评者看来，这一规定可能产生制度漏洞，例如大型数字平台企业可以通过设立新的创业公司以成为法律意义上的“中小企业”，从而同样能够在不展开基本权利影响评估的前提下测试、应用高风险人工智能。³⁰

2.3.2. 加拿大

在人工智能治理方面，加拿大采取了与欧盟类似的制度设计逻辑，即通过制定一般性法律规范来兼顾风险规制和促进创新的双重目标，而其代表性法律是正在制定过程中的《人工智能与数据法案》（Artificial Intelligence and Data Act, AIDA）以及已经施行的《自动决策指令》（Directive on Automated Decision-Making, DADM）。但与欧盟试图在人工智能治理领域复制“布鲁塞尔效应”的立法理念不同，加拿大明确表达了与欧盟、美国、英国保持制度“对齐性”（align approaches）的观点，以确保加拿大公民能够在全球市场受到保护而加拿大公司能满足各国治理要求。³¹在此目标指引下，加拿大人工智能治理同样采取了基于风险分类的规制思路，并形成了一系列的制度设计。

（1）加拿大人工智能治理的基本框架

加拿大人工智能治理框架以 AIDA 和 DADM 两部法律为主要支柱。AIDA 是由信息、科学和产业部（Minister of Innovation, Science, and Industry, MISI）在 2022 年 6 月作为《数字宪章实施法案》（Digital Charter Implementation Act, DCIA）的一部分而提出，而 DCIA 的初衷及目的是改革加拿大的数据隐私保护制度、落实数字宪章要求，以确保加拿大公民能够信任数字化转型进程中的技术创

²⁹ <https://artificialintelligenceact.eu/wp-content/uploads/2023/08/AI-Mandates-20-June-2023.pdf>

³⁰ <https://www.socialeurope.eu/the-ai-act-deregulation-in-disguise>, 388d

³¹ <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>

新与产业应用。考虑到人工智能与数据的紧密关联，AIDA 便作为重要组成部分纳入到了 DCIA 框架下，其主要对私人部门的人工智能研发与应用行为作出规范。

³²与之相比，DADM 主要聚焦公共部门中的自动决策算法应用规范，其在 2019 年 4 月正式开始实施，试图通过完善制度规范来提升公共部门的算法应用与自动化决策水平。

从制度内容来讲，AIDA 的基本逻辑与欧盟《人工智能法案》一致，其都建立在风险分类框架之上，并重点关注“高影响”（high impact）人工智能应用的规制问题。但与欧盟不同，AIDA 并没有对“高影响”人工智能给出明确定义，而只是宽泛提出通过利益相关方咨询的方式决定高影响人工智能的范畴。不过即使如此，AIDA 也给出了判断高影响人工智能的参考建议，这又具体体现在以下几个维度：对健康、安全以及基本人权会造成风险的证据，潜在风险的严重程度，人工智能应用规模，已出现风险的性质，退出人工智能系统的难度，人工智能应用环境的分化结构，以及在现有法律框架下能够被有效规制的程度等。³³围绕这几个维度，AIDA 明确提出，其所关注的治理风险主要体现在两方面：对于公民个体权益的侵害（Harms to individuals），以及存在偏差或歧视的输出结果（biased output）。基于此也不难发现，相比于欧盟《人工智能法案》在基本权利影响评估中较为宽泛地将环境、财产权、人权等包括在内的立法思路，AIDA 所针对的问题集合显得更为聚焦和狭窄。不过根据加拿大信息、科学和产业部的规划，AIDA 不会早于 2025 年出台，³⁴而这也意味着该框架下的“高影响”人工智能范畴究竟将包括哪些内容仍然还存在较大不确定性。

DADM 同样沿袭了风险分类的治理框架，但其并不仅仅针对“高影响”自动

³² <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>

³³ <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>

³⁴ <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>

决策系统提出规制要求，而是一般性地覆盖多个风险类别。在 DADM 框架下，风险分类的维度主要沿两个方向展开：一方面，以跨领域的横向视角为指导，考虑是否可逆、持续时间长短等自动决策系统的共性特征；另一方面，以基于领域的纵向视角为指导，考虑健康福利、经济权益、可持续发展等不同领域权益。二者结合起来以作为风险判断和分类的基本标准。作为结果，自动决策系统被划分为无影响（little to no impact）、有限影响（moderate impact）、高影响（high impact）、超高影响（very high impact）这四个类别，而每个类别也相应匹配以差异化的规制要求和措施。

从规制要求的角度来看，AIDA 和 DADM 与欧盟《人工智能法案》还存在一定差异。加拿大决策者采取了公、私分立的监管态度，AIDA 主要针对私主体的人工智能开发、应用活动，因此尽管 AIDA 要求“高影响（High Impact）”人工智能系统开发、应用者需要向监管者汇报，但其并不像欧盟《人工智能法案》般要求“合规性评估（Conformity Assessment）”以作为市场准入前提条件。对于公共部门而言，DADM 则明确要求自动决策系统在落地应用之前需要进行算法影响评估，并根据评估结果来匹配相应的规制措施。

（2）加拿大人工智能算法影响评估的制度逻辑

与欧盟类似，算法影响评估是加拿大人工智能治理制度框架的基础，AIDA 和 DADM 都将其纳入制度范畴，并扮演了基础性作用。

在 AIDA 框架下，算法影响评估是确定“高风险”人工智能的前提，而这也是基于此展开后续治理活动、提出后续规制要求的基础。但考虑到 AIDA 仍然处于制定过程中，究竟如何评定“高风险”还存在立法不确定性：谁来评定、如何评定、经历何种程序来评定、评定结果应如何更新等问题还需要立法的进一步探索与回答。

相比之下，DADM 框架下的算法影响评估已经较为成熟和体系化，其事实上构成了 2.2.3 节所描述的“问题列表清单评估”的特色模式。考虑到 2.2.3 节已对

此做了较为充分的分析，本节不再赘述。

（3）加拿大人工智能算法影响评估的总结与反思

基于上述分析，不难发现加拿大人工智能算法影响评估制度的三个主要特征。

第一，加拿大是以公、私分立的态度推进人工智能治理，而针对不同主体开展人工智能算法开发、应用活动的影响评估，形成了不同的关注重点。对于公共部门而言，DADM 形成了较为完善的“问题列表清单评估”模式，以形成四种类型的风险评估结果；对于私人部门而言，AIDA 仅要求区分“高影响”人工智能，而不对其他类型作出区分。

第二，相比于欧盟，加拿大更为明显地体现出了试图平衡风险监管与创新激励的双目标视角，并更加侧重于后者。与欧盟对私人部门施加较多的影响评估压力（例如基本权利影响评估制度）相比，加拿大在很大程度上削弱了私人部门的影响评估责任。AIDA 仅对“高影响”人工智能提出规制要求，而“高影响”的判断标准又要求征求人工智能开发者、应用者意见，由此可能形成较为宽松的监管环境。

第三，加拿大较为明显地体现出了将国内治理与全球治理紧密联系在一起的立法态度。AIDA 明确提出与欧盟、美国、英国的人工智能监管框架保持一致，其目的则在于为加拿大国内的人工智能产业提供协同化的国际监管环境，降低合规负担。

正是基于上述三点总结，我们也不难发现加拿大人工智能治理，以及更具体的人工智能算法影响评估制度框架的不足，这又具体体现在以下两方面。

一方面，批评者认为加拿大人工智能立法（主要针对 AIDA）过于保护人工智能产业或企业而轻视了基本权益保护。AIDA 所提出的针对“高影响”人工智能开发者、应用者的规制要求，可能在宽泛的指导原则下沦于形式，而不能真正实现以算法影响评估来“倒逼”产业或企业提升治理水平的立法目的。同时，更有批评者直接提出，信息、科学和产业部作为产业发展部门，其作为人工智能规制立法提出者的身份是不合适的，强调发展的部门逻辑将自然忽视基本权益保护，

因而应由其他部门代替。³⁵

另一方面，批评者同样认为 DADM 所形成的“问题列表清单”评估模式可能并不能有效应对人工智能风险治理要求。“问题列表清单”评估模式的封闭性、自上而下特征，很可能难以应对人工智能技术创新和产业应用本身的动态演化逻辑，且其内在制度逻辑的一致性同样遭受颇多质疑。在此背景下，AIDA 究竟是否会采纳、模仿 DADM 的算法影响评估模式，仍然具有不确定性，而这事实上也为加拿大人工智能治理制度带来了不确定性，反而不利于产业发展。

2.3.3. 美国

作为人工智能技术创新与应用大国，美国在人工智能治理领域的政策选择始终其他国家关注的对象。但与欧盟、加拿大、中国等已经出台立法的国家和地区相比，美国人工智能治理政策仍然尚未上升到法律层面，而主要采取了两种路径以回应人工智能治理的迫切需要：一方面，在联邦层面发布一般性的治理原则或准则，通过行政命令或规则标准的方式来约束、引导公共部门或私人部门在人工智能领域的开发与应用行为；另一方面，垂直领域的法律规制机构正在关注人工智能发展与应用背景下所出现的新的治理需求和治理挑战，并试图通过各垂直领域的法律更新来作出回应。在人工智能治理领域，美国当前的模式特征是否有效，及其未来的动态更新是我们需要关注的重点内容。

（1）美国人工智能治理的基本框架

美国人工智能治理框架主要围绕前述两条路径展开，并分别形成了一些代表性的治理规则文本或治理行动。

就联邦层面的治理进展而言，类似于加拿大，美国人工智能治理同样采取了公共部门、私人部门分开推进的模式特征，但也同时明确提出了以公共部门应用人工智能的治理规范来引导私人部门治理进程的内在影响机制。

对于公共部门的人工智能治理而言，主要以特朗普、拜登政府连续发布的行

³⁵ <https://www.canadianlawyermag.com/practice-areas/privacy-and-data/critics-say-artificial-intelligence-and-data-act-needs-to-focus-more-on-rights-not-just-business/380552>

政命令为主，并以此为基础进一步延伸至联邦管理与预算办公室（Office of Management and Budget, OMB）发布的更为系统的“促进联邦机构推动人工智能治理、创新、风险管理的备忘录”（Proposed Memorandum for the Heads of Executive Departments and Agencies on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence）（以下简称“备忘录”）。

2019 年特朗普政府连续发布的《行政命令 13859：维持美国在人工智能领域的领先地位》（Executive Order 13859: Maintaining American Leadership in Artificial Intelligence）和《行政命令 13960：促进联邦政府应用可信赖人工智能》（Executive Order 13960: Promoting the Use of Trustworthy AI in the Federal Government），都明确提出要为联邦机构应用人工智能确立治理规范，从而在提升人工智能在联邦机构应用水平的同时促进公众信任。特朗普政府发布的行政命令在拜登政府时期得到了延续，后者主要采取了两项具有影响力的政策策略。一方面，拜登政府在 2022 年发布了《人工智能权利法案路线图》（Blueprint for an AI Bill of Rights, BOR），确定了在人工智能应用时代保护美国公众权利与自由的五项基础性原则。³⁶从在整个治理框架中的定位来看，BOR 类似于欧盟框架下的“可信赖人工智能的伦理指导意见”（Ethics Guidelines for Trustworthy AI），其主要表达了政府在人工智能治理方面的基本理念和原则，并为后续更进一步、更实际的规则制定奠定基础。另一方面，拜登政府在 2023 年 10 月发布了《安全与可信赖人工智能的发展与使用行政命令》（Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence），而在该行政命令发布后的两天，白宫 OMB 办公室即发布了“备忘录”草案以征求公众意见。

³⁶ 具体包括安全与有效系统（Safe and Effective System）、算法歧视保护（Algorithmic Discrimination Protection）、数据隐私（Data Privacy）、通知与可解释（Notice and Explanation）、人为决策作为可选项（Human Alternatives, Consideration and Fallback）等五项。

对于私人部门的人工智能治理而言，美国的代表性治理行动主要体现在被反复提出（2019、2022、2023 三次被提案）但尚未通过国会立法的《算法责任法案》（Algorithm Accountability Act, AAA），以及由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）。这两个文本都是针对一般性人工智能应用而形成的治理文本，但在 AAA 立法通过之前，二者都不具有约束力，而只是作为鼓励私人部门采用的自愿性、引导性文本。

从内容上讲，AAA 关注的规制对象是“能够辅助或支撑具有法律或现实影响的决策系统”（augmented critical decision process），其广泛覆盖教育、工作、健康、司法、金融等多个领域。值得注意的是，在规制对象范畴框定的过程中，AAA 同样明确了“技术中立”原则，即仅仅只是作为决策系统支撑的“消极计算基础设施”（passive computing infrastructure）并不被包括在内，典型案例例如支撑算法运行的云计算基础设施。而就部署、使用算法的组织实体而言，AAA 将联邦贸易委员会（Federal Trade Commission, FTC）作为监管机构而非单独设立，在 FTC 的监管框架下，其主要将年收入超过一定限额或处理公民个体信息的数量超过一定限额的大型企业作为监管标的，由此为中小企业创新提供了法律保护。³⁷值得注意的是，AAA 的这一措施与欧盟《人工智能法案》存在差异，后者并没有将中小企业作为例外而豁免在监管之外，而是强调为中小企业提供相匹配的合规资源。³⁸究竟哪种模式是更好的模式，仍然需要在实践中观察。同样重要的问题还在于 AAA 所提出的规制要求，其主要涵盖算法影响评估和年度报告这两项，本报告将在下一节做更详细的描述。

上述分析都体现了美国试图在人工智能治理方面构建跨领域一般性规制框架的努力，但这并非美国决策者的唯一特征，特别是考虑到上述框架都尚未被上

³⁷ <https://www.holistica.com/blog/us-algorithmic-accountability-act>

³⁸ <https://artificialintelligenceact.com/title-v/article-55/>

升到法律层面，其在执行层面的效力仍然值得反思。一方面，对于已经在公共部门产生影响力的行政命令及“备忘录”而言，其作为行政部门自身管理规则的局限性，使得在政府换届后很可能不被认可并因此废止；另一方面，对私人部门而言，AAA 到底能否被通过立法仍然具有未知数，这也反映了立法部门（国会）对是否将一般性立法作为人工智能治理框架的思路存在质疑。与这两方面担忧并行的，是美国各垂直领域监管者对人工智能创新应用而引发新挑战的回应。

例如，FTC 在 2021 年即发布了政策指引，对不同法律在人工智能系统应用方面的适用性进行了解释，其主要关注了三点。³⁹第一，FTC 法案（FTC Act）第五款禁止不公平或欺诈性商业措施，因此售卖或使用种族歧视算法自然将被涵盖在该法案范畴内；第二，公平信用报告法案（Fair Credit Reporting Act, FCRA）旨在规范消费者信用报告（评级）产业，以确保消费者得到公平对待，而当算法被应用于公众在就业、住房、信用、保险或其他福利申请的评估时，其自然被纳入到该法案的管辖范畴；第三，平等信用机会法案（Equal Credit Opportunity Act）要求向消费者提供贷款或消费信用的政府机构、企业或个人，在作出授信决定时不能由于申请人的性别、婚姻状态、种族、宗教信仰、年龄等因素做出歧视性的授信决策，因此当算法被用于授信决策或公共福利申请决策时，自然也将被纳入该法案范畴。FTC 作为监管机构对上述法律在人工智能时代的适用性解读，充分体现了美国试图强化在垂直领域解决人工智能治理问题的努力。换言之，无论 AAA 等跨领域的一般性法律规范通过与否，FTC 作为监管机构都将利用现有法律框架推进人工智能治理工作。

（2）美国人工智能算法影响评估的制度逻辑

在上述对于美国人工智能治理框架的梳理中不难发现，算法影响评估在美国框架下尚没有凸显出其基础性地位。事实上，对于垂直领域的监管努力而言，算法影响评估并不是最核心内容，因无论算法影响如何，一旦其落入已有法律的监管框架，其都会面临需要满足合规性要求的责任。只有对于跨领域的一般性治理

³⁹ <https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>

规则而言，算法影响评估的重要性才凸显出来，而这又分别体现在针对公共部门和私人部门的治理框架之中。

一方面，拜登政府 OMB 办公室在 2023 年 10 月发布的“备忘录”基本上可被视为对联邦机构应用人工智能确立了较全面和完善的治理框架，其中也明确提出了算法影响评估的关键地位。⁴⁰该备忘录提出了强化人工智能治理体系（Strengthening AI Governance）、促进负责任人工智能创新（Advancing Responsible AI Innovation）、管理人工智能应用风险（Managing Risks from the Use of AI）这三项主要目标，并在每项目标下都提出了较为具体的改革行动建议。其中，与本报告关注的人工智能算法影响评估紧密相关的部分在第三项目标，即管理人工智能应用风险。“备忘录”在该环节体现了与他国不一样的治理思路。第一，“备忘录”将人工智能应用区分为“安全影响”（Safety Impacting）和“权益影响”（Rights Impacting）两种类型，前者是指可能在人员生命或福利、气候或环境、关键基础设施、战略性资源等方面产生影响的人工智能应用，而后者是指可能会作为决策基础以对公民权利与自由、平等机会、关键资源获取等产生影响的人工智能应用。第二，基于这两类区分，“备忘录”提出了两类规制要求：一是要求联邦部门在决定应用人工智能时需要采取“成本-收益”分析，在证明收益大于成本后才能应用；二是对不同类型人工智能应用提出了一系列最低规制要求。第三，最低规制要求包括人工智能影响评估、在真实环境下测试人工智能、独立评估人工智能、持续监督并设置人为审查“门槛线”、移除或减轻风险的措施、确保适当的人力资源训练和评估、对于高影响情况的人为介入、通过人工智能应用库来实现公众告知与解释等系列措施。特别的，对于“权益影响”的人工智能应用，“备忘录”还进一步提出了确保人工智能促进公平、纳入受影响群体反馈、持续监督并减轻因人工智能而产生的歧视结果、通知受到负面影响个人、维系人为干涉和救济流程、在可能情况下提供退出渠道等新的最低规制要求。由此不难看出，“备忘录”仍然采取了基于风险的规制框架，但在风险类型的划分、

⁴⁰ <https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf>

规制要求的提出（以列表清单式）等方面仍然体现了差异化特点。

另一方面，对于私人部门监管而言，AAA 所提出的主要规制要求都围绕影响评估展开，涵盖算法影响评估和年度报告这两项。AAA 规定，被监管者需要开展算法影响评估，并主要涵盖以下 11 项基本要求：流程评估（Process Evaluation）、利益相关方咨询（Stakeholder Consultation）、隐私评估（Privacy Assessment）、绩效评估（Performance Evaluation）、（对使用者的）训练与教育（Training and Education）、防护与使用限制（Guardrails and Limitations）、数据记录（Data Documentation）、用户权利评估及透明和可解释要求（Rights, Transparency and Explainability）、负面影响评估（Negative Impact Assessment）、关键节点的文本记录（Documentation and Milestones）、优化决策的资源储备（Resource Identification）。⁴¹与欧盟和加拿大所提出的算法影响评估要求相比，其特点是比较具体的集中在公民权益方面，而并未将环境影响等更广泛内容纳入其中；同时，AAA 的算法影响评估又更具体地列举了缓解风险措施，其中“对于使用者的训练与教育”等要求都是具有创新意义的。

（3）美国人工智能算法影响评估的总结与反思

基于上述分析，不难发现美国在人工智能算法影响评估制度建设方面的三个主要特征。

第一，相比于欧盟，美国更为狭窄地定义了监管对象和监管范畴，并在算法影响评估要求上更为宽松。就前者而言，其定位于“大企业”的立法目的不仅不同于欧盟，也事实上构成了对于中小企业的“安全港”原则；就后者而言，其并没有宽泛地将环境、人权等纳入算法影响评估范畴，而是聚焦在法律权益方面，这也有别于欧盟路径。

第二，美国当前并没有形成具有约束力的人工智能监管规则，所有的治理努力都仍然致力于引导、促进公共部门、私人部门自我治理水平的提升。特别的，

⁴¹ <https://www.holistica.com/blog/us-algorithmic-accountability-act>

美国试图通过规范公共部门人工智能治理行为以影响私人部门，其影响机制可以是政府采购等。

第三，美国尚没有形成较为清晰的算法影响评估模式，尽管对评估内容、公开性要求等关键环节都已经作出了一定规范，但都没有提出具体要求。这也使得美国人工智能治理制度在未来还保持有较大的开放性，其究竟往哪个方面发展还需要进一步的观察。

2.3.4. 英国

作为新一代人工智能强国，英国同样将人工智能的发展与应用置于了国家战略位置，并充分认识到了人工智能治理在促进公众信心从而推动人工智能发展方面的关键作用。不过与欧盟、美国、加拿大相比，英国采取了较为不同的治理路径，其更主要体现了依托各垂直领域监管者灵活反应的治理思路，并通过建设跨领域的治理原则和协同性的治理机制来确保各领域治理措施的一致性、连续性。而这就明显不同于欧盟、加拿大试图建构统一监管框架的努力，同时也不同于美国即使强调垂直领域监管但仍然在联邦层面形成自愿性、引导性监管框架的尝试。

（1）英国人工智能治理的基本框架

英国人工智能治理思路主要体现在其发布的《以促进创新路径来规制人工智能》（Establishing a pro-innovation approach to regulating AI）的政策报告中。

⁴²与欧盟的“可信赖人工智能的伦理指导意见”（Ethics Guidelines for Trustworthy AI）或美国的《人工智能权利法案路线图》（Blueprint for an AI Bill of Rights, BOR）不同，该报告尽管阐述了英国人工智能治理的主要理念，但其重点不在于对人工智能治理的基本原则作出阐释（例如“可信赖”的具体内涵，或者“权利法案”的具体内容），而是重在对英国人工智能框架作出解释和说明。正是基于该报告，英国依托垂直领域监管并形成跨领域对话、协调机制的人工智能治理框架才被明确建立起来。

该报告将英国人工智能治理框架命名为“促进创新的治理路径（A New Pro-

⁴² <https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement>

innovation Approach）”，其又具体包含四条原则：基于场景的治理（Context-specific）、促进创新和基于风险（Pro-innovation and risk-based）、连贯一致的（Coherent）、比例原则和可调整的（Proportionate and adaptable）。首先，基于场景的治理便确定了英国人工智能治理依托垂直监管领域的基本路径，因其认为人工智能的治理风险需要在应用层面才能得到确认和评估，因而实际发生的影响（actual impact）只有在特定场景下才能被充分发现。其次，基于风险和比例原则意味着垂直领域的监管者需要建立与风险大小相匹配的治理措施，而这也意味着英国将采取同欧盟、加拿大相类似的主要聚焦“高风险/高影响”人工智能应用监管的基本思路。第三，可调整原则意味着人工智能治理是一个敏捷反应的过程，而与之相对应的确定的、中心化治理风险清单既容易过时，也不利于发挥各垂直领域监管者的经验和专长。最后，一致性原则意味着需要通过跨领域协同机制来避免各垂直领域监管实践的冲突和碎片化，而这又主要体现在形成统一的宽泛监管原则与建设协同性监管机构这两个层面。

该报告认为，虽然人工智能存在差异化的技术特征与应用特征，但其仍然存在一定共性，并因此需要建立基础性的宽泛监管原则，而这些原则在不同领域可以得到差异化的解释，但统一的基础性原则有利于引导各领域监管者朝着相同方向探索具体监管实践，并避免监管冲突现象。就此而言，该报告主要提出了六条基础原则：确保人工智能应用的安全性（Ensure that AI is used safely）、确保人工智能在技术上是安全的并按照既定功能在运行（Ensure that AI is technically secure and functions as designed）、确保人工智能在合适层面上是透明的且可被解释（Make sure that AI is appropriately transparent and explainable）、将公平性嵌入人工智能之中（Embed considerations of fairness into AI）、将人工智能治理责任落实到法律人（Define legal persons' responsibility for AI governance）、为救济或抗争提供明确渠道（Clarify routes to redress or contestability）。

除了共性原则的基础性、宽泛性要求之外，英国还建立了“数字监管合作论

坛”（Digital Regulation Cooperation Forum, DRCF）这一跨部门协同机制。⁴³在2020年6月，英国市场竞争部（The Competition and Markets Authority, CMA）、信息委员会办公室（Information Commissioner's Office, ICO）以及通信办公室（Office of Communication）共同成立了 DRCF，随后金融指导机构（Financial Conduct Authority, FCA）在2021年4月又正式加入。DRCF的初始目的是在监管在线平台方面推动跨部门合作，而当前也逐渐延伸至人工智能监管领域。DRCF并非法定机构，其属于各部门间自愿成立，其作用既包括共同开展探索性监管研究，也包括分享监管经验和标杆实践案例等。

为更好促进各垂直领域监管部门在人工智能治理方面的实践探索，英国政府还拨付了超过1亿英镑的资金以支持各部门探索监管工具创新，并同时要求每个部门在2024年4月之前都需要就如何应对人工智能治理风险与机遇发布工作计划。⁴⁴其中，较有代表性例如科学、创新和技术部发布的《数字监管：促进增长与释放创新》政策报告⁴⁵，全面阐释了该部门将要推行的监管改革创新，将积极促进创新（actively promote innovation）、实现向前看且前后一致的监管结果（achieve forward-looking and coherent outcomes）、以国际视角释放机会并应对挑战（Exploit opportunities and address challenges in the international arena）这三项确定为监管改革目标，同时较为明显地表达了对于技术解决方案（例如强调“十项技术优先战略”⁴⁶）、对于促进中小企业创新（例如强调成立数字市场

⁴³ <https://www.gov.uk/government/collections/the-digital-regulation-cooperation-forum>

⁴⁴ <https://www.gov.uk/government/news/uk-signals-step-change-for-regulators-to-strengthen-ai-leadership>

⁴⁵ <https://www.gov.uk/government/publications/digital-regulation-driving-growth-and-unlocking-innovation/digital-regulation-driving-growth-and-unlocking-innovation>

⁴⁶ 即 Ten Tech Priorities，包括建设世界领先的数字基础设施（Rolling out world-class digital infrastructure nationwide）、释放数据潜能（Unlocking the power of data）、建设技术常识国家（Building a tech-savvy nation）、保持在线安全（Keeping the UK safe and secure online）、开启创新创业新阶段（Fueling a new era of startups and scaleups）、释放技术和人工智能的变革能力（Unleashing the transformational power of tech and AI）、促进自由和公平数字贸易（Championing free and fair digital trade）、推动围绕技术治理的全球对话（Leading the global conversation on tech）、促进数字繁荣（Leveling up digital prosperity across the UK）、推动数字创新赋能零排放（Using digital innovation to reach Net Zero）等十项，参见

单位作为新型监管机构以限制大企业市场垄断能力⁴⁷)等的政策偏好。当然,对此更全面的分析还需要在各部门都发布报告之后才能全部完成,不过基于《数字监管:促进增长与释放创新》政策报告我们也不难发现其坚持以完善监管来促进创新的核心目的。

在英国促进垂直领域监管改革的同时,一个具有共性的跨领域治理措施是推进人工智能保证生态系统(AI Assurance Ecosystem)的形成与繁荣。所谓人工智能保证生态系统,是指创造出一系列治理工具和服务以确保人工智能系统能够按照它们所预期的行为而工作,从而提升消费者和公众信心以支持人工智能变革潜力的释放。在英国政府看来,人工智能治理本身并不是要对人工智能施加“枷锁”,其反而构成了人工智能创新生态的重要组成部分。于是,英国数据伦理与创新中心(Centre for Data Ethics and Innovation, CDEI)在国家AI战略(National AI Strategy)的要求下发布了“人工智能保证生态系统路线图(AI Assurance Ecosystem Roadmap)”,对英国在该领域的改革措施做出了全面规划。⁴⁸该路线图明确了三个相关主体:负责任的人工智能开发者等作为第一主体、人工智能应用者作为第二主体、人工智能审计者等作为第三主体。第三主体为第一主体提供证明、为第二主体提供信息,而第二主体对第一主体的采购为第一主体提供信任证明。在此框架下,人工智能算法影响评估、算法审计等相关活动便被置于了核心位置,更具体的内容将在下一小节得到阐释。

(2) 英国人工智能算法影响评估的制度逻辑

从整体上看,由于英国坚持以垂直领域监管作为人工智能治理思路,因而人工智能算法影响评估并没有取得其在欧盟、加拿大等国人工智能治理体系中的类似地位,后者以风险分类为基础的治理路径十分依赖人工智能算法影响评估的结果,而英国框架下,各垂直领域的监管逻辑并不完全依赖风险分类,这也使得算法影响评估并不一定会成为制度基础。但这并不意味着英国不重视人工智能算法影响评估,其在前文提及的人工智能保证生态系统、数字监管合作论坛等具体机

<https://dcms.shorthandstories.com/Our-Ten-Tech-Priorities/index.html>

⁴⁷ <https://www.gov.uk/government/collections/digital-markets-unit>

⁴⁸ <https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem>

制中仍然占据着重要位置。

“人工智能保证生态系统路线图（AI Assurance Ecosystem Roadmap）”文件对人工智能算法影响评估所包含的丰富工具箱进行了全频谱的勾勒，按照评估对象与标准清晰程度从模糊到清晰的程度不同，其可被划分为影响评估（Impact Assessment）⁴⁹、影响评价（Impact Evaluation）⁵⁰、歧视审计（Bias Audit）⁵¹、合规审计（Compliance Audit）⁵²、认证（Certification）⁵³、合格性评估（Conformity Assessment）⁵⁴、绩效测试（Performance Testing）⁵⁵、正式认证（Formal Verification）⁵⁶等八种。⁵⁷但值得注意的是，该路线图并没有对不同领域监管框架下，需要进行何种算法影响评估、评估结果将如何被用于匹配差异化的治理要求等问题作出进一步的确认，而是将此留给了各领域监管者自身来决定。换言之，路线图仍然只是提供了人工智能算法影响评估的“工具箱”，具体采用哪种工具、在何种情况下采用此种工具仍然是属于各领域监管者自由裁量的权限范畴。

在英国政府层面发布的与人工智能算法影响评估相关的文件是由数字监管合作论坛机制发布的政策报告《审计算法：当前进展、监管者角色和未来展望》

⁴⁹ 被用于预测该系统对环境、平等性、人权、数据保护或其他产出的影响（Used to anticipate the effect of a system on environmental, equality, human rights, data protection, or other outcomes）

⁵⁰ 类似于影响评估，但是在事后以反思性形式展开（Similar to impact assessments, but are conducted after a system has been implemented in a retrospective manner）。

⁵¹ 对算法系统的输入和输出进行评估以确定其是否会产生不公平的偏差（Assessing the inputs and outputs of algorithmic systems to determine if there is unfair bias in the input data, the outcome of a decision or classification made by the system.）

⁵² 对开发、应用人工智能系统的法人主体的内部流程或外部合规性要求进行审计（A review of a company's adherence to internal policies and procedures, or external regulations or legal requirements. Specialised types of compliance audit include system and process audits and regulatory inspection）

⁵³ 由独立主体证实某个产品、服务或组织是否满足特定治理要求（A process where an independent body attests that a product, service, organisation or individual has been tested against, and met, objective standards of quality or performance）

⁵⁴ 证明某个产品、服务或系统满足特定要求，其包括测试、监测、认证等多个环节（Provides assurance that a product, service or system being supplied meets the expectations specified or claimed, prior to it entering the market. Conformity assessment includes activities such as testing, inspection and certification.）

⁵⁵ 对系统满足预先确定的量化指标的程度进行评估（Used to assess the performance of a system with respect to predetermined quantitative requirements or benchmarks）

⁵⁶ 使用正式的统计学方法对系统满足特定要求的程度进行解释说明（Establishes whether a system satisfies some requirements using the formal methods of mathematics）

⁵⁷ <https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques>

（Auditing algorithms: the existing landscape, role of regulators and future outlook）。⁵⁸基于组成 DRCF 的四个部门之间的经验交流，该篇报告对算法审计在算法治理中的地位、算法审计的不同类型及其优缺点、算法审计的未来制度框架等重要问题进行了回答。

该篇报告将算法审计的作用定位于解决信息不对称问题。尽管强制要求信息公开、市场准入前的系统审核、明确的合规标准等都是应对信息不对称问题的治理工具，但算法审计仍然有其独特价值，特别是考虑到人工智能算法技术的动态变化特征，算法审计可能是与之较匹配的治理工具。根据被审计对象的不同，该篇报告将算法审计总结为三种类型：治理审计（Governance Audit）、经验审计（Empirical Audit）、技术审计（Technical Audit）。治理审计关注的是相关治理制度、政策、流程是否满足合规要求，经验审计针对算法系统的输入、输出并考虑其具体影响，技术审计则进一步切入系统内部以发现代码本身问题。DRCF 在该篇报告中没有对算法审计的整体性流程或框架作出说明或勾画，但其非常重要的一点是强调了监管者在促进算法审计中的重要作用。报告认为，算法审计不能由监管者来完全界定，监管者的作用主要体现在以下五个方面：明确算法审计的启动条件（Stating when audits should happen）、确立算法审计的最佳示范（Establishing Standards and Best Practice）、作为更好审计的推动者（Act as an enabler for better audits）、确保审计所发现问题得到救济⁵⁹（Ensure action is taken to address harms identified in an audit）、对可能误导的审计结果进行纠正或解释（Identifying and tackling misleading claims）等。

（3）英国人工智能算法影响评估的总结与反思

基于上述分析，不难发现英国在人工智能算法影响评估制度建设方面的三个特征。

⁵⁸ <https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook>

⁵⁹ 例如增强算法系统的可接入性以使得更多利益相关方能更有效的参与算法审计。

第一，与欧盟、加拿大、美国相比，英国没有就人工智能算法影响评估制度框架及其内容提出任何具体政策设计或政策要求，其仍然坚持了各垂直领域监管者因地制宜形成差异化治理框架的基本思路，要求各领域监管者推动本领域算法影响评估制度的形成与完善。在整体性政策层面，英国当前工作更多是就人工智能算法影响评估制度的不同路径、优缺点等“元制度要素”进行了分析，以为各垂直领域监管者提供参考。

第二，英国在强调各垂直领域监管者推动算法影响评估制度建设方面重要性的同时，也对监管者的具体职责提出了具体要求。在英国决策者的定位中，各垂直领域监管者的重要职责并非具体界定算法影响评估制度内涵，而是要承担“掌舵者”角色，即在明确算法审计要求、推动其他利益相关方更好实现算法审计效果等方面承担相应职责。

第三，英国在算法影响评估制度建设的具体内涵方面，较为强调技术方案的重要性，并试图以此为基础来促进、催生一个新兴产业生态，而这也同样符合英国以“促进创新”作为人工智能治理主要思路的基本原则。

2.3.5. 中国

在形式上，与欧盟、加拿大不同但与美国、英国类似，中国人工智能治理的基本框架并没有通过整体性法律规则的形式制定下来，这也意味着中国仍然体现了“自下而上”而非“自上而下”的治理路径。不过与美国、英国强调各垂直领域的治理革新又有所区别，尽管中国各垂直领域监管机构也在探索传统治理议题在人工智能时代的新变化、新要求，但中国人工智能治理更主要体现出了对人工智能新出现议题的敏捷回应，而这也构成了当前中国人工智能治理的新特色。

（1）中国人工智能治理的基本框架

以2017年国务院发布的《新一代人工智能发展规划》（简称《规划》）为标志，近年来中国人工智能治理快速推进，在顶层设计和具体治理方案上都形成了丰富成果。

一方面，就顶层设计而言，以《规划》为代表，中国决策者在人工智能治理理念原则、体系框架、基础方案等方面都已经出台了重要性政策文本。《规划》首先以时间路线图的形式对中国推进、形成人工智能法律法规、伦理规范和政策

体系提出了基本要求，后续工作都可被视为对该《规划》的具体落实。其次，紧接着的 2019 年 2 月，中国科技部率先推进了机构改革，设立了“国家新一代人工智能治理委员会”，而该委员会在同年 6 月发布了“新一代人工智能治理原则——发展负责任的人工智能”，并在 2021 年 9 月发布了《新一代人工智能伦理规范》，它们都可被视为奠定了中国人工智能治理的基本理念和原则。第三，中国在将人工智能治理工作细化为几个基础方向后，科技部、国家网信办等相关机构则积极展开了统筹规划工作，这又具体包括科技部在 2019 年 10 月启动的“国家新一代人工智能实验”项目、国家网信办在 2020 年 6 月发布的“国家新一代人工智能标准体系建设指南”等。它们都应被视为顶层设计层面覆盖人工智能治理各个领域、各个方向的一般性治理努力，而非垂直领域、针对具体议题的治理方案。例如《国家新一代人工智能标准体系建设指南》在明确伦理安全标准重要性的基础上，就概念术语、数据算法、系统服务、测试评估等人工智能研发应用关键环节的标准建设工作做出了重点部署，但并未对标准的具体内容作出规定。第四，与中国国内人工智能治理工作相并行的，则是中国在人工智能全球治理方向的积极努力，这又特别以 2022 年中国外交部发布的《中国关于加强人工智能伦理治理的立场文件》、2023 年发布的《全球人工智能治理倡议》等为标志。

另一方面，与顶层设计相匹配的，则是中国各部委针对人工智能治理具体议题的治理方案改革与更新，这既包括传统治理框架对于人工智能时代传统议题以新面貌出现时的适时回应，也包括针对人工智能新议题而建构的新治理框架。就前者而言，国家互联网信息办公室出台的《网络信息内容生态治理规定》、《常见类型移动互联网应用程序（App）必要个人信息范围》等文件，以及 2021 年 7 月人社部等八部委发布的《关于维护新就业形态劳动者劳动保障权益的指导意见》都涉及到了人工智能系统的使用边界问题。不过更能体现中国人工智能治理特色的重要治理进程，还是国家网信办出台的一系列针对人工智能治理新议题的治理文件，其主要以具有约束力、执行力的法律文件所构成，具体包括《互联网信息服务算法推荐管理规定》（2021 年 12 月发布）、《互联网信息服务深度合

成管理规定》（2022 年 12 月发布）、《生成式人工智能服务管理暂行办法》（2023 年 7 月发布）这三个为典型代表。这些垂直领域的法律文件提出了算法备案、算法自评估、安全评估等多种新型治理措施，而这也构成了对于一线创新、应用者的实质性规制约束要求。

（2）中国人工智能算法影响评估的制度逻辑

从整体上看，人工智能算法影响评估主要是在国家网信办出台的针对人工智能治理新议题的规制文件（即《互联网信息服务算法推荐管理规定》、《互联网信息服务深度合成管理规定》、《生成式人工智能服务管理暂行办法》等三个）中被系统提出，它们都明确了开展人工智能算法影响评估的必要性，并要求人工智能创新和应用主体开展自评估或第三方评估。同时，作为人工智能算法影响评估的重要组成部分，算法备案的制度框架已经被完全建立并得到了具体落实，而这也成为中国人工智能治理区别于他国的重要特征。

《互联网信息服务算法推荐管理规定》首先提出并落实了算法备案的制度要求，国家网信办建立了全国统一的算法备案网站系统（2022 年 3 月 1 日上线），并要求应用算法推荐技术提供互联网信息服务的相关主体，就服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息履行备案手续。在此规定要求下，截止到 2023 年 5 月，全国范围内 100 余家平台公司已经就 262 种算法提交了备案信息，从而真正落实了算法备案这一制度要求。

在监管者要求提供的算法备案信息中，备案者需要对该算法可能产生的风险影响进行预评估，并就其预防、治理措施作出解释说明。但在公开的备案信息中，算法影响评估的内容并不被包括在内，而直将算法基本原理、算法运行机制、算法应用场景、算法目的意图等内容向公众公开。值得注意的是，中国当前的人工智能治理框架主要是针对私人部门展开，针对公共部门的人工智能算法治理要求尚未出台。

（3）中国人工智能算法影响评估的总结与反思

基于上述分析，不难发现中国在人工智能算法影响评估制度建设方面的三个特征。第一，中国已经将算法影响评估落实到具体机制中，算法备案要求包含了

对于算法影响进行评估的制度内涵，不过究竟哪些对象被纳入评估范畴（例如人工智能算法的环境影响是否是被评估对象）、评估结果将如何被嵌入治理框架（例如是否只关注评估结果为高风险的人工智能治理议题）、评估的流程与结果性要求（例如是否要求评估结果可公开以及向哪些利益相关方公开）等都还尚不清晰。第二，尽管多个规则文本已经将人工智能算法影响评估作为必要的制度要求，但其在人工智能治理整体框架中的制度定位还不是很清晰。例如算法备案系统在治理框架中的定位和价值都还不甚清晰，其是作为系统准入的前提，还是作为事后追责的凭证，仍然具有制度不确定性，而这也将会影响企业开展算法影响评估、落实算法备案要求的具体合规行为（如果作为前提，那自然要求算法备案越细越好；但如果作为事后追责凭证，则企业很可能会避免写的过细）。第三，由于中国正在考虑制订统一的人工智能治理法规，人工智能算法影响评估的制度框架也可能伴随此“自上而下”的统一框架而逐渐成形。但就目前而言，中国仍然体现了不同于欧盟、加拿大和美国、英国的另一条治理道路特点。

3. 人工智能算法影响评价的治理逻辑：监管者与被监管者的合作

在第一部分对人工智能算法影响评价的制度价值进行理论分析、第二部分对人工智能算法影响评价的三种模式（环境影响评价模式、数据保护影响评价模式、问题列表清单评价模式）以及人工智能算法影响评价的国别差异（欧盟、加拿大、美国、英国、中国）进行实践梳理的基础上，第三部分将再次回到理论分析，讨论人工智能算法影响评价的制度逻辑。在该节，我们要讨论的核心问题是，**人工智能算法影响评价究竟如何实现其治理逻辑，而何种因素又将影响其治理逻辑的实现与否？**

在第一部分我们曾指出，人工智能算法影响评价制度的价值在于，其跳出了“市场-政府”的二分边界，而将监管触手延伸至于市场主体内部，并同时聚焦结果的治理目标调整为两方面的过程性要求：对算法治理的过程进行记录以累积治理经验与知识，以及在风险评价的基础上要求算法设计与应用者及时修改生产流程以作出敏捷回应。在第二部分的实践梳理中，我们也可窥见各国施行的人工智能算法影响评价制度为实现前述两方面目标的准备和努力。但第三部分试图提出的理论问题在于：**人工智能算法影响评价的治理实践，是否能够真正实现其所宣称的两方面目标（这也即“治理逻辑”问题）？到底何种因素又将影响人工智能算法影响评价治理实践的具体运行？**

宽泛而言，人工智能算法影响评价治理逻辑的关键是被监管者能够按照监管者要求认真履行评估义务和要求，并将评估结果反馈至一线流程以有效控制风险、回应治理要求。换言之，人工智能算法影响评价治理逻辑能否成立的关键在于监管者与被监管者要形成“合作”。但另一方面，必然存在的信息不对称乃至“共同无知”现象，使得监管者与被监管者的“合作”并不必然能达到预期效果，而导致二者“分离”的因素也自然成为影响人工智能算法影响评价治理逻辑失效的重要原因。本部分将基于已有文献对于合作监管理论的讨论，在对人工智能算法影响评价治理逻辑进行梳理的同时，解释可能导致其成功或失效的原因，以推进对该制度的深入理解。具体而言，与此相关的研究视角大致可区分为协同治理、管理主义、软治理这三条路径，每一条路径都可被作为人工智能算法影响评价的治理逻辑，但每一条路径都存在其局限性，并可能存在制度漏洞。

3.1. 协同治理 (Collaborative Governance) 视角下的人工智能算法影响评价

算法影响评估在制度上要求将被监管者主体纳入治理范畴,由此构成了多元主体(监管者与被监管者及第三方等)协同以实现治理目标的治理格局,这也自然使得协同治理理论可被视为算法影响评估的理论基础之一。

协同治理理论往往是作为介于政府规制(regulation)和自规制(self-regulation)之间的第三条道路而出现,试图通过避免二者弱点以实现更大的监管灵活性,从而应对日益复杂的社会治理环境。从内涵上讲,协同治理理论通常是指将非政府主体(特别是作为被监管对象的私人主体)纳入治理过程,而这一结构变化事实上包含了规制理论的两方面深层次变革:要求监管者与被监管者的迭代互动以提供监管信息从而克服信息不对称或“共同无知”的复杂治理环境,同时要求从结果监管转向过程监管乃至延伸至被监管者的生产治理过程以回应新形势下的治理需求。但这两方面的治理变革并不会仅仅因为在治理过程中纳入非政府主体而自动实现,监管者与被监管者能否实现“监管合作”才是决定协同治理理论的理论假设能否成立的关键,这在具体机制层面则涉及“灵活性”(flexibility)与“责任性”(accountability)的平衡。

一方面,所谓“灵活性”,是指监管过程需要给予被监管对象以自主决策、自主行动空间,这既包括其对如何收集信息、分享信息、处理信息的自主决策,也包括其对如何探索可行解决方案以回应治理需求、创新治理方案的自主决策,它们都意味着仅仅依靠政府部门作为规制者已经不能满足日益复杂环境的治理需求;但另一方面,灵活性并不意味着协同治理等同于自规制,非政府主体(私人部门)仍然需要像政府主体(公共部门)一般保持对于公众的“责任性”,即其需要在不同程度承担程序性(例如信息公开)或实质性(例如非歧视)的治理责任。但在很多时候,灵活性与责任性是相互冲突(或者此消彼长)的⁶⁰,并因此需要机制上的改革创新以实现二者的平衡。从已有文献的讨论来看,相关机制主要体现为以下三种类别。

⁶⁰ Estlund, C. (2005). Rebuilding the Law of the Workplace in an Era of Self-Regulation. *Colum. L. Rev.*, 105, 319.

第一类机制是透明性机制（Transparency），通过将监管信息向公众开放以实现公开监督，从而在为多元主体提供自主决策空间的同时也将其置于“阳光”下，避免监管俘获并提升多元治理主体的责任性。但值得注意的是，追求绝对意义的透明并不一定有利于协同治理绩效的发挥，因其可能遭到被监管对象的反对或抵制而导致“监管合作”的失败。正因为此，在诸多机制的创设中，透明性虽被纳入其中但并没有被要求向公众全面公开。

第二类机制是惩罚机制，即当非政府主体“不负责任”的时候对其施加惩罚，而根据惩罚类型的差异又可进一步分为“螺旋式惩罚（escalating penalties）”⁶¹或“惩罚性默认（penalty default）”⁶²等。前者会推动非政府主体为避免可能出现的顶格惩罚而主动承受一定成本来接受或参与合作，后者是指设定惩罚的自动触发条件，使得监管者并不作为惩罚发起者并因此避免了与被监管者的对立，进而监管者更是可以与被监管者“合作”以共同避免达到自动触发条件。

第三类机制是激励机制，即为非政府主体的合作参与提供正向激励条件，以使之在考虑合规成本扣减之后仍然能实现正收益。换言之，激励机制需要推动被监管者同样能在共同监管的环境中受益，而不仅仅是付出合规成本。考虑到市场竞争的必然存在，激励机制的实现往往需要改变市场环境，而非仅仅针对被监管者个体施加约束条件。

由此，从协同治理要求纳入非政府主体的基本内涵，到其在机制设计上需要平衡灵活性和责任性的关键，再到实现二者平衡的三类机制（透明、惩罚、激励），本小节完整回顾了协同治理的基本理论框架。以此为起点，我们也不难发现算法影响评估制度的治理逻辑。事实上，算法影响评估的治理实践在多个方面都体现了协同治理的理论内涵及理论要求。

在宏观理念上，协同治理作为政府规制与自规制之外第三条道路的理论定位，正与算法影响评估的制度初衷不谋而合。算法影响评估同样是在面对人工智能算法应用环境复杂性时的一种必然选择，本报告第一章是对此的完整阐释。

⁶¹ Ayres, I., & Braithwaite, J. (1991). Tripartism: Regulatory capture and empowerment. *Law & Social Inquiry*, 16(3), 435-496.

⁶² Sabel, C., & Simon, W. H. (2017). Democratic experimentalism. *Searching for Contemporary Legal Thought* (Cambridge and New York, Cambridge University Press, 2017), *Columbia Public Law Research Paper*, (14-549).

同时，平衡灵活性和责任性的三种具体机制，也都在不同程度上反映在了算法影响评估的各类实践中。例如 2.2.1 节讨论的环境影响评价“四个关键点”事实上同样扮演了“惩罚性默认”的机制作用，因如果在前三个关键点能满足合规，则将自然避免在第四个环节仍然需要完成冗繁沉重的评估工作；类似的，2.2.2 节讨论的数据保护影响评估放弃了“向公众公开”的透明性要求，因唯有如此才可能将被监管对象拉入“监管合作”空间而非走向针锋相对。

不过同样值得注意的是，协同治理看似完整、精巧的理论设计并不代表其必然能够有效，其不足仍然十分明显，并主要体现在以下两点。

一方面，从本质上讲，协同治理在实践中往往被视为一个程序性要求（procedural requirement）而非实质性要求（substantive requirement）。这一倾向在美国环境影响评价制度的实践中被最高法明确地表达并被执行了下去。⁶³ 尽管 NEPA 在法律上同时包含了程序性要求和实质性要求⁶⁴，但出于对司法权力与行政权力分割边界的尊重，美国司法机构不愿意干涉行政决策过程⁶⁵，因而其并不对案件中联邦机构所开展环境影响评价工作能否实现环境保护目标的实质性问题作出审查，而主要关注其是否执行了相关程序性要求，并按照程序性要求完成或简化完成了环境影响评估工作。在此意义上，诸多评论者认为司法机构在事实上简化并扭曲了立法者意图（即忽视了环境保护这一实质性目标的实现与否）⁶⁶，而这也构成了我们对影响评估这一类制度究竟能否实现实质性目标的担忧和质疑。

另一方面，协同治理试图将非政府主体纳入治理过程的基本设想很可能是难以实现的，其在具体执行机制上会面临诸多挑战和困难。究竟谁是利益相关方（并因此将其纳入“协同”范畴），谁掌握监管信息，如何能够推动监管者与被监管

⁶³ Orts, E. W. (2019). Reflexive environmental law. *Environmental Law*, 347-460.

⁶⁴ NEPA 中的程序性要求条款较多，实质性要求条款主要是第 101 款（Section 101）以及第 102 款第 1 条（Section 102 (1)）。参见 Yost, N. C. (1990). NEPA's Promise-Partially Fulfilled. *Envtl. L.*, 20, 533.

⁶⁵ 在相关判决书中，最高法官提出司法不能成为行政决策的一个环节（decision-making unit of the agency）。参见 Ferester, P. M. (1992). Revitalizing the National Environmental Policy Act: Substantive Adaptations from NEPA's Progeny. *Harv. Envtl. L. Rev.*, 16, 207. P.218.

⁶⁶ Yost, N. C. (1990). NEPA's Promise-Partially Fulfilled. *Envtl. L.*, 20, 533; Hodas, D. R. (1999). NEPA, Ecosystem Management and Environmental Accounting. *Nat. Resources & Env't.*, 14, 185.

者的反馈沟通而非“一锤子买卖”，都是协同治理理论尚未回答、难以回答的具体问题。正因为此，算法影响评估也同样不能被视为包治百病的“银弹”，对其局限性的考量仍然需要被纳入到人工智能算法治理的整体框架之中。

3.2. 管理主义（Managerialism）视角下的人工智能算法影响评价

协同治理反映了算法影响评价要求多主体参与的治理格局，而为了促进多主体参与并发挥各自作用，协同治理认为应重视能够平衡“灵活性”和“责任性”的机制创新。如果我们按照协同治理理论指引完善制度框架，算法影响评价制度建设是否就能够实现其释放技术变革潜力、控制技术应用风险的治理目标呢？对此的答案很可能也是否定性的，原因不仅在于操作层面（例如能否确认利益相关方范畴等技术性问题），还在于制度逻辑层面。

在 3.1 节结尾的相关反思中我们曾提到，协同治理（以及更具体的算法影响评价制度）在实践过程中往往被视为程序性的，而非实质性的。导致这一结果的原因并不仅仅因为美国语境下行政权力与司法权力分立的制度背景（也是其制度特殊性），还在于协同治理（或算法影响评价）作为管理主义的具体体现而包含的制度前提、制度假设存在瑕疵。事实上，当前大部分关于协同治理的文献都可被视为“管理主义”（Managerial）视角及理论的具体延伸，而作为一种理论流派，管理主义本身同样有其存在的时代价值以及随时代变化而体现出理论缺陷的时代局限性。正因为此，从管理主义视角对算法影响评价制度的反思将更全面展现出其治理逻辑的优势与不足。

从概念上讲，“管理主义”可被定义为“通过一系列流程设计以提升批量解决问题的效率的一套方法论，但却往往忽视了结构性或系统性不公平等实质性问题的解决”⁶⁷。管理主义的出现，被认为缘起于 20 世纪早期社会危机的应对，并因此与“新自由主义（Neo-Liberalism）”意识形态紧密关联，而在此理论关联性中我们也不难理解“管理主义”作为一种制度设计理论的进步与局限。⁶⁸20 世纪早期的社会危机引发了公众对于“自由放任资本主义（laissez faire capitalism）”的不满，但西方理论家同时也对走向集体主义、社会主义的革命

⁶⁷ Cohen, J. E. (2019). *Between truth and power*. Oxford University Press. p.143.

⁶⁸ Hanlon, G. (2018). The first neo-liberal science: Management and neo-liberalism. *Sociology*, 52(2), 298-315.

路径持有怀疑态度，如何在维系市场竞争的同时保护社会秩序便成为引导理论创新的问题窗口。在此背景下，管理主义作为新的理论道路被提出，并在埃尔顿·梅奥（Elton Mayo）、马克思·韦伯（Max Weber）等的推动、丰富下最终演化为体系化的理论框架。

对于梅奥而言，其关心的核心问题是在生产过程中如何实现“自发的社会合作（spontaneous cooperation）”，即既能以分散化自由决策为基础（从而发挥市场机制的价值以避免政府干涉扭曲市场行为），同时也能实现集体生产的规模性和社会性（从而避免市场对社会的排挤）。梅奥的答案便是“管理主义”，其一方面通过管理者作为精英的知识创造来设计出“常规性合作（routine cooperation）”的内容和框架，使得大众不需要深入思考即可参与生产过程并避免绝对民主的混乱，另一方面大众也在被管理的过程中重新塑造自身的预期和行为，并最终接受管理主义以形成新的社会联系和社会框架。⁶⁹与之相比，韦伯虽然从不同角度切入，但在“官僚制”的思想内核上仍然与梅奥有着内在的一致性。韦伯将官僚制作为社会理性的化身，但其同样也认识到官僚制对个体自由以及市场竞争的侵蚀⁷⁰，而体现“克里斯马权威”的官僚领导者正是平衡官僚制进步与不足的关键所在。在韦伯的视野下，管理者与大众的分离同样是平衡市场竞争与社会团结的前提。

基于以上梅奥和韦伯观点的回溯，已有文献将管理主义的核心特征总结为三点，而这也是其回应 20 世纪早期社会危机的主要“药方”：对自发行为的持续性干涉、强调个体间竞争、精英领导的必要性。⁷¹这三点特征与“新自由主义”的基本主张如出一脉：首先，与自由放任资本主义不同，新自由主义同样强调市场竞争需要在不断的干涉和保护下才能得到维系和扩展，因竞争中形成的利益集团总是试图逃避市场竞争的压力；其次，为避免市场与社会的分离，甚至是市场对社会的排挤，竞争仍然需要被视为首要的回应机制，这既是缘于竞争作为一种激励方式而对自发动力的激发（进而将劳动者转化为自我负责的“企业家”），同时也是缘于竞争作为一种管理方式而对大众施加的工作压力（从而使之避免走向社

⁶⁹ Hanlon, G. (2018). The first neo-liberal science: Management and neo-liberalism. *Sociology*, 52(2), p.10.

⁷⁰ 如其所指出，资本主义与国家官僚利益集团之间的联盟是对市场竞争的巨大威胁。参见 Hanlon, G. (2018). The first neo-liberal science: Management and neo-liberalism. *Sociology*, 52(2), p.12.

⁷¹ Hanlon, G. (2018). The first neo-liberal science: Management and neo-liberalism. *Sociology*, 52(2), 298-315.

会革命的集体行动)；第三，精英作为管理者的领导与组织工作，对于实现前二者必不可少，这又具体体现为工作的选择与招聘、职业路径的构建、发展愿景的勾勒、常规工作实践的塑造等各个方面⁷²。正是由于这三点共性特征，“管理主义”与“新自由主义”才被视为同一个思想内涵在不同领域、不同层次的差异化表现，但也正因为此，针对“新自由主义”的反思与批判恰恰也适合被用于“管理主义”，而这又主要体现在两个方面。

一方面，管理主义（或新自由主义）作为一种“亲市场”的制度规范往往将社会导向政策置于市场导向政策目标之下，效率会优先于正义目标并因此可能不利于后者的实现。⁷³这一批评尤其在数字时代可能将更为凸显。乔治敦大学法学教授朱莉·E·科恩将信息资本主义（informational capitalism）时代所涌现的治理问题特征总结为“无形资产侵害（problem of intangible harms）”和“大规模侵害（problem of numerosity）”两类。为同时回应这两类问题，司法部门采取了管理主义的改革路径，针对不同问题提供了差异化的解决方案，由此有别于工业革命时期所形成的统一裁判框架。这一改革思路固然提升了问题处理效率，却不可避免地带来了结构性分化结果：典型案例就是知识产权或其他产权性问题因有利于市场主体而在互联网商业化历程中备受关注，但隐私保护、劳动者权益等议题却几乎被忽略。⁷⁴在管理主义的理论视角下，这一分化几乎是不可避免的，亲市场的制度规范将自然限制社会性政策目标的实现。

另一方面，管理主义（或新自由主义）为精英管理者提供权力与自主决策空间进而维系市场与社会新平衡的解决思路，仍然可能简化了管理的复杂性，并使得手段与目标“南辕北辙”。梅奥和韦伯之后的管理理论研究，已经较为丰富的指出，管理者本身也可能存在“道德风险”，其并不一定会为了微观组织的生存乃至整个社会的稳定而勤勉工作，管理者与劳动者的权力分离同样可能使前者为了自身个体利益而损害集体利益，如何在机制上保证管理者的可信承诺同样应成为前述三个管理主义核心特征之外的第四个关键机制。⁷⁵基于后续理论的讨论我们不难发现，管理主义本身并非逻辑自洽的统一理论，其同样可能演化为管理者

⁷² Hanlon, G. (2018). The first neo-liberal science: Management and neo-liberalism. *Sociology*, 52(2), p.5.

⁷³ Cohen, J. E. (2019). Between truth and power. Oxford University Press. p.155. (stating, managerialism as a form of institutional discipline that has gradually but inexorably swept the judicial system into its orbit)

⁷⁴ Cohen, J. E. (2019). Between truth and power. Oxford University Press. p.145.

⁷⁵ <https://mp.weixin.qq.com/s/EklGffk1Zxv8UxIT5oK7QQ>

模糊（甚至替代）治理目标的“托词”。

上述对于管理主义的理论梳理与反思分析，对我们理解算法影响评估的治理逻辑同样具有启发价值。作为一种理想模型，算法影响评估的制度逻辑被解释如下：其反对自由放任的算法技术创新与应用行为，并要求将监管压力延伸至算法开发与应用过程；同时，考虑到算法治理风险及治理方案的不确定性，一线主体的自主探索是受到鼓励的，监管者希望被监管者之间形成治理竞争格局，以在相互学习的过程中积累治理知识。但从管理视角的反思来看，这一理想模型的制度逻辑很可能是难以成立的。

一方面，从效率优先的目标来看，算法影响评估仍然可能忽视或相对忽视社会性目标的实现。事实上，从 2.3 节的分析不难看到，在已有的政策实践中，算法影响评估的主要标的仍然聚焦消费者法定权益（包括隐私权、价格歧视等），因其不仅涉及违法违规的惩罚问题，也是维系市场主体声誉的关键变量；但相比之下，算法对环境的影响、算法对社会公平的影响等都是较为模糊而未引起充分重视的治理议题。这一反思同样体现了算法影响评估制度建设过程中可能出现的议题分化、议题偏差现象。

另一方面，利益相关方仍然有可能质疑算法影响评估的真实效果，一线主体是自发、主动开展算法影响评估进而提升算法治理水平，还是仅仅将此当作“洗白”借口与托词。已有研究已经开始批评人工智能技术创新和应用企业将人工智能伦理治理作为逃避“硬监管”借口的虚假行为，同时在部分实证文献中，我们也逐渐发现人工智能伦理治理原则并没有在事实上改变技术人员的开发与应用行为，而利益相关方在很大程度上也并不信任企业的人工智能算法伦理治理效果。换言之，当技术企业管理者不能向社会给出可信承诺时，算法影响评估的制度效果仍然将大打折扣。

值得指出的是，管理主义视角的上述分析更多是将被监管者的自利行为视为前提来考察算法影响评价制度设计的不足与漏洞，其假设算法影响评价是被监管者的负担并因此会受到后者的抵制，因此我们才需要监管者的干涉以促进被监管者行为的改变。但另一方面，我们在现实中同样可观察到的现象是，作为被监管者的企业或个人往往也会采取超出规制要求范畴的主动行为，以减少其行为对其他利益相关方的影响。例如在环境影响领域，造纸厂会主动投入合规要求之外的

更多资源来消除恶臭⁷⁶；在隐私保护领域，我们同样看到苹果公司超出合规要求来强化操作系统的数据保护程度，从而使得每一个应用在索取用户数据时都需要得到用户的明确同意⁷⁷。面对这些超出合规要求的治理行为，我们同样需要解释其背后的治理逻辑：即在具有约束力的法律规则及监管者压力之外，哪些因素还会影响企业或个人等一线主体的治理行为？

该领域的零散研究对此提出了多种解释因素，但总结起来，其大致可被认为是企业在面对社会期待（social pressure）与经济状况（economic constraints）双重约束条件下的理性选择。⁷⁸例如在针对卡车产业的环境影响治理行为研究中发现，对于竞争压力大的企业而言，社会期待和规范性压力并不能促使其增加环境治理投入；而对于拥有更多资本的大公司而言，出于声誉考虑其将进行超出合规要求的环境治理投入。⁷⁹不过这一解释框架却与算法影响评价制度的具体实践不完全一致：在 2.3.1 和 2.3.3 节对于欧盟和美国人工智能治理框架的比较中可看出，美国将中小企业作为例外而豁免在人工智能监管范畴之外，但欧盟《人工智能法案》则强调为中小企业提供相匹配的合规资源而并不将其排除于监管范畴。从超出合规行为的制度解释研究来看，大企业相比于中小企业会更多考虑社会期待并因此有更大的合规动力，因而豁免中小企业的美国模式并不一定有效，欧盟模式反而是符合这一理论框架的政策实践。

不过我们仍然需要注意的是，社会期待与经济状况的交互影响还并非解释一线主体超出合规要求治理行为的唯一制度逻辑。从行动者视角（activism）来看，

“政策企业家（policy entrepreneurs）”的作用同样不可忽视。换言之，即使对于相同经济规模并面临类似社会压力的企业而言，企业内部是否存在“政策企业家”以推动治理行为至关重要。例如针对人工智能产业中伦理治理实践者的研究表明，公平人工智能（AI Fairness）研究往往是企业内部个别倡议者推动临时性

⁷⁶ Gunningham, N., Kagan, R. A., & Thornton, D. (2004). Social license and environmental protection: why businesses go beyond compliance. *Law & Social Inquiry*, 29(2), 307-341.

⁷⁷ <https://www.forbes.com/sites/kateoflahertyuk/2021/01/31/apples-stunning-ios-14-privacy-move-a-game-changer-for-all-iphone-users/>

⁷⁸ Gunningham, N., Kagan, R. A., & Thornton, D. (2004). Social license and environmental protection: why businesses go beyond compliance. *Law & Social Inquiry*, 29(2), 307-341.

⁷⁹ Thornton, D., Kagan, R. A., & Gunningham, N. (2009). When social norms and pressures are not enough: Environmental performance in the trucking industry. *Law & Society Review*, 43(2), 405-436.

议程的结果，而并不一定是企业自上而下的有组织行为。⁸⁰正因为此，就组织管理而言，为“政策企业家”提供活动空间可能是比直接推动治理改革更有效的手段。

3.3. 制度同化（Institutional Isomorphism）视角下的人工智能算法影响评价

无论是认为被监管者会逃避算法影响评估的监管压力，还是认为被监管者会采纳超出规制要求的治理行为，3.2 与 3.3 节分析的治理逻辑都是以其作为理性自利主体为基础，试图寻找影响被监管者“成本-收益”计算的相关因素，并在此框架下探索如何通过制度设计来改变被监管者的理性计算环境，从而引导其配合开展算法影响评价的合规行为。但制度主义的分析路径不仅仅只包含理性计算，以制度同化理论为代表的非理性路径同样构成了解释制度形成及演化过程的重要视角，这也由此构成了理解人工智能算法影响评价制度逻辑的第三种视角。

制度同化理论的基本假设是认为，某个组织是否采纳以及如何执行某项制度，并不取决于其对所在环境的理性判断以及在此判断基础上作出的最优决策，而是缘于社会化过程，即在受到社会环境影响过程中而不断同化的过程。在制度同化理论下，相同环境下的所有组织所采纳的制度框架及其执行范式都将走向趋同化。导致趋同的机制包含三类：强制（coercive），模仿（mimetic）与规训（normative）：第一，强制机制是指外部监管压力将迫使相同环境下的所有组织采取类似的治理行为，而无论组织面临的具体“成本-收益”结构是否相同；第二，模仿机制是指当组织面临不确定环境而难以计算“成本-收益”结构时，其更多选择模仿同行选择以获得合法性或者避免追究责任；第三，规训机制的基础是行动者视角，其认为各类组织将雇佣从同一专业体系下规训出来的技术专家，而后者的类似思维将同样影响组织本身的同化进程。⁸¹再次需要强调的是，这三类机制都与特定

⁸⁰ Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020, April). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

⁸¹ P.172

组织的理性计算没有必然联系——事实上，基于该三类机制所形成或演化的制度范式，很可能与理性计算的结果相反。

制度同化理论视角对于我们理解人工智能算法影响评价制度在各个企业的建立、执行过程及其最终绩效，同样具有启发意义。考虑到算法影响的不确定性，究竟对哪些对象进行评价、评价到何种程度、评价结果应如何公开等制度要素都存在极大模糊空间，在此意义上一线主体事实上很难对评价行为本身的“成本-收益”结构作出精确判断，按照规制者要求的强制执行、模仿同行、在内部成立评价部门或者寻找类似第三方展开评价等都是可能被观察到的制度同化现象。同时，如果沿袭制度同化的理论视角，2.3 节所描述的不同国家采取的多样化、差异化算法影响评价模式可能只是阶段性现象，未来各国制度是否会趋同仍然是值得观察的重要现象。加拿大与欧盟的类似性、英国与美国的类似性以及中国是否会参考欧盟模式等，都是具有一定现实基础且可能发生的未来演化。

值得指出的是，制度同化理论的分析视角对于我们未来改革、优化算法影响评价制度也具有启发价值。如果考虑到不同组织之间存在相互模仿现象，那么首先推动行业领军企业（例如大型数字平台企业）开展算法影响评价并探索其具体模式，可能是具有示范意义的，而其他企业则可能在模仿过程中提升算法治理水平；同时，在人员层面推动职业伦理教育、形成行业规范模式，也可能是促进算法影响评价制度成型并扩散的重要机制。

4. 人工智能算法影响评价的未来改革与方向

基于前三个部分的分析，本报告第四部分将聚焦人工智能算法影响评价的未来改革问题。需要指出的是，考虑到算法影响评价的实践性较强，存在较为丰富的治理工具可供选择，在影响评价的若干环节中，技术社群和产业业界也涌现出了诸多实践策略，而这些实践细节都难以在一篇报告中被全面综述。本报告第四部分更多定位于未来改革思路和方向的分析，并在此基础上对算法影响评价体系架构和关键环节提出建议。这既是建立在前三个部分分析基础上的自然延伸，也为理解算法影响评价的整体改革路径提供了参考，因而同样具有其价值意义。

4.1. 人工智能算法影响评价的制度要求：开放性、合作性、责任性

前三个部分对人工智能算法影响评价的实践模式、各国策略、制度逻辑都做出了较完整的阐释。基于正、反两个方面的对比分析，我们不难看出，一个完备的人工智能算法影响评价制度，其本身需要包含开放性、合作性、责任性这三方面核心特征。

4.1.1. 开放性

开放性是指人工智能算法影响评价应在评价与被评价主体、评价对象、评价方法、评价工具等多个维度上避免边界的封闭与割裂，持续与所处环境发生关联，并在此关联过程中累进式地积累治理经验。之所以要求开放性，可从以下三个方面得到解释。

第一，就治理目标而言，人工智能算法影响评价制度是具有双重性的，其一方面要求针对特定治理风险进行评估从而为救济做准备，另一方面则是要求积累治理经验以持续性地提升治理水平。考虑到人工智能算法治理领域仍然存在诸多“已知的未知（known unknowns）”和“未知的未知（unknown unknowns）”，第一个目标的实现事实上是有难度的，而第二个目标才更为实际。在此背景下，人工智能算法影响评价的意义即在于“自下而上”地构建治理知识生成与积累的通道，该通道因此也必然要求是开放性的，使得不同主体、不同方法都可以被纳入其中。

第二，就治理环境而言，人工智能算法影响评价制度要面对的是一个复杂、多样、动态的治理环境，理论上预先划定的治理边界在面对现实治理环境时往往难以匹配，并因此需要开放面对所有可能性以作出调整应对。2.2.3 节所提到问题列表清单模式是典型的封闭模式，尽管设计者可以尽职准备以尽可能的穷尽问题从而为评估者提供框架，但这一努力在现实治理环境面前仍然将面临本质上的挑战。不同治理环境下相同的治理问题与答案可能会包含不同含义，同一个治理问题在不同环境下可能会出现差异化的治理答案，这些现象都要求人工智能算法影响评价需要保持对环境的开放性，以维持与环境的动态匹配。

第三，就治理机制而言，人工智能算法影响评价要求形成动态、迭代的反复

过程，而这一过程的前提即要求制度上的开放性，以使得利益相关方能够以多重渠道、不同形式参与其中。换言之，开放性是合作性和责任性的前提与基础。

4.1.2. 合作性

人工智能算法影响评价的“合作性”主要是指不同主体间要形成动态、反复、迭代的持续治理过程，并在此过程中累积治理知识、回应治理问题。合作性是决定人工智能算法影响评价成功与否的关键，这又主要体现在监管者、被监管者、第三方这三个治理主体的不同侧面。

对监管者而言，“合作性”意味着监管权力的分散，以及在此过程中与被监管者关系的改变。虽然算法影响评价可以作为—种监管压力施加在被监管者身上，但本报告第一部分提及的算法技术研发与应用的特殊逻辑，使得监管者难以沿袭传统的“命令—控制”模式来实现治理目标。被监管者主动寻找评价方案和风险救济措施，不仅是重要的，也是必要的。在此意义上，监管者的工作重心便从传统意义上监督、惩罚被监管者，转变为了以机制创新激励或“倒逼”被监管者的主动行为，后者也即“合作性”的本质。

对被监管者而言，“合作性”既意味着其自主性的维系，但同时也意味着其需要接受合作框架的约束并承担责任。—方面，自规制模式下的自主性在算法影响评价体系下仍然是必要的，信息不对称和共同无知现象的存在，仍然使得更具有资源和信息相对优势的被监管者需要扮演更为积极主动的作用；另—方面，自规制模式缺少约束力的不足仍然使之可能流于形式，包括惩罚性默认在内的诸多机制创新则为推动被监管者进入合作关系建立了边界条件。在此意义上，“合作”对于被监管者而言，不仅是权利，也是义务。

对第三方而言，“合作性”意味着需要发挥更多主体的作用以建立更丰富、多层次的治理框架与体系。监管者与被监管者尽管是算法影响评价的核心主体，但考虑到二者各自“道德风险”的存在可能性，第三方主体的纳入才可能真正推动合作的实现：监管者可能被“监管俘获”，被监管者可能出于保护商业秘密的需要而拒绝开放，第三方主体的引入—方面可能监督监管者行为，另—方面可能为被监管者提供利益平衡的渠道（例如向第三方开放而不向公众开放）。

4.1.3. 责任性

就制度属性而言，开放性、合作性都属于程序性要求，但正如 3.1 和 3.2 的反思性分析所体现的，影响评价作为一种制度，其对于程序性要求的强调可能淡化实质性目标的实现，而这无疑是需要未来改革加以关注的重点，责任性要求便可被视为对此不足的补充和完善。

责任性要求是指算法影响评价需要在一定程度上回应实质性治理目标的实现与否，通过设定明确的、最低限度的合规标准以确保被监管者落实治理责任。从治理逻辑来讲，责任性要求的提出是对影响评价制度的一种修正。事实上，如果考虑到影响评价的制度初衷是激发被监管者参与治理行为的动机，那么其前提必然是监管制度本身存在宽泛、模糊地带，从而能够为被监管者提供可自由裁量与探索的空间，而责任性要求对实质性治理目标的纳入将削弱这一前提，使得影响评价制度具备了一定刚性——这也是为什么责任性要求仅包含最低限度合规标准的原因所在。

但另一方面，这种修正是否会从根本上动摇影响评价制度的根基，仍然是值得讨论的重要议题。具体而言，影响评价制度绩效的关键是要实现监管者与被监管者的合作关系，而实质性目标的设立将使二者转变为“敌对”关系，即监管者需要监督被监管者是否按照实质性目标标准严格落实。在此意义上，被监管者便可能止步于该目标的实现与否，而不再做进一步的治理努力和探索。若如此，所谓“最低限度合规标准的责任性要求”反而将成为影响评价制度崩溃的根源。对此质疑的回应需要回到 3.2 节讨论的被监管者超出合规要求而开展治理行为的原因分析中，而究竟哪种机制（停留于最低合规要求，还是超出合规要求）会扮演主要作用则更多表现为实证命题而非理论演绎命题。

在捋清责任性要求制度逻辑的基础上，究竟如何充实其制度内涵便是第二个重要问题。考虑到算法影响评估的宽泛性，到底哪些内容应该被纳入最低限度的实质性治理目标范畴，以及如何设定“最低限度”，仍然是需要探索的实践性问题。从 2.3 节对各国具体实践的梳理来看，个人基本权益保护（例如隐私保护、反歧视保护等）是各国算法影响评价制度具有共识性的要素选择，而在公共安全、环境保护等方面则体现出了差异性。因此，如果从寻求最大共识程度来讲，个人

基本权益保护应被视为责任性要求的必要内涵，但这仍然并非是唯一选择；如果从寻求最广覆盖度的角度来讲，公共安全、环境保护则同样应被纳入其中。究竟如何选择仍然是留给决策者的实践问题，但在算法影响评价制度中纳入实质性目标的责任性要求，应成为各国决策者的基本共识。

4.2. 通用人工智能算法影响评价的体系框架：理论基础与层次结构

在 4.1 的基础上，接下来的实践性问题便在于如何构建人工智能算法影响评价制度。在很大程度上，对该问题的回答仍然需要回到实践中去总结和反思，而本报告定位于制度分析的基本逻辑，使得我们并不致力于人工智能算法影响评价操作手册的撰写，而仍然试图提供整体性的框架思考与具有启发性的理论引导。围绕这一目标，4.2 节将从三方面展开：4.2.1 的“理论基础”将介绍本报告认为具有前沿启发性的理论创新，以为实践者搭建人工智能算法影响评价制度体系和操作手册提供指引；4.2.2 的“层次结构”将从技术、组织、社会三个层面解构人工智能算法影响评价的基本框架和思路，报告的主要观点是人工智能算法影响评价是一个涉及技术、组织、社会三个维度的系统性工程，需要对它们进行综合考量；4.2.3 的流程指标将以人工智能算法应用对权益、秩序、规则的影响为例，在操作层面提出影响评价的流程和指标构建。

4.2.1. 理论基础：心理计量学与机器行为学的理论融合创新

人工智能算法影响评价作为制度要求在各国的普遍提出，与之匹配的则是技术社群与企业主体为落实该制度要求的实践探索。事实上，围绕人工智能算法治理的不同目标（可靠性、安全性、公平性等）已经发展出了丰富的评价指标与工具，但这并不意味着人工智能算法影响评价已经形成了从制度到工具的完整体系，人工智能技术本身的快速发展仍然需要理论与实践的及时回应。

就当前而言，人工智能较为关键的进展便在于通用人工智能相对于专用人工智能的进步与应用。以 2022 年下半年 ChatGPT 的横空出世为标志，原本聚焦垂直领域、以完成特定功能为目标的专用性人工智能，已经在大数据、大算力、大

模型的联合推动下“涌现”出了通用能力，并形成了通用人工智能的雏形。尽管仍然存在诸多批评⁸²，但不可否认的现象是，以 ChatGPT 为代表的大型语言模型表现出了跨领域的通用能力，在体现出对人类社会常识“了解”的基础上，能够将其用于各个场景并完成多样化任务。在此背景下，传统地以特定场景下特定功能实现与否、特定价值遵循与否为标的人工智能算法影响评价工作，需要适应通用人工智能的兴起与变迁。具体而言，在难以提前预测人工智能应用场景和功能的前提下，我们应如何开展人工智能算法影响评价工作，评价什么、如何评价？

针对此问题，前沿研究提出应借鉴心理计量学的知识框架，以引导通用人工智能的影响评价工作的探索。但同时，心理计量学的运用依赖于我们对于人类行为模式的理解与建构，而人工智能算法作为被评价主体，其“行为”模式与人类存在较大差别。在此意义上，将“机器行为学”引入并与心理计量学结合，可能可为未来人工智能算法影响评价工作提供新的理论指引，而这也是本报告的创新所在。

（1）心理计量学对于通用人工智能算法影响评价的启示

与垂直领域人工智能算法影响评价相比，通用人工智能算法影响评价工作存在三方面的具体挑战：第一，通用人工智能的应用场景、应用方式存在不确定性，因而难以通过提前预定其应用目的与方式而设定测试环境，也不能通过其在具体任务上的表现来评价其在其他任务（或在一般任务）上的影响；第二，通用人工智能影响评价的指标体系需要重新建构，针对垂直领域人工智能影响评价的现有指标体系不能满足治理需要；第三，考虑到 ChatGPT 的输出与输入提示词紧密关联，存在针对输入的敏感性，因而如何对“通用人工智能影响评价结果”进行评价，仍然存在方法论上的挑战。针对上述三个问题，心理计量学的已有研究可以给予回应。

第一，心理计量学在其长期发展过程中已经抽象出了一系列的一般/通用概念，而这些概念能够被用于通用人工智能算法影响评价。垂直领域人工智能算法

⁸² 对 ChatGPT 的批评较有代表性的两位例如加里·马库斯（Gary Marcus）和杨立昆（Yann LeCun）。马库斯认为 ChatGPT“只是语言顺序的模型，而并非世界如何运行的模型”。换言之，尽管 ChatGPT 能够生成合乎语言逻辑的内容，但这些内容并不一定是现实本身。杨立昆则认为，大型语言模型只是“浅层理解（shallow understanding）”而远不能接近人的智能。他还进一步提出，“在通往人类级别的智能道路上，大型语言模型是一条歧路（off-ramp）”。

影响评价的传统思路是针对具体任务展开，其并不重视能够横跨多个任务领域的一般/通用概念的抽象与提炼，而后者正是心理计量学的发展重点，并体现为“概念导向的评价（construct-oriented evaluation）”。在心理计量学看来，这些概念反映的是人类行为的“隐变量”，其是具体外化行为的决定因素，并反映了不同外化行为的共性特征。举例而言，当我们需要对某个人的学术潜力进行评价时，心理计量学并不着力于该人在具体学科、具体知识上的绩效表现，而是通过“个性(personality)”、“认知能力(cognitive ability)”、“价值取向(value orientation)”这三个隐变量来加以评价。换言之，无论其在具体学科、具体知识上的表现如何，通过这三个一般/通用变量，心理计量学也能够对该人的学术潜力做出评价，因为正是这三个变量才决定了其在任意学科、任意知识领域的绩效表现。在此意义上，当我们需要对通用人工智能影响进行评价时，心理计量学已经发展出的这一套一般/通用概念（或“隐变量”）仍然是具有启发意义的。

第二，基于对人类行为模式、规律的把握，心理计量学对一般/通用概念所作出的理论化、体系化解构，并在此基础上形成的一套层级化指标体系，将有利于通用人工智能影响评价体系的细化与建构。在其长期发展历程中，心理计量学不仅梳理、总结了能够体现人类行为一般模式、特征的通用概念（隐变量），围绕这些概念/变量的测度也形成了具有内在逻辑关联、层次化、可供测度的指标体系。例如“福利（well-being）”作为社会科学的重要概念，在心理计量学范畴下即形成了多个层次的测度指标体系，其具体包括个体层面的快乐（与开心、高兴相关）、满足（与目的、意义相关），以及集体层面的社会福祉（与集体心理有关）等不同侧面。借鉴心理计量学已经发展出的指标体系，我们同样可以发展出针对通用人工智能算法影响评价、具有可操作性的指标体系。

第三，通用人工智能的输入敏感性问题，在心理计量学范畴下可以通过“信度”与“效度”的一般方法论来加以回应。“信度”是指某个概念的测度指标在多大程度上接近该指标的全部内涵，而“效度”是指多次测度结果的稳定性与一致性程度。围绕这两个概念而形成的方法论体系，同样可以用来评价“影响评价”本身。

正是基于上述三点解释，我们不难发现心理计量学对于通用人工智能算法影响评价的启示价值。当然，究竟应该如何建构通用人工智能算法影响评价的概念（隐变量）体系，以及如何围绕该概念（隐变量）体系形成可测度的指标框架，进而采用何种方法对算法影响评价工作的“信度”和“效度”做出解释性说明，仍然需要在具体工作中才能进一步展开，这取决于我们如何定位所要评价的对象，以及如何解释被评价对象的行为规律、特征。举例而言，在心理计量学中常被应用的理论模型是“项目反应理论（Item Response Theory, IRT）”，该理论被用于评估某项测试与被测试者某项能力（隐变量）之间的关系。IRT 同样可以被用于对通用人工智能进行评价，不过此处的差别在于，通用人工智能不同于人类行为的特质，使得我们虽然可以沿用 IRT 的基本思路，但其项目特征曲线可能是存在本质差异的。换言之，在心理计量学之外，我们还需要另外的一般性理论对人工智能的行为模式、特征进行概括，从而在此基础上才能衍生、发展出人工智能算法影响评价的完整模型。结合当前的理论发展，“机器行为学”（machine behavior）可能能够满足此种要求。

（2）“心理计量学+机器行为学”的理论融合创新

“机器行为学”是借鉴行为学的研究框架以分析作为人造物的机器的形成、演化与影响规律，其理论源头可被追溯至赫伯特·西蒙在 1960s 年代关于“人工科学”问题的讨论。在西蒙看来，“尽管不同于自然现象服从于自然法则而具有一种‘必然性’，人工现象由于易被环境改变而具有‘权变性’，但人工现象不会完全适应环境仍然反映了其同样具有凌驾权变性之上的必然性”。也正是因为这种必然性的存在，才使得我们可以以“科学”研究的方法来探究“人工现象”的基本规律。人工智能算法作为“人工现象”在数字时代的典型代表，自然也不例外。

另一方面，以“科学”研究的方法探究“人工现象”，并不意味着仅仅依靠旨在解决特定功能性问题的自然科学视角即可完成，人造物被置于特定环境之中而与环境相互影响的复杂机制，往往使得对于“人工现象”规律的把握需要跳出自然科学的边界。以算法为例，尽管程序员往往是以达成特定功能性目标（例如

人脸识别准确率)作为算法设计目的,但算法在特定环境下的应用结果仍然并不完全取决于设计者(例如人脸识别针对不同人种的识别准确率可能存在系统性偏差)。针对人造物应用过程的环境复杂性,旨在讨论人或其他生物与所处环境相互影响关系的“行为科学”便具有了重要价值,其分析问题的框架、层次也自然可能作为解构人工现象基本规律的重要参考,这便成为了“机器行为学”的理论起点。换言之,“机器行为学”的核心主张是将机器所处环境纳入考量范畴,通过经验式的研究方法分析机器在特定环境下的表现“行为”,从而探究其内部机制、把握其运行规律、预测其可能结果。“机器行为学”更多体现了以结果“倒推”原因的逆向思维,而非通过探究机器设计过程而以原因“界定”结果的正向逻辑。机器应用于社会环境的复杂性,决定了“机器行为学”此种视角的重要性。

就具体的分析框架而言,沿袭行为科学对于人或动物行为规律的探索,机器行为学同样从“触发(Generation)、发展(Development)、功能(Function)、进化(Evolution)”四个议题出发,完成对于机器生命周期的分析与建构。

首先,“触发”问题关注特定机器行为产生的具体原因和机制,而这又往往涉及机器本身的设计原理及其所处环境的激发条件这两个方面。例如无人驾驶汽车为何会出现启动、变道、停车等特定驾驶行为便是“触发”问题的具体体现,而其原因便可归结为无人驾驶汽车的技术设计,以及操作者和运行环境所给予的特定输入。其次,“发展”问题关注机器形成某类特定特征的原因和机制,这既可能源于机器设计者的特定设计,也可能取决于机器运行所处环境的特定结构。例如人脸识别算法对于白人男性的识别正确率系统性的高于黑人女性,其原因在于该算法训练所用数据集中白人男性照片远远多于黑人女性,从而导致了算法偏差这一特定“特征”。第三,“功能”问题关注不同机器在应用于特定环境下出现特定选择结果的原因和机制。从行为科学的视角,特定功能的形成源于不同机器对于所处环境具有不同的适应度,与环境适应度高的“功能”将被保留和扩散,而不适应的“功能”则将消失。例如亚马逊电商平台上,卖家使用的动态调价算法主要围绕卖家的星级、好评比例、用户回复数量、与最低价的差值等因素展开,而并未考虑销售量、用户差评比例等其他因素,其原因就在于前者与亚马逊电商平台的广告排序算法更匹配。第四,“进化”问题关注机器功能的演化过程,其并非完全决定于当前环境,而同样受到历史因素和机制的影响。相比于生物进化

更多取决于基因突变和环境选择的缓慢过程，机器行为因不同机器之间的网络化连接和迭代化创新而可能呈现出更快的进化速度和更广的普及范围。与此同时，不同政治经济环境可能对机器进化施加不同约束，这也将导致差异化机器进化路径的形成。

“触发、发展、功能、进化”四个议题的探索构成了对于机器生命周期的整体描绘，而“机器个体、机器群体、人-机混合”则构成了分析机器行为的三个不同层次。第一，“机器个体”行为层次的研究聚焦于单个机器本身的内生属性或行为规律，其或者比较单个机器在不同环境下的行为差异，或者分析不同机器在相同环境下的行为变化。前者例如探究基于特定区域数据的犯罪风险评估模型在被应用于其他区域时是否发生变化，后者例如比较分析不同电商平台上的动态调价算法是否存在差异。第二，“机器群体”行为层次的研究聚焦机器群体作为一个整体所体现出的特征或规律，其或者关注机器之间的行为互动，或者关注机器群体的结构或属性。举例而言，维基百科的自动编辑机器人可能产生不同于人类的“互相斗争”行为特征，而社交媒体机器人也会产生“核心-边缘”结构的不同分工。第三，“人-机混合”行为层次的研究聚焦人类与机器共存环境下的行为特征与规律，其既包括机器对于人类行为的影响，也包括人类对于机器行为的塑造，还包括“人-机”协作过程所体现的特殊属性。新闻推送算法对于人类观念以及社会结构的影响、特定政治经济环境对于机器演化结果的选择、“人-机”共存情况下信息传播规律和特征的变化都是上述三点的典型案例。

由此，机器行为学从行为科学角度对人造物展开的分析视角，以及在此基础上从四个议题、三个层面提出的分析框架，构成了探索“人工现象”科学规律的新理论，而这也自然成为理解算法现象的新起点。换言之，机器行为学的本质在于探究人工智能算法作为一种“行为”的模式与特征，而对此的把握自然可以作为进一步展开人工智能算法影响评价的基础。

具体而言，机器行为学从触发、发展、功能、进化四个角度对行为模式的分类，可作为心理计量学建构隐变量的基础框架。例如“触发行为”关注的是引起算法输出变化的特定因素。从影响评价的角度来看，利益相关方关心的是某些特定因素“不应该”引起输出变化（例如性别因素引发的歧视结果），以及某些特

定因素“应该”引起输出变化（例如性别因素引发的分化需要被缩小）。在此意义上，“不应该被区分对待的行为结果”以及“应该被区分对待的行为结果”就作为两种机器行为模式被归纳出来，而其背后的一般/通用因素（即算法行为的隐变量）也因此“在心理计量学的范畴下需要被界定出来，进而以此为基础完成测度指标的构建。这一思路同样可被用于“发展、功能、进化”这三个议题，而由此形成的理论框架则可作为人工智能算法影响评价实践探索的理论指引。

4.2.2. 层次结构：技术、组织、社会

在 4.2.1 节的论述中，本报告已经体现，但却没有清晰论述的一个逻辑是算法影响评价的层次结构。即，心理计量学与机器行为学的核心目标在于梳理、界定人工智能算法行为的规律与特征，但这一工作并不是仅仅针对人工智能作为一种技术系统本身而展开，其行为规律与特征与技术、技术开发应用所处的组织、该组织所处的社会环境这三个逐渐递进的层次紧密相关。也正是在此意义上，我们对于人工智能算法行为规律与特征的勾勒，事实上应该包含技术、组织、社会这三个层面；相应的，人工智能算法影响评价的对象也需要同时包含技术、组织、社会这三者。

但接下来的问题在于，在算法影响评价的具体实践中，这三者究竟应该如何被涵盖其中，每一个层次的影响评价应该包含哪些内容，而这三个层次的相互关系是什么，这种相互关系又应该如何被纳入影响评价的体系之中？

对这些问题的回答需要从技术社会学、技术政治学等理论沿袭中寻找答案，而不同学者所提出的丰富理论模型便是吸取营养的土壤。就此而言，本报告选择了以下五种具有代表性的理论模型进行解释，以体现分析思路。但需要指出的是，这并不意味着算法影响评价对技术、组织、社会三者的考虑只有这五种模型，在不同环境下可被应用、发展的理论模型可能还有其他，这仍然是一个开放性问题的（参见 4.1.1 节的论述）。本节的主要目的仍然在于启发改革的思路，而非提供改革的方案。

（1）简·芳汀与技术执行框架

作为数字政府治理领域被广泛引用的理论源头，简·芳汀关于技术执行框架

的论述同样可被视为对“技术-社会”系统的经典表达。简·芳汀理论创新的起点是对技术决定论、理性行动者与渐进主义、系统分析等已有理论观点的批判，而这些理论都可被视为割裂“技术”与“社会”后的单方面表达。

技术决定论认为技术能够自动作用于个人、制度与社会，并因此带来或积极、或消极的影响，技术相对于社会而言是一个外生变量，而这一观点明显忽略了技术及其影响产生之间的复杂过程，特别是不能解释相同技术所带来的差异化影响。理性行动者与渐进主义都强调个人或组织基于理性分析而对技术的渐进接纳、使用，在此过程中个人或组织将会选择“最好的”技术，并不断矫正、修正以服务于其目的，但这一观点又完全忽视了社会行为的“非理性”因素，即我们往往是在路径依赖效应的影响下累计“错误”而不是“学习”经验，由此导致的技术更迭或重建成本极为昂贵。系统分析尽管承认了技术与社会相互影响过程的复杂性，但其强调“制度性滞后”（即制度变化滞后于技术创新）的隐藏假设则仍然是认为技术的新发展总是好的、总是应该且尽快被采纳，但人类认知的稳定性以及根深蒂固的制度并不一定能够很好的吸纳变化，而且流程再造工程往往失败率很高，在此意义上系统分析究竟在多大意义上接近现实也是存疑的。

在对上述观点展开反思的基础上，简·芳汀提出了技术执行的理论框架，其核心贡献是区别了“客观技术”与“被执行的技术”，并认为客观技术需要通过组织、制度的双重“过滤”后方能成为“被执行的技术”，而作为“被执行的技术”，其本身已经包含了足够分量的主观因素，并因此其执行结果可能是不确定的、多重的、理性或者非理性的。同样值得强调的是，技术执行框架并不认为制度是不可改变的，正是因为其对主观性的重视，才使得该理论框架的重要假设之一是建立在行动者视角基础上。综合来看，简·芳汀认为技术执行是认知、文化、结构、政治嵌入的结果：一方面组织行动者倾向于利用技术来维持现有社会或网络关系，但另一方面，技术作为一个媒介触发者同样也影响着新变化的发生，并推动行动者去改变结构与制度。

技术执行框架对于算法影响评估的启发意义在于，其提供了评估的基础框架和要素对象。一方面，客观技术（硬件或软件）、组织形式（等级结构、全县、标准化程度、稳定性、社会资本等）、制度安排（认知、文化、社会、法律等）

三者被视为塑造“被执行技术”的重要因素，并因此产生了具有影响作用的“结果”。在此视角下，客观技术、组织形式、制度安排自然应被视为评估的对象或被纳入评估范畴，因只有对这三者展开评估并基于评估产出才可能去影响“结果”；另一方面，客观技术、组织形式、制度安排并非单方面决定“被执行的技术”，后者同样可能触发行动者行为并因此改变组织形式或制度安排，由此导致技术与结构相互影响的结果呈现出不确定性。但即使如此，行动者行为仍然是可以把握、追踪并评估的对象。在此视角下，算法影响评估同样需要关注推动算法技术创新、应用的“行动者”，并在对其行为可能影响做出评估的过程中，干涉算法结果、积累治理经验。

技术执行框架的价值是不言而喻的，但其不足同样也是十分明显。“客观技术”与“被执行技术”的区分在多大程度上能够适用于对人工智能算法的分析，仍然是可讨论的。特别的，将组织形式、制度安排作为隔离“客观技术”与“被执行技术”边界的思路，仍然是将“技术”视为外在、确定对象然后经由组织形式、制度安排进行裁剪的传统思路，但人工智能算法从技术研发到应用的连续性过程可能会突破这一框架。在此意义上，我们可能仍然需要寻找一个更连续而非离散的“技术-社会”系统分析框架。

（2）“委托-代理”系统框架下的人与算法

在人工智能技术演化背景下考虑“技术-社会”系统的第二个视角是将人工智能算法视为人类生产生活活动的替代者，并因此与人类社会形成了“委托-代理”关系。同样根据“委托-代理”关系中信息不对称种类和强度的不同，人工智能算法作为决策系统的影响可被划分为不同类型，并因此被要求承担不同责任。⁸³在经典的“委托-代理”框架下，委托者依赖代理者去完成特定任务，因而其收益受到代理者行为的影响；而由于信息不对称的存在，代理者出于自身利益考量而可能掩盖其动机或行为，并因此损害委托者利益。由此，“委托-代理”框架在揭示二者关系的同时，致力于寻找解决信息不对称的责任机制。

⁸³ Krafft, T. D., Zweig, K. A., & König, P. D. (2022). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, 16(1), 119-136.

在人工智能算法创新和应用背景下，算法决策系统便可被视为人类“代理者”以完成各项任务，由此形成的“委托-代理”框架可被界定如下：考虑到人工智能算法决策的目的往往是为用户推荐广告、对用户评级等，因而此处的“委托者”应被视为算法决策系统的服务对象（并同时受到算法决策影响），与之相对的“代理者”应涵盖算法决策系统本身及其设计者、运营者。在该框架下，信息不对称现象主要体现在三个层面：“代理者”可能掩盖其“行为”（委托者可能需要花费很高成本才能看到算法决策结果），“代理者”可能掩盖所掌握的“知识”或“信息”（即使委托者能够观察到算法结果，但可能因欠缺足够信息并因此不能对该结果的影响展开完整评估），“代理者”可能掩盖“动机”或“偏好”（即是委托者能够观察到结果并了解信息，但不能确定算法系统的动机，例如算法设计中故意嵌入的有偏参数或者从数据中学习习得的偏见）。

以上三种类型的信息不对称现象都可能导致“代理者”与“委托者”价值的不对齐，而针对它们也相应存在差异化的责任机制。例如针对可观察到的“行为”偏差问题，基于市场的责任机制很可能就能充分有效，即通过鼓励市场竞争、用脚投票的方式，即能够约束算法系统作为“代理者”相对于“委托者”的偏差；对于“信息/知识”偏差问题，市场机制已经不够充分，对“代理者”提出透明度要求可能成为必须，即算法系统需要解释所采用的算法机制、算法设计与应用流程、可能存在的风险及预防措施等，以回应“委托者”的质疑；对于“目标/动机”偏差问题，即使是一般性的透明度要求也已不够充分，回溯性要求可能成为必需，此时不仅要对算法系统的基本原理进行解释，还需要全流程的评估、数据集监测等干涉性措施。

借鉴“委托-代理”框架不难看出，算法影响评估的目的事实上服务于不同类型的信息不对称现象，而其宗旨都在于确保算法系统作为“代理者”相对于用户作为“委托者”的价值与权益对齐。在此框架下，算法影响评估事实上明确了其作为算法治理特定环节的价值定位，并基于该价值定位而可能衍生出的相匹配的责任机制。但该框架的不足之处在于，其并未明确区分“算法技术系统”与“算法技术系统的开发者、应用者”这两类的责任区别，而只是宽泛地将二者都视为“代理者”加以讨论分析。在现实治理场景中，为避免抑制创新，我们需要将二

者区别对待。在此意义下，“委托-代理”框架仍然可能掩盖了一部分可能的制度创新空间。

（3）“社会-生态”系统理论（Social-Ecological Systems）

将技术、组织、社会纳入统一考量的第三个视角是在各个领域都有所发展的“社会-生态”系统理论，其具体又可能表现为不同形式，较有代表性的例如卢曼关于系统功能主义（systems functionalism）的理论建构，以及帕森斯的“结构-功能”主义（structural functionalism）。“社会-生态”系统理论的分析起点是试图从系统论视角对整个人类活动与生态环境的相互影响关系做出结构性解释，在划分不同子系统并厘清边界的基础上，抽象出内在机理与机制，从而为在行为、组织、制度层面的人为干涉提供参考。以帕森斯的“结构-功能”主义理论为例，在其理论演化过程中，帕森斯抽象出了文化（culture）、社会（规范）（social norms）、个性（personality）、环境（资源）（environmental resource）以及元系统（telic system）五个子系统，并认为它们构成了解释行动者行为结果的结构因素，而行动者行为结果的不同组合又最终构成了人类社会运行的各种“功能”。

尽管“社会-生态”系统理论能够极大简化人类社会与生态环境相互影响关系的复杂度，但其适用性仍然面临着诸多批评。例如以哈贝马斯为代表的批判理论学派即认为，“社会-生态”系统理论本身秉持的系统论视角存在“中立性”假设，并因此难以对权力、政治、价值等概念做出充分回应。类似的，“反自然主义”（antinaturalism）同样认为，“社会-生态”系统理论尽管提出了结构性解释，但其在很大程度上仍然不能挖掘深层次的行为因素，例如人类的创造力、想象力等非物质要素并没有被涵盖其中。

从探索算法影响评估制度建构的视角和目的来讲，围绕“社会-生态”系统理论正反两方面的讨论同样具有启发性。一方面，算法影响评估的本质是试图解释人工智能算法系统作为一种“人造物（artefacts）”而对社会环境、自然环境

带来的影响，并基于评估结果反过来对此“人造物”的设计与应用进行干涉或修正。就此而言，算法影响评估的制度建设逻辑与“社会-生态”系统的理论探索目标是一致的。也正因为此，“社会-生态”系统（例如帕森斯的“结构-功能”主义理论），同样有利于以结构化视角帮助算法影响评估建构起可解释、可操作、可量化的制度框架。但另一方面，考虑到“社会-生态”系统本身的局限性（及引致的批评），算法影响评估制度建设同样需要做出适应性调整。例如在算法影响评估中受到高度重视的价值对齐问题（涉及公平、歧视等价值诉求），是否以及如何被纳入系统论视角之下而得到充分讨论，仍然是需要探索的理论突破问题。在此意义上，“社会-生态”系统理论仍然只是提供了一个基础性起点，而非可直接复制使用的模版。

4.3. 结论与讨论

作为一种已经受到普遍认可和共识的算法治理机制，算法影响评价制度正陆续在各国逐步展开。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，提出了要求对生成式人工智能进行算法影响评价的制度要求。为进一步落实该文件，2024年2月，全国信息安全标准委员会发布了全国网络安全标准化技术委员会技术文件《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。更为重要的是，《安全基本要求》第一次对我国算法影响评价的对象、范畴等实质性内容做出了明确规定。尽管其是以技术文件的形式发布，不具有强制性，监管部门究竟将如何利用该文件尚存在不确定性，但如果不考虑短期的合规性问题，而是从算法影响评价制度体系的建构与执行角度来讲，《安全基本要求》往前迈出了实质性的一步，其可能极大加速算法影响评价制度的完善进程。

在此背景下，本研究围绕算法影响评价制度逻辑的梳理与分析，将具有重要意义。一方面，当我们意识到算法影响评价的制度形式可能存在多种选择的前

提下，明确其制度目标、厘清其制度逻辑将有助于决策者清楚辨析不同模式的优劣，并在此基础上选择、建构适合本国需要的算法影响评价制度；另一方面，算法作为一种新兴技术现象，以及算法应用作为一种新型“技术-社会”系统而体现出的特殊规律，使得传统的影响评价理论可能并不适用于此，本报告在第三、四章展开的理论探索，将同样有助于未来改革的进一步推进。当然，更具体的算法影响评价制度内涵、要求的分析已经超出了本报告的内容，而这也是未来进一步需要完善之处。当理论与实践形成相互促进的良性循环之后，算法影响评价才可能实现其制度初衷、提升其制度绩效。