

发展、权利、安全的三元博弈和抉择 ——中美欧人工智能治理比较与中国路径

摘要: NextX 系列 (7): 超级智能时代 (五)

- **美国强调创新优先, 区分民用与国家安全领域实施差异化监管; 欧盟以统一硬法和风险分级为基础, 重视公民基本权利保护。**美国形成白宫统筹、联邦部门执行、行业监管嵌入和跨部门协调相结合分散式治理体系。近年来, 治理重心由拜登政府时期的风险前置转向特朗普第二任期的创新竞争与国家安全导向; 民用领域依靠行业自律和事后执法降低合规负担, 对军事国防、前沿模型和网络安全等关键领域实施针对性管控, 并通过算力、能源、资本、人才等产业政策以及出口管制、盟友合作和全栈技术输出巩固技术优势。欧盟以《人工智能法案》为核心, 建立统一风险分级硬法治理框架, 针对通用人工智能增设专项合规义务。在强化基本权利保护的基础上, 配套监管沙盒、服务平台及合规简化措施平衡监管约束与产业创新, 同时依托布鲁塞尔效应实现治理规则外溢。
- **中国人工智能治理构筑起以国家战略统筹、敏捷安全监管为特征, 兼顾产业应用落地与全球治理参与的复合型治理体系。**我国依托党中央集中统筹、多部门协同机制, 以“人工智能+”行动推动技术向制造、医疗、能源、消费等领域加速落地, 围绕算力、数据、核心技术、资本、人才和开源生态完善产业发展基础。风险治理遵循敏捷迭代思路, 监管对象由算法推荐、深度合成拓展至生成式人工智能, 并正面向智能体和具身智能延伸。建立风险分级管理机制, 综合运用算法备案、伦理审查、行业准入等事前管理工具, 叠加国家安全审查、专项整治和事故追责形成事中事后闭环。对外秉持“以人为本、智能向善”理念, 积极参与全球人工智能治理, 构建包容普惠的国际治理秩序。
- **受治理目标、产业基础和制度环境影响, 中美欧三方在治理逻辑、监管工具、产业扶持和国际博弈的选择上存在明显分化, 同时在风险共治和标准协调上仍存在合作空间。**美国优先维护技术和产业竞争优势, 在民用领域偏向软法和事后监管, 在国家安全关键领域强化约束; 欧盟以基本权利保护为核心, 通过统一硬法和风险分级建立强制性合规体系; 中国兼顾发展与安全, 以基础法律、专项规章和国家标准构建多层次治理框架。产业支持路径上, 美国充分激活市场与资本力量, 欧盟依托公共资源补齐产业短板, 中国发挥战略统筹优势系统配置创新要素。在国际治理方面, 美国输出技术生态、欧盟输出规则体系、中国推广多边治理倡议, 而全球治理也面临规则碎片化、企业跨境合规成本抬升等现实挑战, 未来人工智能竞争将是技术、规则和产业生态相互协调的综合博弈。
- **中国可立足于国内产业迭代和企业出海的现实诉求, 优化治理工具、补齐新型风险防控短板, 并深度参与全球人工智能治理。**国内治理层面, 可借鉴域外风险管理的成熟实践, 将原则性治理要求转化为可量化、可落地的风险评估和测试流程, 扩大监管沙盒试点, 打通第三方测试、行业自律与政府监管的衔接机制。风险防控层面, 前瞻研判智能体、具身智能等新技术形态的潜在风险, 深化场景差异化监管, 完善模型全生命周期治理, 强化模型迭代后的持续监测、预警与应急处置能力, 夯实算力、数据集等产业创新要素。全球治理与企业出海层面, 推动国内治理规则与美欧国际标准映射, 探索合规等效互认。搭建面向海外市场的公共合规服务体系, 实现产品、测试工具和治理标准协同出海, 依托多边平台开展能力建设合作, 提升中国人工智能治理实践国际影响力。
- **风险提示:** 人工智能产业增长不及预期、人工智能治理规则分化加剧、人工智能技术风险演变超预期、中国人工智能治理政策落地不及预期。

产业研究中心



朱峰(分析师)

021-38676284



zhufeng@gtht.com



登记编号

S0880522030002



汪玥(分析师)

021-38031030



wangyue8@gtht.com



登记编号

S0880525080001

往期回顾

中美建设性战略稳定关系中的能源产业布局与影响 2026.09.30

商业航天中美对比: 美国全链引领, 中国加速崛起 --商业航天系

列五 2026.09.28

6G: 迈向万物智联, 重构智能经济网络底座

2026.09.26

中美对比研究: 中国主导硬件供应链, 美国引领具身模型创新——人形机器人系列 (二)

2026.09.22

并购启新程, 产业开新局——从中美周期对比、全球龙头实践看中国并购市场机遇 2026.09.22

目录

1. 概述	6
1.1. 美欧人工智能治理体系总体特征	6
1.2. 中国人工智能治理体系总体特征	7
1.3. 中美欧人工智能治理模式对比分析	8
1.4. 中国人工智能治理政策启示	8
2. 美国人工智能治理体系	9
2.1. 整体统筹：多部门协同形成多层次的人工智能治理架构	9
2.1.1. 联邦内部：白宫统筹与跨部门协同机制	9
2.1.2. 联邦与州级：联邦优先权与州级碎片化立法之间的博弈	11
2.2. 两任政府治理模式演变：从风险治理转向竞争驱动	11
2.2.1. 拜登政府：强调人工智能安全监管，推动风险前置与制度化管理	11
2.2.2. 特朗普第二任期：重视创新驱动发展，去监管化和国家安全导向间的回摆	13
2.3. 民用领域：采取自愿标准和行业自律，形成轻合规治理模式	14
2.3.1 NIST 迭代自愿性标准框架，为企业提供可裁剪的风险管理流程	15
2.3.2 行业传统监管部门嵌入人工智能治理，以事后执法为主	16
2.3.3 鼓励企业自愿承诺，通过市场自律形成公私合作机制	16
2.4. 关键领域：军事与前沿模型领域实施严格监管和前置审查	17
2.4.1. 军事国防：人工智能战略地位升级，强制安全标准纳入采购流程	17
2.4.2. 前沿模型：治理规则持续迭代，采取风险评估、前置审查与安全管控	18
2.4.3. 网络安全：推动人工智能融入网络防御体系，强化关键基础设施安全治理	19
2.5. 注重创新：多措并举扶持人工智能产业技术领先优势	21
2.5.1. 完善算力基建与能源配套，夯实人工智能发展的技术底座	21
2.5.2. 加大联邦资金投入，引导社会资本协同助力产业创新	23
2.5.3. 重视专业人才教育，构建覆盖全民的培养体系	24
2.6. 对外博弈：持续升级出口管制，全栈技术输出以巩固盟友体系	25
2.6.1. 出口管制：从芯片禁运到模型权重限制、由广泛限制转向精准管制	25
2.6.2. 规则与技术输出：通过治理规则输出和全栈技术出口强化盟友绑定	26

3. 欧洲人工智能治理体系.....	28
3.1. 顶层治理模式：由欧盟专门机构统筹部署，成员国具体执行落地.....	28
3.1.1. 欧盟统筹：依托人工智能办公室和人工智能委员会强化统一治理.....	28
3.1.2. 成员国执行：建立国家监管机构以落实市场监管和企业合规.....	29
3.2. 监管思路：《人工智能法案》建立以风险分级为核心的统一强制性监管约束.....	30
3.2.1. 《人工智能法案》的通过与实施.....	30
3.2.2. 风险分级监管：形成从禁止不可接受风险到低风险豁免的差异化监管.....	31
3.2.3. 监管对象：从高风险系统向通用人工智能拓展.....	32
3.3. 配套措施：推动监管要求落地，强化基本权利保护.....	32
3.3.1. 通过辅助合规措施与自愿倡议，降低统一监管框架的落地成本.....	33
3.3.2. 研制并采纳协调标准，构建技术合规基准.....	33
3.3.3. 强调保护基本权利，将隐私、公平和非歧视纳入人工智能治理.....	34
3.4. 监管与创新并重：通过监管沙盒和产业政策缓解高合规要求对创新的约束.....	36
3.4.1. 设置监管沙盒与测试设施，以可控试验降低企业创新的合规成本.....	36
3.4.2. 产业政策导向从监管优先，转向兼顾强化技术主权和产业竞争力.....	37
3.4.3. 支撑算力、数据、科研和人才发展，强化欧盟技术主权发展的产业基础.....	38
3.5. 国际影响：通过规则外溢和价值引领扩大欧盟治理话语权.....	39
3.5.1. 借助统一市场和规则输出，实现人工智能监管标准和价值观念外溢.....	39
3.5.2. 依托多边机制和数字外交，拓展人工智能的技术竞争力和全球影响力.....	40
4. 中国人工智能治理体系.....	42
4.1. 整体统筹：政府主导、多部门协同推进人工智能治理体系建设.....	42
4.1.1. 强化顶层战略引导，推动人工智能发展方向的统一.....	42
4.1.2. 多部门协同治理，共同推动人工智能的政策协调与落地实施.....	43
4.2. 创新要素：围绕算力、数据、技术、开源生态、资本和人才完善人工智能发展基础.....	45
4.2.1. 算力基建：推进智能计算基础设施布局，提升人工智能发展支撑能力.....	45
4.2.2. 数据资源：推动数据资源开发利用，完善人工智能训练数据供给.....	47
4.2.3. 核心技术：加强关键技术攻关，提升人工智能产业自主创新能力.....	48

4.2.4. 开源生态：推动开源开放合作，促进人工智能技术扩散应用	49
4.2.5. 资本投入：发挥政府与市场作用，引导人工智能产业投资布局 ...	50
4.2.6. 人才培养：加强人工智能人才供给，支撑产业长期发展	50
4.3. 应用落地：推动人工智能与实体经济深度融合	51
4.3.1. 政策部署：实施“人工智能 +”行动，促进重点行业的智能化转型	51
4.3.2. 行业应用：推动人工智能在重点领域的加速落地	51
4.4. 风险监管：构建覆盖人工智能全生命周期的动态治理机制	54
4.4.1. 监管对象持续演进，从算法管理逐步拓展至生成式人工智能治理	54
4.4.2. 通过备案制提升治理精准度，实现高风险模型的持续跟踪	55
4.4.3. 强化全链条内容标识，通过强制溯源提升人工智能应用的透明度	55
4.4.4. 全面推进科技伦理审查，落实研发周期前置防控措施	56
4.5. 安全保障：强化数据与关键领域管控，筑牢人工智能发展边界	57
4.5.1. 建立人工智能风险分类管理机制和技术安全基准	57
4.5.2. 以基础法律强化数据安全治理和个人信息保护	58
4.5.3. 构筑国家安全审查体系，将人工智能纳入安全部署	59
4.5.4. 完善重点领域的安全准入机制	60
4.6. 国际参与：提出人工智能治理的中国方案	61
4.6.1. 积极发布国际倡议，提出人工智能治理理念和合作方案	61
4.6.2. 倡导多边合作，依托国际机制拓展人工智能治理网络	62
5. 中美欧人工智能治理体系比较分析	64
5.1. 治理逻辑：从发展目标出发形成差异化的治理模式	64
5.1.1. 美国：以创新驱动发展的治理模式	64
5.1.2. 欧盟：以规则约束和统一立法为核心的治理模式	65
5.1.3. 中国：形成以统筹发展为核心的治理模式	66
5.2. 治理体系：治理主体与制度工具的不同特征	66
5.2.1. 治理主体：分散协同、超国家协调与集中统筹并存	66
5.2.2. 制度工具：硬法监管、软法治理与复合平衡之间的不同选择	67
5.3. 风险治理：风险关注点从技术安全拓展至社会安全	68
5.3.1. 风险治理范围：从技术安全向国家、社会与个人安全扩展	68
5.3.2. 风险治理工具：覆盖事前介入、过程监管、事后处置的全链条差异化治理	69
5.4. 产业发展：通过完善创新要素配置建设人工智能产业生态	70
5.5. 全球人工智能治理趋势	71
5.5.1. 国际竞争：形成技术优势、规则输出和产业生态建设的综合竞争模式	71

5.5.2. 全球治理存在规则碎片化特征，增加跨境协调成本	72
5.5.3. 治理模式由静态规则制定向敏捷动态监管转变，以适应技术的快速迭代.....	72
5.5.4. 人工智能治理将持续关注产业竞争与社会影响的平衡	73
6. 治理启示：国际经验与国内展望.....	73
6.1. 借鉴国际经验：完善人工智能治理工具体系	73
6.1.1. 建立监管沙盒机制，为人工智能监管规则提供试错空间	74
6.1.2. 将原则性风险治理要求转化为可落地的风险管理流程	74
6.1.3. 建立健全人工智能标准化测试评价体系	75
6.1.4. 采用渐进式立法和敏捷监管的治理逻辑	75
6.1.5. 拓展企业自律机制，推进与政府监管的衔接	76
6.2. 展望国内治理：立足中国发展实际，补齐新型风险与治理能力短板 ..	76
6.2.1. 前瞻识别新型风险，向智能体和具身智能等自主行为安全延伸 ...	77
6.2.2. 厘清人工智能责任链条，明确开发者、部署者和使用者责任分担机制.....	77
6.2.3. 聚焦高风险应用场景，进一步深化行业和场景差异化治理	77
6.2.4. 事前监管基础上，同步关注持续监测和事故追责等事中、事后监管措施.....	78
6.2.5. 统筹算力与数据资源，筑牢人工智能发展的基础设施根基	78
6.3. 展望全球治理：推动治理规则国际化，构建企业出海支撑体系	79
6.3.1. 推动国内治理标准向国际标准转化.....	79
6.3.2. 建立人工智能企业出海合规支撑体系	80
6.3.3. 推动技术、产品与治理标准配套协同出海	80
6.3.4. 通过标准互认和多边合作降低规则碎片化	81
7. 风险提示	81
7.1. 人工智能产业增长不及预期.....	81
7.2. 人工智能治理规则分化加剧.....	81
7.3. 人工智能技术风险演变超预期.....	81
7.4. 中国人工智能治理政策落地不及预期.....	81

1. 概述

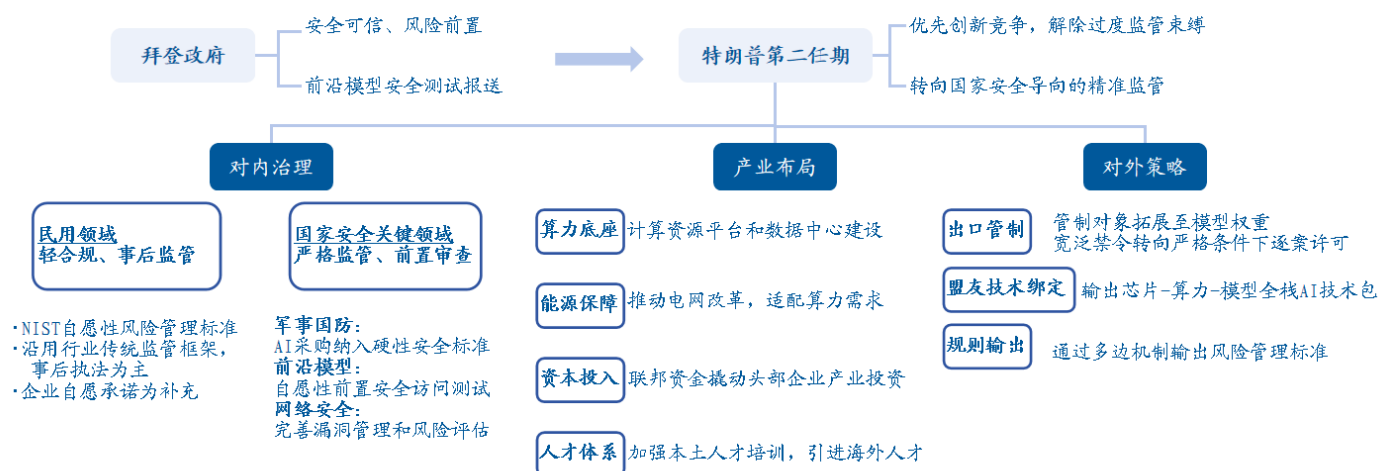
全球人工智能技术正在经历快速迭代并向社会各领域渗透，治理体系从早期围绕模型安全、数据保护等技术风险的单一监管，逐步开始兼顾产业发展、国家安全和治理等更为广泛的领域。美国、欧盟和中国基于自身技术基础、制度环境和治理理念，形成了差异化的人工智能治理体系：美国强调创新竞争和技术优势，欧盟侧重规则约束和基本权利保护，中国则更加注重统筹发展与安全。三方虽然在治理模式上存在明显差异，但均面临如何在技术快速发展的背景下兼顾创新动力与风险管控、降低跨境治理摩擦等现实挑战。通过对比域外治理经验，能够为我国人工智能治理优化提供参考。

1.1. 美欧人工智能治理体系总体特征

美国人工智能治理总体呈现出创新优先、分散式治理的特征，治理重心由风险防范转向技术竞争和国家安全导向。美国未设立统一的人工智能监管机构，而是依托白宫统筹、联邦部门执行、行业监管机构嵌入以及跨部门协调形成分散式治理体系。在推行联邦优先的基础上，由各州进行立法补充。拜登政府在治理中强调风险前置，对前沿大模型设置安全测试报送、政府采购规则等制度安排，重点防范算法歧视、网络安全等安全风险。特朗普第二任期政策出现明显转向，优先保障产业创新竞争力，撤销拜登时期行政令框架，通过行业自律和事后执法等方式降低民用领域合规负担。但是，在国家安全相关领域转向较为严格的精准监管，对军事国防、前沿模型和网络安全领域保留风险管控手段。

美国对内依靠多元政策工具完善人工智能发展所需的创新要素，对外采取技术出口管制和盟友规则输出，试图依靠技术优势带动产业生态和国际影响力。在产业端，美国持续推进算力基建、能源配套、资本投入和人才体系建设，通过国家科研算力平台、电网配套改革、公私资本协同、人才培育和引进巩固技术发展基础。在对外层面，持续更新出口管制规则，管制对象从人工智能芯片硬件延伸至前沿大模型权重、相关海外投资活动。还依托 G7、TTC 等多边平台向盟友输出美国治理规则，面向伙伴国家输出芯片-算力-模型全栈人工智能技术包，构建具有排他性的技术联盟。

图1：美国人工智能治理体系框架图



数据来源：美国白宫、联邦公报、美国 NIST、美国国防部、美国能源部、美国劳工部等，国泰海通证券研究

欧盟在统一硬法《人工智能法案》的基础上，采取以风险分级为框架、保护基本权利为特征的强制性监管模式。由欧盟人工智能办公室负责通用人工智能模型合规监管与协调，成员国设立市场监督、通知机构负责本土监管实施。按照风险等级将人工智能划分为不可接受风险、高风险、有限风险和最小风险，匹配差异化的合规义务，还针对通用人工智能（GPAI）设置专项强化合规要求。欧盟还配套建立协调标准、监管沙盒、测试设施、AI 工厂和数据实验室等支持创新的工具，并推出一系列政策措施和数字化服务平台，以简化监管合规流程。同时，以“布鲁塞尔效应”发挥法案域外效力，借助多边外交输出人权导向的治理规则。

图2：欧盟人工智能治理体系框架图



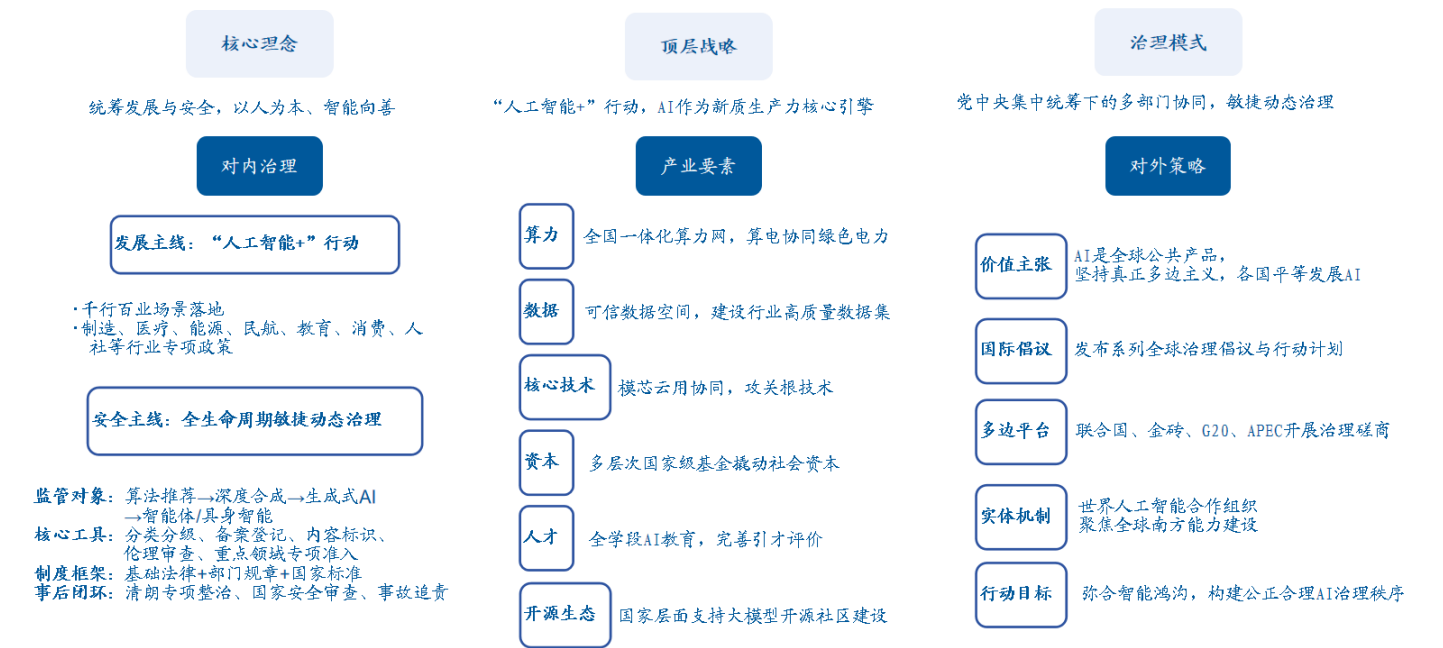
数据来源：欧盟委员会官网、EUR-Lex 欧盟法律数据库、欧盟委员会数字战略司、欧洲议会官网等，国泰海通证券研究

1.2. 中国人工智能治理体系总体特征

中国人工智能治理体系以国家战略统筹为引领，以“人工智能+”行动推进产业应用落地。党中央对人工智能治理进行集中统一领导，由国务院、核心主管部门和行业监管部门协调分工执行，并由标准化组织和技术支撑机构提供支持。从《新一代人工智能发展规划》到《“十四五”规划纲要》以及《关于深入实施“人工智能+”行动的意见》，人工智能的战略定位持续提升，并逐步从技术研发向制造、医疗、能源、消费、教育、交通、人社等具体应用场景延伸。在产业发展方面，中国围绕算力、数据、核心技术、资本、人才和开源生态进行系统布局，通过算力基础设施、高质量数据集和国家级资金推动人工智能技术发展。

中国治理体系伴随技术发展形成敏捷治理路径，在发展与安全之间寻求动态平衡，并对外构建公正合理的全球人工智能治理秩序。我国监管对象已经从算法推荐、深度合成扩展至生成式人工智能，进一步向智能体、具身智能等新型技术形态延伸。初步形成风险分级管理制度，为高风险领域强化监管、低风险领域保留创新空间提供制度基础。在安全治理上，建立算法备案、内容标识、伦理审查、重点行业准入等事前审查工具，辅以国家安全审查、专项整治行动和事故追责形成事后闭环，形成基础法律、专项规章和国家标准相结合的多层次治理框架。同时，中国发布全球治理倡议与行动计划，通过多边机制积极参与全球人工智能治理，强调“以人为本、智能向善”的理念，主张以真正多边主义保障各国平等发展人工智能的权利，将发展中国家纳入治理体系。

图3：中国人工智能治理体系框架图



数据来源：中国政府网、中央网信办、工业和信息化部、国家发改委官网、外交部官网、全国网安标委 TC260 等，国泰海通证券研究

1.3. 中美欧人工智能治理模式对比分析

中美欧人工智能治理均兼顾技术发展与风险防控，但受治理目标、产业结构和技术基础等因素影响，在治理逻辑和监管工具选择上存在差异。美国优先保障技术与产业竞争优势，以创新驱动为主，在民用场景以软法为主，关键领域实行刚性约束；欧盟强调保护公民基本权利，依靠统一立法确立规则，实现硬法强制约束；中国兼顾发展与安全双重目标，以基础法律、专项规章和标准构成多层次制度体系。在风险治理方面，美国更关注国家安全外溢风险，民用场景以事后执法和企业自律为主，国家安全相关领域实施刚性管控；欧盟突出个人基本权利和社会秩序，围绕高风险 AI 构建覆盖全链条的合规义务；中国则覆盖技术、社会 and 认知等多维风险，强调事前准入环节治理，兼顾事中、事后形成动态闭环治理。

除风险治理外，产业竞争和国际规则博弈成为三方的共同议题，技术、规则和产业生态日益绑定。产业发展方面，美国重视资本和企业力量，欧盟通过公共资金和基础设施投资补足产业短板，中国由国家战略统筹推动创新要素系统布局。美国依托技术和产业链优势，推动规则标准和技术体系向外扩散；欧盟依靠统一市场规模和《人工智能法案》形成规则外溢；中国则通过发布国际倡议和多边合作机制参与国际治理。不同治理模式和监管要求带来规则碎片化，增加了企业跨境部署的合规成本。中美欧在技术标准层面存在有限协调空间，但治理框架层面的实质性合作目前仍然有限。未来，人工智能治理将成为技术能力、治理规则与产业生态协同的综合竞争，随着人工智能技术的持续发展，各国也需要更加关注治理规则的敏捷动态调整，以及产业发展与社会价值之间的平衡。

1.4. 中国人工智能治理政策启示

在国内治理工具层面，可考虑在借鉴国际经验与盘活国内实践两个方向上同步推进。一方面，可参考美欧等主要经济体以规则、标准和技术生态扩大影响力的有益经验，但不简单复制其治理模式，将原则性的治理要求转化为可量化、可执行的风险评估和测试流程，针对不同行业形成风险评估模板。另一方面，可结合我国金融科技已有实践，在各领域探索人工智能监管沙盒，以真实应用验证数据使用、责任划分和安全要求，并通过行业自律、第三方测试与政府监管衔接，降低一般场景的制度性合规成本。

面向人工智能自主化和应用范围扩大趋势，应补齐新型风险识别和全生命周期治理能力，将监管重点延伸至自主行为及其现实后果。未来风险可能更多产生于人工智能自主决策、任务执行以及与现实环境交互的过程中，因此需要前瞻识别自主行为安全、产业链责任划分等新型问题，并根据金融、医疗、就业、制造、交通等场景设置差异化监管要求。同时，应强化模型更新后的持续评估、运行监测、风险预警、事故上报和应急处置，推动事前准入、事中监测与事后问责形成更加完整的闭环，并在算力和数据领域完善全国一体化配置、算电协同、高质量数据集和数据质量评价等基础支撑。

在全球治理和企业出海层面，可考虑推动国内治理标准与国际规则双向对接、互认互通，由单纯支持产品出海转向产品出海与治理、合规能力建设并重。具体来看，可推动国内治理规则与美欧治理规则及国际标准进行映射和对照，探索合规等效认定和标准互认，降低企业在不同市场出海的重复合规认证成本。同时，应建设面向海外重点市场的人工智能合规公共服务体系，提供合规检查清单、法律咨询、前置评估和应急响应等服务。在此基础上，推动人工智能产品、产业服务与测试工具、治理标准协同出海，并通过多边平台开展治理规则和能力建设合作，在扩大中国治理实践国际影响力的同时，为企业进入不同市场提供稳定的制度支撑。

2. 美国人工智能治理体系

2.1. 整体统筹：多部门协同形成多层级的人工智能治理架构

2.1.1. 联邦内部：白宫统筹与跨部门协同机制

美国未建立统一的人工智能监管机构，而是依托白宫战略统筹、联邦部门执行、行业监管机构嵌入以及跨部门协调机制，形成分散式的多层级治理体系。该模式能够充分发挥美国的行政体系和产业生态优势，根据不同应用场景快速调配政策资源，避免单一监管机构可能带来的僵化约束。但也会导致各治理机构监管权责边界模糊、跨部门政策协调成本较高和风险监控碎片化等潜在问题。

白宫体系是美国人工智能治理的决策中枢，主要负责制定国家人工智能治理的战略方向并协调联邦政府部门间的科技资源。美国人工智能治理不存在单一的监管机构，而是主要由总统通过行政命令、人事任命和预算提案在政治领导和方向设定上发挥核心作用，同时白宫科学技术政策办公室（OSTP）、国家科学技术委员会（NSTC）以及总统科学技术顾问委员会（PCAST）共同承担顶层设计职能。其中，OSTP 负责向总统提供人工智能等关键科技领域的政策建议，监督国家科技政策议程的制定和执行，是美国人工智能政策制定的重要牵头部门；NSTC 是美国行政管理部门内部统筹协调科技政策的核心机构，下设有人工智能专项委员会，推动制定跨部门的研发战略；PCAST 是由联邦政府外部杰出人士组成的独立科学顾问机构，为总统提供人工智能技术趋势、风险评估和政策建议，2025 年特朗普政府通过行政令重组 PCAST，规定总统科技助理（APST）和“人工智能与加密货币特别顾问”担任联合主席。此外，国家人工智能咨询委员会（NAIAC）也是美国的政策咨询机制之一，由学术专家、科技企业负责人及社会组织代表组成，负责向总统和国家人工智能倡议办公室提供人工智能研发、安全和经济竞争力等方面的独立评估和建议。

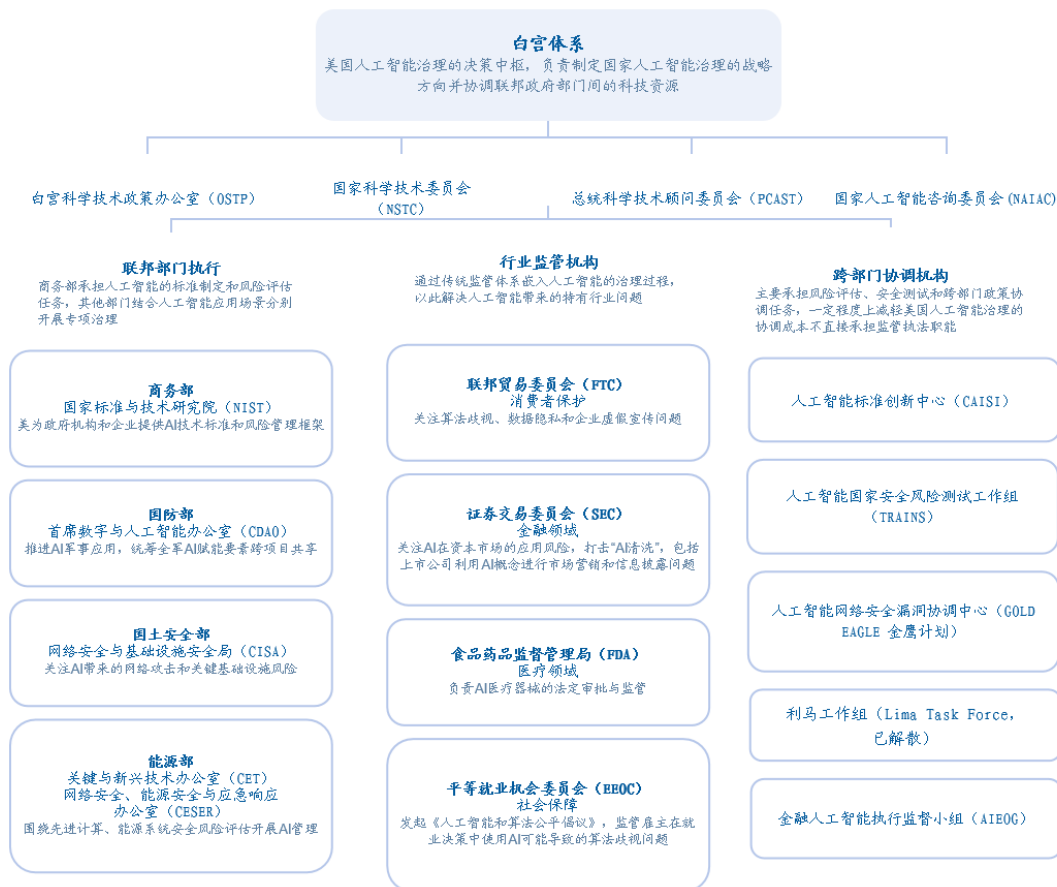
商务部及下属机构承担人工智能的标准制定和风险评估任务，国防、国土安全和能源等部门结合人工智能应用场景分别开展专项治理。商务部是美国人工智能产业治理的重要部门，其下属的国家标准与技术研究院（NIST）于 2023 年发布了自愿性人工智能合规框架《人工智能风险管理框架》（AI RMF 1.0），为政府机构和企业提供人工智能技术标准和风险管理框架。国防部通过首席数字与人工智能办公室（CDAO）推进人工智能的军事应用，统筹全军人工智能赋能要素跨项目共

享；国土安全部通过网络安全与基础设施安全局（CISA）关注人工智能带来的网络攻击和关键基础设施风险；能源部则通过关键与新兴技术办公室（CET）和网络安全、能源安全与应急响应办公室（CESER），围绕先进计算、能源系统安全风险评估开展人工智能管理。

行业监管机构通过传统监管体系嵌入人工智能的治理过程，以此解决人工智能带来的特有行业问题，但是不同行业监管机构之间的协调治理相对欠缺。在消费者保护领域，联邦贸易委员会（FTC）主要关注算法歧视、数据隐私和企业虚假宣传问题；在金融领域，证券交易委员会（SEC）关注人工智能在资本市场中的应用风险，将“AI 清洗”列为执法重点，包括上市公司利用人工智能概念进行市场营销和信息披露的问题；在医疗领域，食品药品监督管理局（FDA）负责人工智能医疗器械的法定审批与监管；在社会保障领域，平等就业机会委员会（EEOC）发起《人工智能和算法公平倡议》，监管雇主在就业决策中使用人工智能可能导致的算法歧视问题。

跨部门专项机构一定程度上减轻美国人工智能治理的协调成本，但不具备监管执法权限，在治理体系中发挥的作用依旧有限。美国商务部于 2025 年 6 月将原人工智能安全研究所（AISI）重组为人工智能标准创新中心（CAISI），设在 NIST 内部；此前，AISI 已于 2024 年 11 月牵头联合国防部、国土安全部等建立了人工智能国家安全风险测试工作组（TRAINS）；由财政部牵头，国土安全部、战争部参与建立了人工智能网络安全漏洞协调中心；在美国财政部召集下，金融服务行业协调委员会（FSSCC）、金融与银行信息基础设施委员会（FBIIC）合作建立金融人工智能执行监督小组（AIEOG）。这些机构主要承担风险评估、安全测试和跨部门政策协调任务，不直接承担监管执法职能。

图4：美国人工智能治理架构示意图



数据来源：美国白宫官网、联邦公报、白宫归档网站、CAISP 智库、美国 NIST、美国国防部、美国能源部 CET 办公室、美国 FDA 官网等，国泰海通证券研究

2.1.2. 联邦与州级：联邦优先权与州级碎片化立法之间的博弈

除白宫统筹与联邦政府内部的跨部门协同之外，联邦政府与各州之间的监管博弈也是美国人工智能治理体系的一大特征。在国家层面的联邦统一立法缺位的背景下，美国各州立足本地产业特征和价值观念，相继出台较为严格的地方性人工智能立法，在前沿模型监管、消费者保护等领域填补监管空白。

具体而言，科罗拉多州于2024年5月17日签署SB 24-205《人工智能消费者保护法》，聚焦算法歧视治理，对用于就业、住房、信贷等重大决策的高风险人工智能系统施加系统性义务，包括部署前影响评估、消费者事前通知、不利决策的人工复核权。加利福尼亚州于2025年9月29日签署SB 53《前沿人工智能透明度法案》，要求训练计算量超过 10^{26} FLOPs的大型人工智能开发者公开发布灾难性风险管理框架，并在可能导致大规模伤亡或重大财产损失的安全事件时向州政府报告。伊利诺伊州于2026年7月6日签署SB 315《人工智能安全措施法案》，要求先进人工智能系统的开发人员公开披露安全管理实践，报告重大安全事件，并建立完备的合规管理流程，成为较早要求前沿人工智能系统开展常态化独立第三方安全审计的州。

在多数州已自行制定人工智能相关规则的背景下，特朗普政府以行政令和立法建议形式推动联邦建立统一的监管框架，试图通过联邦优先原则取代州级的分散立法和过度监管。该治理原则可以为企业发展人工智能保留创新空间，降低企业跨区域经营的合规负担，避免地方碎片化监管对技术扩散造成的影响。但这也可能削弱州政府对地方特有风险采取的针对性保护措施，降低监管体系在早期防范和化解新型风险、为全国治理提供合理参考的可能性。

- 2025年12月11日，第14365号行政令《确保人工智能国家政策框架》宣布建立司法部人工智能诉讼专项小组，负责在联邦法院挑战与联邦政策冲突的州人工智能法律。要求商务部长在90天内发布全美现有州级人工智能法规评估报告，识别合规负担过重的法规，并威胁对出台严苛人工智能法规的州削减或取消“宽带公平接入与部署”（BEAD）专项补贴。
- 2026年3月20日，白宫发布的《国家人工智能立法框架》系统提出立法建议，核心内容包括建立统一的联邦人工智能监管标准，确立联邦法优先于州级法规的原则，但同时保留各州在未成年人安全保护、防止欺诈和消费者保护、人工智能算力与数据中心基建、州政府自身采购使用人工智能等领域的合法监管权限。

2.2. 两任政府治理模式演变：从风险治理转向竞争驱动

近三年来，美国人工智能治理政策经历了由拜登政府强调安全监管，到特朗普第二任期重新强化创新竞争和战略优势的政策调整。拜登政府提出完善风险评估和建设安全标准，试图在事前介入人工智能发展过程，以降低技术扩散带来的社会风险；特朗普政府则更加重视技术领先和产业竞争，通过放松监管限制以释放企业的创新活力。两届政府的治理路径均以维护美国人工智能全球竞争优势为终极目标，但政策重心的摇摆可能使得企业面临政策预期及合规方向不确定的难题。

2.2.1. 拜登政府：强调人工智能安全监管，推动风险前置与制度化管理

拜登政府时期，美国人工智能治理重点逐步转向建立安全可信的监管体系，通过行政命令、风险管理框架和政府采购规则强调人工智能治理的风险前置和合规管理。随着人工智能技术的快速突破，美国政府开始关注人工智能可能带来的潜在风险。为降低人工智能技术应用的不确定性，拜登政府提前建立风险管理机制，逐步形成以“安全、可靠、可信任”为核心的治理理念，提高政府对人工智能潜

在风险的识别和干预能力。

- 2022 年 10 月 4 日，美国白宫科学技术政策办公室发布《人工智能权利法案蓝图：让自动化系统为美国人民服务》（Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People），围绕系统安全有效、算法歧视保护、数据隐私、通知与解释、人工替代与退出五个方面提出指导原则，为后续人工智能立法和监管提供参考框架。该文件不具有强制法律效力，主要通过自愿性、原则性指导和政策倡议来推动保护公民权利和隐私的人工智能系统建设。
- 2023 年 10 月 30 日，拜登签署第 14110 号行政令《关于安全、可靠和可信任地开发和使用人工智能的行政命令》，是拜登政府对于人工智能的首部综合性行政文件，覆盖了人工智能的研发、部署和应用全过程。该行政令从以下方面提出政策要求：

表1：《关于安全、可靠和可信任地开发和使用人工智能的行政命令》具体要求概览

政策文件	具体要求
《关于安全、可靠和可信任地开发和 使用人工智能的行政命令》	1) 加强人工智能安全评估，要求开发前沿人工智能模型的企业向政府提交相关安全测试结果及其他关键信息；
	2) 规范联邦机构采购和使用人工智能，各联邦机构设立“首席人工智能官”（CAIO）；
	3) 推动人工智能技术标准建设，设立美国人工智能安全研究所（AISI）；
	4) 建立先进的网络安全计划，开发人工智能工具查找和修补关键软件漏洞；
	5) 增加人工智能研究投资，扩大人工智能人才引进；
	6) 防止人工智能在住房、就业、福利方面的歧视，推动消费者权益保护和隐私保护，评估人工智能对劳动力市场的影响；

数据来源：美国联邦公报，国泰海通证券研究

- 2024 年 3 月 28 日，美国管理和预算办公室（OMB）发布 M-24-10 号备忘录《推进机构使用人工智能的治理、创新和风险管理》，是落实第 14110 号行政令的核心配套文件。要求各联邦机构在 60 天内指定首席人工智能官，在 180 天内提交并公开发布合规计划，同时建立人工智能治理委员会。对“影响安全”（Safety-Impacting）和“影响权利”（Rights-Impacting）的人工智能应用设定了最低风险管理实践，未达标机构需停止使用不合规人工智能系统。
- 2024 年 9 月 24 日，美国管理和预算办公室发布 M-24-18 号备忘录《推进政府负责任采购人工智能》，围绕跨职能协作、人工智能风险与绩效管理、竞争性人工智能市场三大目标展开。要求对“影响权利”和“影响安全”的人工智能采购设定强制性合规要求，包括隐私保护、偏见评估、持续测试和 72 小时事故报告；针对生成式人工智能和 AI 生物识别系统设立的专门采购规范，要求供应商披露训练数据来源，并通过防止供应商锁定等措施培育竞争性人工智能市场。

2.2.2. 特朗普第二任期：重视创新驱动发展，去监管化和国家安全导向间的回摆

特朗普第二任期将人工智能视为影响国家竞争力的重要战略资源，治理重点从风险防范转向技术创新，通过大幅放松监管约束和简化合规要求，为企业创新释放制度空间。同时，政府助力基础设施建设和技术应用，提升美国人工智能的产业优势。这种治理模式有助于激发企业研发积极性，加快技术商业化落地，但弱化监管要求可能会增加人工智能快速扩散过程中的安全风险，形成模型滥用、数据安全、隐私泄露等造成社会影响的问题。

- 2025 年 1 月 23 日，特朗普签署第 14179 号行政令《消除美国在人工智能领域领导地位的壁垒》，要求撤销拜登时期第 14110 号行政令中的人工智能监管措施，废除或放宽阻碍人工智能创新的非必要规章制度，强调维持美国作为人工智能全球领导者的地位。此外，还要求总统科学与技术助理（APST）、AI 与加密货币特别顾问、国家安全事务助理（APNSA）在 180 天内联合制定新的人工智能行动计划。
- 2025 年 4 月 3 日，美国管理和预算办公室发布 M-25-21 号备忘录《通过创新、治理和公众信任加速联邦对人工智能的使用》，正式废除并取代拜登政府时期的 M-24-10 号备忘录。备忘录确立了联邦使用人工智能的三大优先事项：创新、治理和公众信任，放弃对于风险管理的优先关注。保留首席人工智能官制度，但新增了战略创新的核心权责；将拜登时期细分的“影响权利 AI”和“影响安全 AI”简化为“高影响 AI”（High-Impact AI），弱化了公众参与和文档披露要求。
- 同日，美国管理和预算办公室发布 M-25-22 号备忘录《推动政府高效采购人工智能》，废除并取代拜登政府的 M-24-18 号备忘录《推进政府负责任采购人工智能》。确立联邦人工智能采购的三大主题：培育竞争性美国人工智能市场、保障纳税人利益、促进跨职能协作。明确“买美国货”的政策导向，最大化采购美国本土开发的人工智能产品；通过标准化合同条款、数据与模型可移植性防止供应商锁定；简化供应商披露要求，降低合规文档负担。
- 2025 年 7 月 23 日，白宫发布《赢得竞赛：美国人工智能行动计划》（Winning the AI Race: America's AI Action Plan），提出加速人工智能创新、建设人工智能基础设施以及扩大美国国际影响力三大政策方向。该计划强调放宽人工智能研发和应用的监管约束，鼓励企业扩大人工智能技术部署；重点推进数据中心、芯片工厂以及配套能源网络的建设；希望通过技术合作和标准输出扩大美国人工智能生态的国际影响力。
- 2025 年 12 月 11 日，特朗普签署第 14365 号行政令《确保人工智能国家政策框架》，旨在搭建合规负担最低的联邦统一人工智能政策体系，减少各州各自立法和出台监管规则的差异。
- 2026 年 3 月 20 日，白宫发布《国家人工智能立法框架》，系统提出未来联邦人工智能立法的总体方向，强调通过建立统一的法律制度支持美国人工智能长期发展。框架围绕人工智能监管放松、研发创新、数据资源开放、基础设施建设、联邦与州监管协调、人才培养以及国际安全保障等方面提出立法建议。

随着特朗普第二任期人工智能政策的不断推进，治理逻辑由执政初期强调全面减少监管约束，逐步转向以国家安全为导向的精准监管模式。在继续支持产业创新的同时，政府对先进模型、关键技术和高风险应用实施了更加有针对性的治理，但是政策措施仍以自愿合作和行业自律为主，依旧面临建立稳定有效安全边界的挑战。

- 2026年6月2日，特朗普签署第14409号行政令《促进先进人工智能创新与安全》，强调推动人工智能创新与国家安全并举，通过以下措施实现国家安全导向的有限监管：
 - 1) 建立自愿合作框架，确立“30天自愿审查窗口”，允许前沿模型开发者在发布前向联邦政府提供访问权限，但明确禁止设立强制性预审批机制；
 - 2) 要求财政部、国家安全局（NSA）和网络安全与基础设施安全局（CISA）在60天内开发保密基准测试流程和“受涵盖前沿模型”认定机制，但明确强调企业自愿参与；
 - 3) 要求CISA在30天内发布约束性操作指令强化联邦系统网络防御，财政部建立人工智能网络安全信息交换中心协调漏洞管理，OMB审查联邦拨款以支持人工智能漏洞检测技术研发，并指示司法部长优先执法打击利用人工智能进行的网络犯罪。
- 2026年6月4日，美国众议院两党议员联合发布《伟大美国人工智能法案》讨论稿，围绕前沿人工智能治理、劳动力影响、网络安全和研发与国际合作四大支柱展开，核心监管对象为“大型前沿开发商”。

图5：美国拜登政府及特朗普第二任期人工智能治理顶层政策导向演进图



数据来源：美国白宫、联邦公报、OMB备忘录、美国众议院议员官网等，国泰海通证券研究

2.3. 民用领域：采取自愿标准和行业自律，形成轻合规治理模式

美国在民用人工智能领域坚持创新优先的原则，主要依托技术标准、行业自律和传统监管机制开展风险治理，形成以低强制合规和事后监管为特征的治理模式。联邦政府未设定统一的合规阈值，而是通过自愿性技术标准引导企业自主评估和缓释风险；各行业监管机构在既有法律框架内嵌入人工智能治理，大多以事后追责方式处理安全问题。类似措施可以降低企业合规成本，为人工智能技术迭代和商业化应用提供较大空间，但这也意味着企业在风险管理上具有较大自主权，强制监管约束有限。在人工智能快速扩散的背景下，仅依靠自愿标准和行业自律可能难以及时覆盖新型风险，在算法歧视、隐私泄露、模型安全等领域出现监管滞后问题。

2.3.1 NIST 迭代自愿性标准框架，为企业提供可裁剪的风险管理流程

美国依托 NIST 在 2023 年发布的自愿性人工智能风险管理框架（AIRMF）构建技术标准体系，通过提供风险识别、评估和缓释工具，引导企业建立人工智能安全治理流程。NIST 同步更新技术指南、发布系列配套文件、推出开源工具和评估平台，兼顾人工智能技术的灵活发展和风险管理需求。然而，框架本身缺乏强制法律效力，不设定统一合规门槛和处罚机制，执行实际效果高度依赖企业主动采用程度。截至 2026 年 9 月，AIRMF 1.0 正在修订中，修订版尚未正式发布。

- 2023 年 1 月 26 日，NIST 发布《人工智能风险管理框架》1.0 版（AIRMF 1.0），将分散的人工智能风险关注点整合为治理（Govern）、映射（Map）、度量（Measure）和管理（Manage）四个核心功能，给出了有效性和可靠性、安全性和弹性、可说明性和可解释性、隐私增强性、公平-偏见管理、可追责和透明性等可信 AI 系统的特征。同时明确该框架为自愿性框架，不设定合规阈值或处罚。

表2: NIST 《人工智能风险管理框架》四大核心功能具体解释

步骤	具体解释
治理（Govern）	建立组织层面的人工智能风险管理制度、权责划分和问责制度，将人工智能风险纳入公司制度和决策流程，明确责任链路。
映射（Map）	梳理人工智能系统的应用场景、数据流和利益相关者，识别潜在危害的具体类别，形成可追踪的风险登记册。
度量（Measure）	采用红队测试、对抗性攻击模拟、公平性指标评估及持续监控等手段，对模型准确性、公平性、安全性进行量化评估和持续监测。
管理（Manage）	根据风险评估结果对风险进行优先级排序，制定缓释措施和突发事件应急预案，并持续跟踪优化。

数据来源：美国 NIST 官网，国泰海通证券研究

- 2023 年 3 月 30 日，NIST 发布《AIRMF 操作手册》完整版（AIRMF Playbook），通过行动建议和相关指导落实 AIRMF 中的四大核心功能。同步发布配套《路线图》（AIRMF Roadmap）和交叉对照表（AIRMF Crosswalks），确定推进 AIRMF 的关键行动并填补理论和实践方面的空白，明确了 NIST 框架的持续迭代机制。此外，成立 NIST 人工智能资源中心（AIRC），整合配套技术文件与资源，为人工智能的测试、评估与核验工作提供支撑。
- 2024 年 7 月 26 日，NIST 发布《生成式人工智能专项指南》（NIST AI 600-1），是 AIRMF 1.0 主框架的配套跨行业专项实施指南，属于自愿采用标准，专门针对大语言模型和多模态人工智能带来的“幻觉”、合成毒性文本、知识产权侵权、人机交互偏差和数据隐私泄漏等 12 大类生成式人工智能专属风险提供具体的技术排查与处置指南。
- 同日，NIST 发布《人工智能标准全球参与计划》（NIST AI 100-5），同步发布开源 AI 测试平台 Dioptra 1.0 正式版，用于评估、分析和追踪人工智能系统面临的对抗性攻击风险，量化逃逸攻击、数据投毒攻击、模型窃取攻击对模型性能的影响，以开源方式落实算法测试要求而非强制指令。
- 2026 年 2 月，NIST 旗下人工智能标准与创新中心（CAISI）启动人工智能智能体标准倡议，旨在让具备自主执行能力的下一代人工智能（智能体）获得大众信任并广泛落地，牵头推动由行业主导的技术标准与开放协议建设，保障智能体能够代表用户安全运行、在各类数字场景中顺畅跨平台互通。

2.3.2 行业传统监管部门嵌入人工智能治理，以事后执法为主

美国未设立专门的人工智能监管机构，而是依托金融、医疗、消费者保护等传统监管部门，在既有法律授权范围内对人工智能应用进行监管。各机构利用现有监管经验，根据不同应用场景实施精准管理，有针对性地提高治理效率。但是，各行业目前阶段的监管措施仍以建议性质或事后追责为主，难以及时识别人工智能系统运行过程中的潜在风险，监管措施存在一定滞后性。

在金融领域，2024 年 3 月 27 日，财政部根据第 14110 号行政令发布《金融部门管理人工智能特有网络安全风险报告》，针对金融机构使用人工智能的运营风险、网络安全和欺诈问题进行全面评估，并提出解决大型与小型金融机构能力差距、缩小欺诈数据鸿沟、完善监管协调机制、建立数据供应链标准化管理体系等方案；2024 年 12 月 19 日，财政部进一步发布《关于人工智能在金融服务中的使用、机会和风险报告》，总结了金融行业人工智能采纳现状，要求金融机构在部署前严格审查人工智能用例是否合规，推动国内外政企监管协同，但仍是建议性质，未设定强制性合规标准。2024 年以来，美国证券交易委员会（SEC）先后对 Delphia、Presto Automation、Nate Inc 等市场主体提起“AI 清洗”执法行动，并于 2025 年 11 月发布的 2026 财年检查优先事项中，将人工智能相关披露的准确性列为重点审查领域，要求企业确保人工智能能力陈述真实、具体且贴合自身业务实际。

在医疗健康领域，食品药品监督管理局（FDA）对人工智能医疗器械的法定审批与监管，沿用了传统医疗企业审批框架。2025 年 1 月，FDA 发布《人工智能赋能医疗器械软件功能：全生命周期管理与上市提交建议》草案指南，首次系统覆盖 AI/LM 医疗器械从设计、验证到上市后性能监测的全生命周期，要求制造商建立上市后的监测计划。2026 年 1 月，FDA 发布《临床决策支持软件最终指南》，界定了人工智能临床决策支持的监管边界，包括软件不得分析医学影像或诊断信号，须为医疗专业人员提供建议而非替代其判断，且须使临床医生能够独立审查建议依据。同时要求披露算法验证信息与数据来源，以缓解自动化偏见风险。

在消费者保护领域，联邦贸易委员会（FTC）将算法歧视、数据隐私和企业虚假宣传列为人工智能执法重点，但主要依据《联邦贸易委员会法》第 5 条“不公平或欺骗性行为”的概括性授权进行事后追责。2024 年 7 月 23 日，FTC 与司法部、英国竞争与市场管理局（CMA）、欧盟委员会发表《关于生成性人工智能基础模型和人工智能产品竞争的联合声明》，针对核心生产要素垄断、头部企业巩固市场支配地位及科技巨头扼杀初创企业等风险设立联合执法审查标准，确立公平交易、互操作性和多元选择三大原则。2026 年 2 月 23 日，司法部反垄断司和 FTC 就竞争对手间合作的潜在指导启动了联合公开调查，对人工智能、算法定价和信息交换更新协作指南，明确提出共同接入一个 AI 模型或共享敏感数据的算法定价合谋属于违法行为。

2.3.3 鼓励企业自愿承诺，通过市场自律形成公私合作机制

美国政府引导企业签署人工智能安全承诺、建立行业联盟和开展安全测试，依托市场自律形成政府与企业共同参与的治理模式。这种机制能够充分调动领先科技企业的技术资源和市场引导能力，构建围绕红队测试与透明度提升的治理生态，具有灵活性强和响应速度快的优势。但安全承诺主要依靠企业自律参与，缺乏统一监督和强制审查机制，不可避免企业治理投入不足、信息披露不充分等问题。

- 2023 年 7 月，白宫召集亚马逊、Anthropic、谷歌、Inflection、Meta、微软和 OpenAI 七家领先人工智能公司签署自愿安全承诺。要求企业加强模型发布前的内外部红队测试，共享安全风险信息、通过第三方评估提高模型透明度、开发水印等技术机制。同时，企业承诺推动内容标识技术发展，并优先研究社会

风险的评估机制。2023 年 9 月 12 日，承诺范围扩展至 Adobe、英伟达等八家额外企业，2024 年 7 月 26 日苹果加入，使企业总数达到 16 家。

- 2024 年 2 月 8 日，美国商务部长宣布创建人工智能安全研究所联盟(AISIC)。AISIC 首批成员超过 200 家，涵盖 OpenAI、谷歌、Anthropic、亚马逊、微软、Meta、英伟达、Palantir、英特尔、摩根大通、美国银行等科技巨头与金融机构，以及卡内基梅隆大学、斯坦福大学、康奈尔大学等顶尖学术机构。该联盟旨在联合 AI 创作者与用户、学术界、政府和产业研究人员及民间社会组织，共同支持安全可信 AI 的开发与部署，重点制定红队测试、能力评估、风险管理、安全与防护以及合成内容水印等方面的指南。至 2025 年 2 月，成员已扩展至超过 290 家。
- 2026 年 5 月，美国人工智能标准与创新中心(CAISI)宣布已与谷歌、微软、xAI 等企业建立合作机制，允许政府在合作伙伴的 AI 模型向公众发布前进行独立评估，并在部署后继续检测。

2.4. 关键领域：军事与前沿模型领域实施严格监管和前置审查

美国在军事国防、前沿模型和网络安全等涉及国家安全的关键领域实施更加严格的监管，通过安全评估、风险测试和技术管控降低人工智能带来的潜在风险。相较于民用领域以自愿标准和行业规范为主的治理模式，美国在国家安全领域更加强调风险管控。面对模型幻觉、对抗攻击、数据投毒等内生风险，各部门着力提高模型的可靠性、鲁棒性和可控性；面对利用人工智能增强网络攻击、生物武器研发等双重用途风险，强调防止技术能力向威胁现实安全的领域扩散。美国通过战略定位升级、前置安全评估、红队测试和漏洞管理等方式建立安全边界。但是当前相关领域治理仍高度依赖行政命令、自愿承诺和临时性措施，对于模型能力快速演进带来的新型风险，监管规则的稳定性和强制约束力仍有待加强。

2.4.1. 军事国防：人工智能战略地位升级，强制安全标准纳入采购流程

美国政府持续强调人工智能在军事作战体系中的重要地位，通过战略计划、采购标准和组织调整提升人工智能在情报分析和作战决策中的应用能力。国防部通过法案和备忘录等行政手段完成人工智能在军事体系中的战略升级。《国防授权法》对人工智能的政府采购实行硬性安全标准。此外，政府完善人工智能系统的安全评估流程，同步关注技术应用的可信性和透明度。

- 2023 年 8 月 10 日，美国国防部宣布成立生成式人工智能工作组，又称利马工作组，以战略性方式评估、利用和提升生成式人工智能技术能力，帮助美军在作战能力、军事管理、卫生健康、战备及政策等领域全面赋能。该工作组已于 2024 年 12 月解散。
- 2023 年 11 月 2 日，国防部发布《数据、分析和人工智能应用战略》，提出改善基础数据管理、提供联合作战能力、加强治理消除政策障碍、投资可互操作联合基础设施、推进合作生态系统建设、扩大数字人才管理六大核心目标，并制定 2024-2025 年度人力资本计划(HCOP)，启动数字人才飞行员倡议以压缩高精尖专家外聘周期。
- 2025 年 6 月，国防部高级研究计划局(DARPA)在 FY2026 预算中大幅扩大人工智能相关研发投资，新设“新兴人工智能研究与增长”(AIFORGE)项目，将国家安全考量融入前沿生成式人工智能开发生命周期，联合高校与前沿人工智能企业资源，解决人工智能在情报分析、战术规划、自主系统等军事应用中的可信性、鲁棒性和透明度问题。

- 2025 年 12 月 18 日，特朗普签署《国防授权法案（FY26 NDAA）》，第 1513 条首次硬性要求国防部针对其采购的 AI/ML 技术制定网络与物理安全框架，并将其直接锁入《国防联邦采购补充条例》（DFARS）和“网络安全成熟度模型认证”（CMMC）项目，所有国防部人工智能供应商和承包商都必须强制执行此人工智能安全硬性标准。此外，法案同时强制要求全军年度网络安全培训必须加入人工智能引起的独特网络安全挑战。
- 2026 年 1 月 9 日，国防部推出“人工智能加速战略”执行备忘录，提出构建“AI 优先”作战力量的全面规划。同日，国防部发布新版“国防部人工智能战略”和“转变国防创新生态系统以加速作战优势”备忘录，明确首席数字与人工智能办公室（CDAO）为人工智能战略落地主体，重新定义“负责任 AI”仅遵守法定强制约束、拒绝额外意识形态类模型调校。
- 2026 年 6 月 5 日，特朗普签署第 11 号国家安全总统备忘录（NSPM-11）《国家安全体系中的人工智能》，将人工智能定性为美国历史上对国家安全最具变革性的技术之一，要求军方和情报体系加快采用人工智能，同时建立人工智能系统的安全评估流程，确立采纳（Adoption）、适配（Adaptation）、保障（Assurance）、问责（Accountability）四大政策方向。

表3：美国第 11 号国家安全总统备忘录四大政策方向具体解释

政策方向	具体解释
采纳（Adoption）	要求国防部、国家情报总监办公室在 120 天内完成 AI 采购流程改革，消除部署障碍；
适配（Adaptation）	在军方与情报体系领域建立多供应商框架，禁止单一供应商主导涉密 AI 业务；建设新一代高安全等级算力基础设施，组建由政府体系外顶尖专家参与的“AI 国家安全战略配置”；
保障（Assurance）	规定 AI 必须稳定、可控且可受指挥意图操控；加入“禁止擅自禁用”条款，禁止商业实体未经批准擅自禁用、削弱或修改已部署的政府 AI 系统。
问责（Accountability）	将 AI 使用责任锚定于军事与情报指挥链内部，而非外部文职监管机构。

数据来源：美国白宫，国泰海通证券研究

2.4.2. 前沿模型：治理规则持续迭代，采取风险评估、前置审查与安全管控

拜登政府强制要求大模型开发者向联邦政府报告安全测试结果，特朗普政府则撤销该强制义务，侧重于以前置审查和风险评估为主的自愿性措施。不同于民用领域的事后监管，在前沿模型治理领域两任政府均提出事前风险评估。但是，NIST 出台的风险管理指南和特朗普政府发布的行政令均非强制性措施，《伟大美国人工智能法案》尚未正式实施，前沿模型监管措施的强制约束力和稳定性不足。近期，特朗普政府对 Anthropic 和 OpenAI 等前沿科技公司发布的最新模型采取了出口管制、访问限制等强制措施，但普遍是临时的应急式管控，缺少常态化风险治理的统一硬性约束。

- 2023 年 7 月 26 日，Anthropic、谷歌、微软和 OpenAI 四家美国前沿人工智能公司联合成立前沿模型论坛，作为独立的非营利性行业组织，专注于推进前沿人工智能安全研究、识别最佳实践、促进政府与学术界的信息共享，亚马逊和 Meta 于 2024 年 5 月 20 日官宣加入。该论坛发布的技术报告涵盖 CBRN（化学、生物、放射、核）和网络风险的前沿能力评估，并制定行业安全标准。

- 2023 年 10 月 30 日，拜登签署的第 14110 号行政令中首次要求前沿 AI 模型开发者（训练算力超过 10^{26} FLOPs）在公开发布前必须向联邦政府报告安全测试结果。该行政令于 2025 年 1 月 20 日被特朗普政府撤销，终止了强制报告安全测试结果的要求。
- 2023 年 11 月，拜登政府设立美国人工智能安全研究所（AISI），2025 年 6 月特朗普政府将其重组并更名为人工智能标准与创新中心（CAISI），作为跨部门协调机构，负责前沿人工智能模型的安全评估、测试标准开发及联邦机构协调工作。
- 2024 年 7 月 26 日，NIST 发布《双用途基础模型滥用风险管理指南》初稿（NIST AI 800-1），针对可被滥用于生物武器、网络攻击等场景的双用途模型提出七大管理目标：识别潜在滥用风险、制定滥用风险管理计划、保护模型免遭未经授权访问、测量部署相关滥用风险、部署前缓解滥用风险、监控与响应滥用事件、披露滥用风险管理实践。虽然指南聚焦于前沿大模型的恶意滥用，但是依旧属于自愿性行业标准，未设置强制监管规范或处罚条款。
- 2024 年 11 月 20 日，AISI 牵头成立人工智能国家安全风险测试工作组（TRAINS），联合国防部、国土安全部等跨部门资源，开展先进人工智能模型的联合风险评估、红队测试和基准开发。
- 2026 年 6 月 2 日，特朗普签署第 14409 号行政令《促进先进人工智能创新与安全》，要求前沿 AI 实验室在发布具备先进网络能力的模型前，向联邦政府提供最多 30 天的发布前访问期，用于基准测试和安全评估。但是行政令明确拒绝了强制性政府许可、预审批或许可证制度，强调相关内容不得被解释为授权建立强制性政府许可。
- 2026 年 6 月 4 日，《伟大美国人工智能法案》的立法草案要求“大型前沿开发商”每半年接受独立第三方安全审计，公开灾难性风险管理框架，15 日内报告关键安全事件，并接受 FTC 执法。
- 2026 年 6 月 12 日，美国商务部以国家安全出口管制为由，要求 Anthropic 限制外国用户访问最新发布的 Fable 5 和 Mythos 5 模型，防止安全漏洞和越狱问题。同年 6 月 30 日，商务部即解除相关管制，Anthropic 作出包括发布前开放政府测试、快速共享安全信息、投入专门资源支持政府研究、制定前沿模型自愿安全规范四项承诺。
- 2026 年 6 月 26 日，OpenAI 发布 GPT-5.6 模型家族，在 EO14409 框架下被限制仅向约 20 家政府批准的可靠合作伙伴提供访问，是美国人工智能公司首次在政府管理的访问列表下推出前沿模型。

2.4.3. 网络安全：推动人工智能融入网络防御体系，强化关键基础设施安全治理

美国将人工智能视为提升网络防御能力的重要工具，同时加强对人工智能自身安全风险的治理，通过漏洞管理、风险评估和关键基础设施保护强化安全治理。政府一方面通过行政手段和战略部署，发挥人工智能在革新网络防御体系中的技术优势，提高网络安全治理效率；另一方面针对人工智能系统面临的模型投毒和数据泄露等安全风险，通过 NIST 风险框架、科研数据集开放、漏洞协调机制和关键基础设施责任划分，构筑人工智能安全防线。

- 2024 年 11 月 14 日，美国国土安全部发布《关键基础设施中人工智能的角色与职责框架》，为美国 16 个关键基础设施领域（能源、通信、公共卫生、交通等）中人工智能的安全开发和部署提供指导。将人工智能生态系统划分为云和计算基础设施提供商、AI 开发员、关键基础设施所有者和运营商、公民

社会、公共部门五个关键角色，明确保护人工智能开发和部署环境安全、推动负责任的模型与系统设计、实施数据治理、确保安全部署、监测性能与影响五大责任领域。

- 2025 年 6 月 6 日，白宫发布第 14306 号行政令《继续努力加强国家网络安全》，新增第五条专门部署“依托人工智能、保障人工智能自身安全”，指出人工智能有望革新网络防御体系。要求商务部长通过 NIST 主任牵头，在兼顾商业保密与国家安全的前提下，最大程度将现有网络防御研究数据集向更广范围的学术研究机构开放。同时要求国防部长、国土安全部长、国家情报总监将人工智能软件漏洞、系统遭入侵处置工作纳入各机构现有漏洞管理流程与跨部门协同机制，覆盖事件追踪、应急处置、信息上报全环节。
- 2025 年 12 月 16 日，NIST 发布《网络安全框架人工智能配套规范》（NISTIR 8596, Cyber AI Profile）初步草案，依托网络安全框架 CSF 2.0 补充人工智能场景安全指引，指导企业和关键信息基础设施实体防范 AI 模型面临的数据投毒、对抗性样本攻击、模型权重被盗等特有网络安全风险。
- 2026 年 3 月 6 日，美国白宫发布《特朗普总统的美国网络战略》，以维持美国在网络空间的绝对领先地位为核心目标。强调推动联邦政府网络安全与现代化，将网络战技术重点集中于后量子密码学和人工智能驱动的安全解决方案。同时提出构建涵盖数据中心的人工智能技术体系，抵制实施审查、监视的外国人工智能平台。
- 2026 年 7 月 14 日，白宫正式启动“金鹰”（GOLD EAGLE）倡议，作为第 14409 号行政令设立的人工智能网络安全漏洞协调中心，由财政部牵头，国土安全部、战争部参与，与人工智能企业和关键基础设施运营商建立自愿合作机制，实现漏洞信息的快速共享、验证和修复。

图6：美国人工智能治理体系特征（1）



数据来源：美国白宫、美国 NIST、美国财政部、美国 SEC、美国 FDA、美国 FTC、美国商务部、美国国防部、美国政府出版局 GPO、美国国土安全局、众议院议员官网、Anthropic 官网、Frontier Model Forum，国泰海通证券研究

2.5. 注重创新：多措并举扶持人工智能产业技术领先优势

美国人工智能产业政策长期围绕保持技术领先和强化创新生态展开，通过完善算力基建、联邦资金投入和人才培养体系建设等方式巩固产业发展根基。政府更多依靠战略引导，将算力、资本、人才等关键创新要素纳入国家战略，以此快速聚集领先企业和科研资源共同参与人工智能创新投入。

2.5.1. 完善算力基建与能源配套，夯实人工智能发展的技术底座

美国支持建设国家级人工智能计算资源平台和大型数据中心建设，并同步完善能源配套改革，持续强化人工智能技术进步的基础保障。这些措施涵盖推出试点计划、发布行政令与战略文件、设立专项投资计划，有效缓解了人工智能训练和推理所需的算力瓶颈，为前沿模型研发提供物理基础和电力保障，同时保障了居民用电的基本权利。但是，为了迎合人工智能算力需求的快速增长，特朗普政府在数据中心审批方面取消原有的土地和环境要求，为技术生态的可持续发展埋下隐患。

- 2024 年 1 月，美国国家科学基金会（NSF）联合 14 个联邦机构和 28 个私营部门合作伙伴正式启动国家人工智能研究资源（NAIRR）试点计划，通过统一门户为全美研究人员和教育工作者提供共享的先进计算资源、数据集、预训练模型和软件工具，普惠开放人工智能研究基础设施的访问。
- 2024 年 4 月 29 日，美国能源部（DOE）发布《人工智能和能源：现代电网和清洁能源经济的机遇》，提出从多方面加快人工智能工作，不仅保持全球竞争力，还要管理人工智能日益增长的能源需求；同时推出 1300 万美元的 VoltAIc 专项计划，自研政策大模型 PolicyAI，用人工智能简化能源项目审批，实现政府监管流程数字化治理。
- 2025 年 1 月 21 日，特朗普总统宣布启动“星际之门”（Stargate Project）人工智能基础设施投资计划，目标是在全国各地建设数据中心，为下一代人工智能提供物理和虚拟基础设施支撑。
- 2025 年 7 月 23 日，特朗普签署第 14318 号行政令《加速联邦数据中心基础设施许可》，具体措施包括：简化大型 AI 数据中心的审批流程；松绑《国家环境政策法》（NEPA）监管；商务部发起为新增用电负荷超过 100 兆瓦、资本承诺不低于 5 亿美元的“合格项目”提供贷款、赠款或税收激励；要求内政部、能源部和国防部识别并授权在联邦土地及军事基地闲置地块上建设数据中心，用于人工智能的训练、推理、仿真和合成数据生成。同时，废除了 2025 年 1 月拜登政府发布的第 14141 号行政令《加强美国在人工智能基础设施领域的领导地位》，取消了原有的清洁能源建设和气候合规要求。
- 2025 年 9 月 2 日，NSF 发布 NSF25-546 号项目征集公告，响应白宫的《美国人工智能行动计划》，计划投入 3500 万美元为人工智能研究资源 NAIRR 设立一个运营中心（NAIRR-OC），推动该试点从试验阶段向可持续的国家基础设施过渡，旨在整合并为研究人员提供算力、数据和模型等研究基础资源。
- 2025 年 10 月 23 日，能源部（DOE）部长向联邦能源监管委员会（FERC）提交《关于加速大负荷电力用户并网的提前制定规则通知》，要求 FERC 行使联邦管辖权，简化并加速超过 20 兆瓦的大型负荷设施接入州际输电系统的审批流程。同时，推动各州为人工智能数据中心创建新的公用事业费率等级，防止居民电价因算力需求激增而飙升。
- 2025 年 11 月 24 日，特朗普签署第 14363 号行政令，启动“创世纪计划”（Genesis Mission），由能源部牵头整合国家实验室、企业、高校数十年积累的海量科研数据资源，对标曼哈顿计划搭建全国统一人工智能科研算力数据平台，用于训练科学基础模型和研发自动化科研智能体。
- 2026 年 3 月 4 日，特朗普在白宫召集七家领先科技公司共同签署《用电户费率保护承诺》，由企业自愿承诺全额承担数据中心所需新增发电资源及电网升级的成本，不转嫁给居民用户，同时与公用事业公司和州政府就独立的费率结构进行谈判。
- 2026 年 6 月 18 日，FERC 发布六项专项指令，依据《联邦电力法》第 206 条正式发出责令说明理由命令，迫使全美六大区域电网运营商在 60 天内证明现行电网定价结构能够容纳人工智能数据中心，或立即进行价格改革；同时在 30 天内提交强制性电网可靠性报告，确保有足够发电能力支持人工智能算力基建。

2.5.2. 加大联邦资金投入，引导社会资本协同助力产业创新

美国通过联邦资金投入和公私合作机制，撬动社会资本共同参与人工智能基础研究，促进政府资金和社会资本的协同发展。美国联邦政府各部门通过年度预算倾斜和专项拨款，向基础研究、技术应用等领域投入大规模资金。此外，依托国家基金会与企业共同出资建设研究院和实验设施，提升基础研究投入的效率。

- 2025 年 1 月 21 日，“星际之门”人工智能基础设施投资计划正式启动，计划未来四年投资 5000 亿美元、立即部署 1000 亿美元，是由甲骨文公司、OpenAI 和日本软银集团联合出资打造。
- 2025 年 5 月 22 日，美国众议院表决通过《大而美法案》（the One Big Beautiful Bill Act）。该法案包含一系列人工智能专项资金安排，例如，国防部专项经费 1.24 亿美元升级测试资源管理中心的人工智能技术能力、2.5 亿美元拓展网络司令部人工智能相关工作、2 亿美元部署自动化和人工智能工具加快财务报表审计工作。
- 2025 年 5 月 30 日，NSF 向国会提交 FY2026 预算请求，将人工智能列为三大“关键活动”之一，申请 655.23 百万美元用于人工智能投资，较 FY2024 当前计划增加 1958 万美元（增幅 3.1%），明确将投资聚焦于机器学习、计算机视觉和自然语言处理等基础研究领域。
- 2025 年 7 月 29 日，NSF 宣布与 Capital One 和 Intel 合作投入 1 亿美元支持五个国家人工智能研究院和一个中央社区中心，重点覆盖人工智能材料研究、药物开发、人机协作和 STEM 教育等方向。同期，NSF 还启动多项配套投资：1 亿美元用于建设可编程云实验室网络（NSF Test Bed），7500 万美元用于开发开放多模态 AI 基础设施以加速科学研究，2000 万美元用于扩展 NSF CloudBank 以增加商业云计算和先进云服务的访问。
- 2025 年 12 月 10 日，美国能源部（DOE）宣布为“创世纪计划”（Genesis Mission）追加超过 3.2 亿美元的投资，用于推进人工智能赋能科学研究。该投资将建设“美国科学云”（AmSC）和变革性人工智能模型联盟（ModCon），并资助 14 个机器人与自动化实验室项目、37 项基础人工智能奖项。

表4：美国本土 AI 及半导体领域公司投资额与投资重点

企业	投资额	投资重点
Meta	6000 亿美元	人工智能基础设施建设与人才扩充
苹果（Apple）	6000 亿美元	制造业务与人员培训
英伟达（NVIDIA）	5030 亿美元	人工智能基础设施与超级计算机
软银、OpenAI、甲骨文（Oracle）（联合）	5000 亿美元	人工智能基础设施（星门计划）
亚马逊（Amazon）	3400 亿美元	人工智能基础设施、数据中心与云业务扩容
美光（Micron）	2030 亿美元	半导体制造、研发
Anthropic	500 亿美元	人工智能基础设施
环球晶圆（GlobalWafers）	40 亿美元	半导体晶圆工厂
NorthMark Strategies	28 亿美元	超级计算中心
安靠科技（Amkor Technology）	20 亿美元	半导体制造工厂

西门子 (Siemens)	4.5 亿美元	人工智能数据中心及制造业务
埃文特科技 (Avant Technology)	2000 万美元	半导体制造

注：本表为各企业向美国政府或媒体披露的美国本土 AI 及半导体相关投资、资本开支承诺，大多为 2025—2030 跨年规划口径，涵盖算力基建、半导体制造、技术研发、云业务、人员培训等数据来源：美国白宫，国泰海通证券研究

2.5.3. 重视专业人才教育，构建覆盖全民的培养体系

美国通过优化人才引进机制、加强人工智能教育普及和职业培训，提升人工智能领域的人才供给能力。政府改革技术岗位聘用流程、放宽移民签证政策吸纳海内外高端技术人才，同时加强基础教育和高等教育阶段的人工智能素养普及，依托国家政策和专项基金推动职业技能培训，填补人工智能时代专业技术人才缺口。

- 2023 年 10 月 30 日，拜登政府第 14110 号行政令将扩大人工智能领域的移民与人才引进作为重要政策内容，简化人工智能人才签证流程。2024 年 3 月 28 日，OMB 发布的 M-24-10 号备忘录，将扩大人工智能人才招聘、简化技术专家入职流程作为重要配套措施。
- 2025 年 4 月 23 日，特朗普签署第 14278 号行政令《为美国人未来的高薪技能工种做好准备》，确立国家战略以投资和发展美国劳动力。劳工部 (DOL) 据此搭建人工智能人才培养框架：面向所有学徒职业普及人工智能素养课程、拓展从事人工智能开发与系统管理的核心技术岗位学徒项目、扩大数据中心配套基建技术工种培训规模。
- 2025 年 8 月 12 日，劳工部、商务部和教育部联合发布《美国人才战略》，提出五大支柱计划，要以产业需求导向开展劳动力培育、提升劳动者人工智能适配能力与职业流动性、建立适配人工智能时代灵活创新的政策工具，为受到人工智能转型影响的工人和社区提供保护。
- 2025 年 9 月 30 日，劳工部 (DOL) 向 14 个州授予超过 8600 万美元的产业驱动技能培训基金，用于人工智能基础设施及相关高需求行业的人才培训。
- 2025 年 10 月 2 日，计算研究协会 (CRA) 在 NSF 支持下启动 NAIRR AI 教育奖学金计划，面向全国选拔 10 名教职员工冠军，推进本科和硕士层次的人工智能教育建设。入选者依托 NAIRR 试点课堂拓展人工智能教育研究写作网络 (NAIRR AIEDU RCN) 开展一年工作，将试点资源用于人工智能课程教学。
- 2026 年 1 月 6 日，劳工部宣布预计拨款 1.45 亿美元用于扩大人工智能基础设施等关键行业的注册学徒制，采用创新的按绩效付费模式以确保资金使用的问责性。
- 2026 年 3 月 20 日，白宫发布的《国家人工智能立法框架》进一步将“全民人工智能教育，培育适配人工智能时代的劳动力”列为政策要点之一，强调通过教育改革和职业培训保障美国在人工智能时代的人才竞争力。
- 2026 年 6 月 4 日，《伟大美国人工智能法案》立法草案明确将人工智能素养纳入 K-12 基础教育体系，支持结合区域产业需求开展职业技术教育，培养应用型人工智能技能人才；同时对受人工智能冲击的失业人员提供帮扶服务，并放宽高端人工智能人才的招录限制。

2.6. 对外博弈：持续升级出口管制，全栈技术输出以巩固盟友体系

人工智能已成为美国维护科技领先优势和国际影响力的核心工具，美国通过出口管制、规则输出与技术生态建设，巩固自身在人工智能全球治理体系中的话语权。美国通过出口管制和投资限制，将遏制范围从半导体硬件扩展至先进模型权重，在促进国内人工智能创新的同时限制他国产业发展。同时，借助多边机制向盟友输出风险评估标准和安全底线，通过芯片供给、数据中心建设和人才培养巩固美国的全球主导地位。

2.6.1. 出口管制：从芯片禁运到模型权重限制、由广泛限制转向精准管制

美国持续扩大人工智能出口管制范围，从先进芯片限制延伸至模型权重和投资活动，以维护美国在人工智能产业链中的技术优势。财政部限制美国主体对人工智能系统的投资，商务部将模型权重列入出口管制，按盟友关系远近实施分级许可。特朗普政府虽撤销了部分广泛合规要求，但转而采取更精准的管制措施，对先进芯片出口实施逐案许可审查，形成了覆盖完整技术链条的全球性限制网络，降低竞争对手获取关键技术的可能性。

- 2024年10月28日，美国财政部发布“关于实施出境投资行政命令(EO 14105)的最终规则”，限制美国主体对特定受关注国家的人工智能系统、尖端半导体和量子计算领域进行投资，明确规定禁止投资用于军事、重大情报监视用途或算力超 10^{25} 次计算操作训练的人工智能模型，对于未达到禁止标准，但使用算力超 10^{23} 次计算操作训练的其他人工智能系统投资，必须向财政部报备。
- 2025年1月13日，美国商务部与工业安全局(BIS)发布临时最终规则《人工智能扩散规则》，旨在控制先进人工智能技术的全球传播。通过对最先进半导体和相关模型权重实施对于目的地、数量和最终用户的管制，以降低恶意行为方获取最先进AI模型的可能性。
 - 1) 将全球国家及地区按照与美国盟友关系的远近分为三个等级，中国(大陆、中国香港、中国澳门)与伊朗、俄罗斯、朝鲜被划属第三等级，几乎完全被禁止采购美国AI芯片。
 - 2) 首次将最先进AI模型的“模型权重”(Model Weights)列入全球出口管制(新设ECCN 4E091)，针对最先进半导体和某些高级闭源AI模型制定全球许可要求。
 - 3) 采用外国直接产品规则(FDPR)，即在境外只要使用美国技术或软件训练出来的最先进AI模型权重，也受美国出口管制约束。
- 2025年5月13日，美国商务部宣布撤销拜登时期的《人工智能扩散规则》广泛合规要求，声明中称该规则将扼杀美国创新并给企业带来繁重的监管负担。同时，BIS发布三项更精准的管制措施：提醒公众允许美国人工智能芯片用于中国人工智能模型训练和推理可能带来的后果；明确在全球任何地方使用华为昇腾(Ascend)芯片均违反美国出口管制规定；指导美国企业防范供应链外流、规范供应链布局。
- 2025年12月8日，特朗普总统宣布美国将允许将H200和类似产品有条件出口给经批准的中国客户，但申请者必须证明不会减少美国本土的半导体产能，买方必须采用严格的出口合规与客户筛选程序，且产品必须在美国本土通过独立的第三方安全与性能测试。然而，美国参议院提出的《安全芯片法案》(SAFE Act)立法提案，要求商务部在30个月内停止向包括中国和俄罗斯等国的买家发放H200、Blackwell等先进人工智能芯片的出口许可。
- 2026年1月，BIS修订了运往中国大陆、中国澳门的先进人工智能芯片的出口审查政策，将目的地的高性能人工智能芯片许可审查政策从推定拒绝转变为有严格条件的逐案许可，旨在平衡对美国人工智能相关技术的出口和采用

的支持与保护美国国家安全利益的需要。

2.6.2. 规则与技术输出：通过治理规则输出和全栈技术出口强化盟友绑定

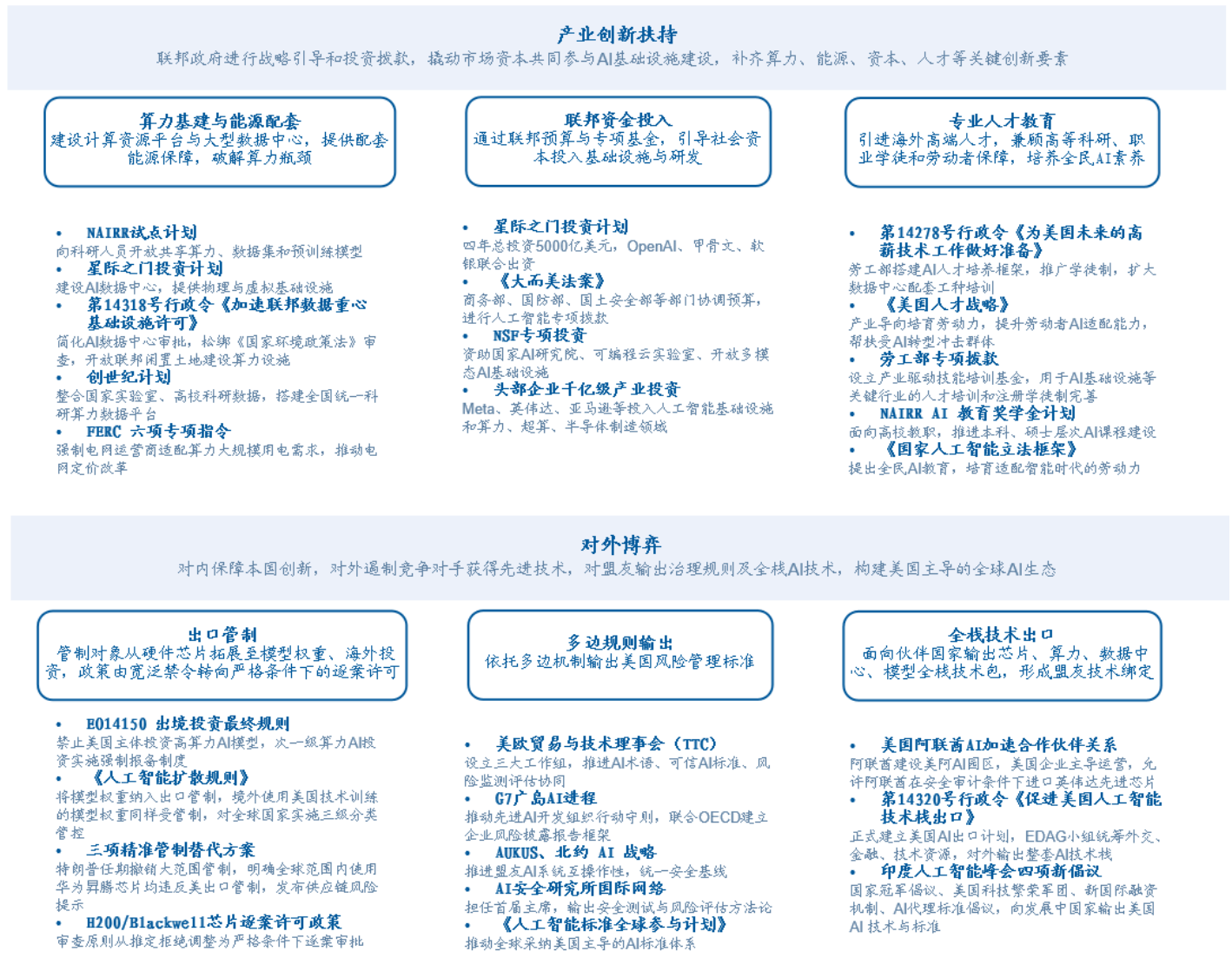
美国积极参与国际治理机制并推广风险管理标准，扩大自身在全球人工智能治理体系中的影响力。美国通过 TTC、G7 广岛进程、人工智能安全研究机构国际网络等多边机制与双边机制，吸纳盟友采纳美国主导的 AI 术语、风险管理标准和国际合作原则，试图构建以美国为中心的人工智能全球治理生态。

- 美欧通过贸易与技术理事会（TTC）推进联合人工智能路线图，设立三大专家工作组，分别聚焦 AI 术语与分类体系、可信 AI 标准及风险管理工具、AI 风险监测与量化评估。双方于 2023 年 5 月出台 65 项 AI 核心术语清单，配套美国、欧盟各自解读口径，以及欧美双方达成共识的统一定义，为理解人工智能风险治理路径提供基础依据。
- 2023 年 5 月，G7 广岛峰会期间，包括美国在内的七国集团领导人正式启动“广岛 AI 进程”（Hiroshima AI Process），旨在应对先进人工智能系统带来的机遇与挑战。2023 年 9 月 7 日，G7 数字与技术部长会议通过《广岛 AI 进程：G7 数字与技术部长声明》，提出制定适用于所有人工智能行为体的国际指导原则、开发先进人工智能系统组织的行动守则，以及基于项目的负责任 AI 工具与最佳实践合作。2023 年 10 月 30 日，G7 领导人发表《广岛 AI 进程领导人声明》，指示相关部长在年底前完成综合政策框架，并与全球人工智能伙伴关系（GPAI）和经济合作与发展组织（OECD）合作，开展多利益相关方外联与磋商，推动建立安全、可靠和可信赖的人工智能系统。
- 2023 年 7 月 29 日，美澳双方部长级磋商（AUSMIN）联合声明中，双方重申了要安全、合乎伦理地发展人工智能等关键新兴技术。澳方认可美国推出的企业人工智能风险自愿管控承诺，明确双方将围绕人工智能安全治理开展协同合作。“美英澳三边安全伙伴关系”（AUKUS）Pillar II 将人工智能与自主系统列为关键优先领域之一，并通过弹性自主人工智能技术（RAAIT）试验推进三边人工智能系统的互操作性与互换性，确保未来联合作战模型遵循统一安全基线。
- 2024 年 7 月 10 日，北约发布修订版人工智能战略，在 2021 年版本基础上纳入生成式 AI 等最新技术进展。新战略明确优先事项包括：推进北约负责任使用原则的实施、增强联盟内 AI 系统的互操作性、促进人工智能与其他新兴颠覆性技术的融合、通过加强与盟友工业界和学术界及志同道合伙伴的合作扩展北约 AI 生态系统。
- 2024 年 7 月 26 日，NIST 发布《人工智能标准全球参与计划》（NIST AI 100-5），旨在推动全球人工智能共识标准的制定实施、合作协调以及信息共享，预期目标包括制定科学合理的人工智能标准、反映多元利益相关方需求、开放透明的制定过程、加强国际关系。
- 2024 年 11 月 20 日，美国商务部与国务院在旧金山联合举办人工智能安全研究机构国际网络（International Network of AI Safety Institutes）首次会议，正式宣告该网络建立。初始成员包括美国、澳大利亚、加拿大、欧盟、法国、日本、肯尼亚、韩国、新加坡和英国，美国暂任首届主席。该网络宣布投入超过 1100 万美元用于合成内容风险研究资金，启动关于人工智能安全和创新的国际合作。

在治理规则输出的基础上，美国通过人工智能产业基础设施部署和全栈技术出口，使得盟友形成对美国人工智能技术的深度依赖。美国通过芯片供给、数据中心建设和人才培养，使得伙伴国家的产业发展环节与美国人工智能技术体系实现战略绑定，形成从硬件到应用层的完整出口体系。

- 2025 年 5 月 15 日，美国和阿联酋联合宣布“美阿人工智能加速合作伙伴关系”（US-UAE AI Acceleration Partnership），在阿布扎比建设美阿大型人工智能园区。该项目由阿联酋 AI 企业 G42 牵头建设，引入 OpenAI、Oracle、英伟达、思科、软银等美国生态伙伴，美国商务部与阿联酋达成初步协议允许其每年进口至多 50 万片英伟达最先进 AI 芯片，出口附带严格的安全协议与审计条件。阿方则同意由 OpenAI 和 Oracle 主导数据中心运营，项目受双边安全工作组的监督。
- 2025 年 7 月 23 日，特朗普签署第 14320 号行政令《促进美国人工智能技术栈出口》，强调通过促进全栈美国人工智能技术包的出口来支持美国人工智能产业，通过支持美国原发人工智能技术的全球部署来减少对他国开发的人工智能技术的依赖，建立美国人工智能出口计划。该行政令指示经济外交行动小组(EDAG)协调技术、金融和外交资源以加速优先人工智能出口包的部署，同时支持伙伴国家营造有利于美国人工智能系统部署的创新监管环境。
- 2026 年 2 月 20 日，在“印度人工智能影响力峰会”上，美国正式阐述通过尖端人工智能技术支持全球盟友的战略，推出“美国人工智能出口计划”框架下的四项新倡议：
 - 1) 国家冠军倡议：将合作国领先 AI 公司纳入美国定制化的人工智能出口体系；
 - 2) 美国科技繁荣军团：作为和平队新倡议向进口伙伴派遣志愿技术人才，为部署强大的人工智能应用提供“最后一公里”支持；
 - 3) 新国际融资机制：财政部在世界银行启动新基金，联合美国进出口银行、国际开发金融公司、国务院及小企业管理局等多机构，帮助发展中国家克服人工智能采用的资金障碍；
 - 4) AI 代理标准倡议：由 NIST 及其人工智能标准与创新中心主导，促进代理式人工智能的互操作性与安全标准开发，推动国际推广。

图7：美国人工智能治理体系特征（2）



数据来源：美国白宫、NAIRR 项目官网、美国能源部、美国 NSF 国家科学基金会、美国联邦能源监管委员会、美国国会官网、美国出版局 GPO、美国劳工部、美国财政部、CRA 计算研究协会、联邦公报、美国工业与安全局、G7 文档库、澳大利亚国防部、北约 NATO、国泰海通证券研究

3. 欧洲人工智能治理体系

3.1. 顶层治理模式：由欧盟专门机构统筹部署，成员国具体执行落地

3.1.1. 欧盟统筹：依托人工智能办公室和人工智能委员会强化统一治理

欧盟人工智能办公室承担起法案实施与统一监管的职能。2024 年 1 月 24 日，欧盟人工智能办公室（European AI Office）宣布成立，隶属于通信网络、内容和技術总局（DG CONNECT）。作为《人工智能法案》的核心执行机构，需要确保法案在成员国间一致适用，对通用人工智能（GPAI）模型进行监督、评估和分类，制定行为准则和法案实施指南。该办公室被赋予实质性执法权限，可以调查侵权行为，要求提供技术文档或访问模型以评估能力，在违规情况下还可要求采取纠正措施、限制模型公开并直接处以罚款。此外，该办公室还承担了促进可信人工智能发展和推动国际合作的职能。

人工智能委员会推动成员国间的监管协同，并在欧盟人工智能政策的制定中发挥作用。《人工智能法案》于 2024 年 8 月 1 日生效，据此设立欧盟人工智能委员会 (AI Board)。委员会由每个成员国的代表组成，欧洲数据保护监管局 (EDPS) 等机构以观察员身份参与。其核心职责是协调各成员国监管机构，以确保法案在成员国间的协调一致。同时，就《人工智能法案》的实施和人工智能创新政策、国际合作向欧盟委员会提供咨询和政策建议。通过制定统一的政策框架，确保人工智能技术惠及所有欧洲公民，同时坚持最高的安全和道德标准。

欧盟同时设置了以独立科学专家小组和咨询论坛为代表的专家机构，为人工智能发展提供技术等外部支持。独立科学专家小组 (Scientific Panel of Independent Experts) 由《人工智能法案》第 68 条设立，由来自各国的人工智能技术专家组成，主要面向系统性风险与通用人工智能 (GPAI) 模型开展专业技术评估，并支持市场监督活动，为人工智能办公室和成员国监管机构提供科学建议。咨询论坛 (Advisory Forum) 由《人工智能法案》第 67 条设立，就通用人工智能向欧盟委员会和人工智能委员会提供咨询建议。机构代表了民间社会、学术界、工业界以及中小企业和初创企业等多元主体的相关利益，成员们就《人工智能法案》的广泛方面提供专业技术知识和建议，包括标准化过程中的潜在挑战。

3.1.2. 成员国执行：建立国家监管机构以落实市场监管和企业合规

《人工智能法案》将日常执法权限下沉至成员国层面，通过法定机构设置和跨境协作机制设计，推动欧盟统一监管规则在各成员国落地实施，但是各国监管资源和执行强度的差距也给制度带来现实挑战。

《人工智能法案》明确了各成员国必须设立的国家主管机构。法案要求各成员国设立至少一个通知机构和一个市场监督机构：通知机构 (Notifying Authority) 负责设立和监督公告机构 (Notified Body)，由公告机构执行完成高风险人工智能系统进入市场前的合规评估。市场监督机构 (Market Surveillance Authority) 有权对构成风险或不符合法案要求的人工智能系统进行干预，实施远程监控，访问企业文件、数据集和源代码，要求企业落实整改并实施处罚。欧盟要求成员国在 2025 年 8 月 2 日之前，指定负责的市场监督机构，并选择一个作为对外单一联系点，面向公众、其他成员国及欧盟相关机构开展对接。例如，意大利指定国家网络安全局 (ACN) 作为市场监督机构与欧盟单一联系点，意大利数字化署 (AgID) 担任通知主管机构，意大利央行、CONSOB 与 IVASS 在各自金融监管职责范围内，也会对金融机构使用 AI 的行为进行监管。

如何缩小各国执行差距、保障跨境人工智能服务监管环境一致，是欧盟人工智能治理体系在成员国落地执行上面临的关键考验。在目前阶段的实践过程中，各国监管资源和执行强度出现分化，截至 2026 年 9 月，27 个成员国中仅少数国家完成正式指定，部分国家错过主管机构的任命截止日期。此外，人工智能服务具备跨境流动属性，成员国监管机构的有效协作对实现统一高效监管尤为重要。欧盟官方指出，依据《人工智能法案》设立的各市场监管机关，将在欧洲人工智能委员会框架下开展工作，推进协作监管与经验交流。同时依托欧盟产品合规网络 (EUPCN) 和行政协作小组 (AdCos) 等现有工具，统一欧盟各国市场监管执行规范。

图8：欧盟人工智能治理架构示意图



数据来源：欧盟委员会官网、欧洲议会官网、EUR-Lex 欧盟法律数据库，国泰海通证券研究

3.2. 监管思路：《人工智能法案》建立以风险分级为核心的统一强制性监管约束

欧盟选择以统一立法的形式推进人工智能治理，通过《人工智能法案》建立起全球首部具有法律约束力的综合性人工智能监管框架。与美国依托自愿性标准和行业自律的治理模式不同，欧盟采取法律强制监管的路线，以风险分级为核心，对人工智能系统实施差异化合规要求，并明确了域外适用效力，试图通过内部市场规则塑造人工智能的全球治理标准。该模式在提升法律确定性和保护基本权利的同时，也面临着合规成本抬升及成员国监管资源差异等现实考验。

3.2.1. 《人工智能法案》的通过与实施

欧盟率先通过统一立法确立了人工智能治理的基本规则，试图通过欧盟层面的强制性制度减少成员国的碎片化监管，建立覆盖人工智能开发、部署和使用的共同监管标准。《人工智能法案》历经了立法磋商进程，在 2024 年内完成了表决、公示与生效的全流程。法案有利于统一欧盟内部市场规制、强化对公民权利的保护，并通过域外使用效力试图向外输出欧盟治理规范。但同时也引起企业合规成本的抬升，实施效果可能受制于各成员国监管机构建设进度的差距。

- 2024 年 3 月 13 日，欧洲议会以 523 票赞成、46 票反对、49 票弃权的表决结果通过了《人工智能法案》的最终文本，旨在保证基本权利、民主、法治和环境可持续性免受高风险人工智能的影响，同时促进人工智能的创新发展。
- 2024 年 7 月 12 日，《人工智能法案》在欧盟官方公报发布，Regulation (EU) 2024/1689 确立了欧盟统一的人工智能风险分级监管体系，同时增设通用人工智能 (GPAI) 模型专项监管规则。明确欧盟人工智能办公室、成员国监管机构权责，设置分阶段落地时间表，还修订多部欧盟既有法规，在保障基本权利的前提下兼顾人工智能创新发展。

- 2024 年 8 月 1 日,《人工智能法案》正式生效,作为全球首部全面监管人工智能的综合性法规,包含了多项核心制度设计:
 - 1) **风险分级体系**: 将人工智能系统划分为不可接受风险、高风险、有限风险(透明度风险)和低风险四个等级,按照风险等级设置差异化的合规义务;
 - 2) **域外适用效力**: 凡是投放至欧盟市场、在欧盟境内使用或对欧盟产生实际影响的各类人工智能系统与通用大模型,均受法案约束;
 - 3) **法律责任与罚款**: 被禁人工智能系统最高面临 3500 万欧元或全球年营业额的 7% 的罚款;违反高风险系统或透明度风险相关义务最高罚款 1500 万欧元或全球营业额的 3%;提供虚假、不完整或误导性信息最高 750 万欧元或全球 1% 的营业额;
 - 4) **全领域覆盖**: 法案治理范围覆盖了生物识别、医疗设备、关键基础设施、教育招聘、信贷金融、执法边境管理、司法民主程序、通用基础大模型等各类场景,对纯粹军事国防用途和科研试验用途存在部分豁免;
 - 5) **通用人工智能专章**: 第五章为通用人工智能(GPAI)模型制定专项监管规则,对具备系统性风险的模型施加额外的模型评估、风险缓释和实践报告义务。
 - 6) **分阶段实施**: 法案采用渐进式落地的战略,法案总条款、不可接受风险禁令、透明度义务、GPAI 模型相关规则、监管沙盒等相关条款分阶段逐步落实,为企业适应统一监管框架预留了调整空间。

3.2.2. 风险分级监管: 形成从禁止不可接受风险到低风险豁免的差异化监管

《人工智能法案》建立了四级风险分类体系,对不同风险等级的人工智能系统实施差异化监管,是具有创新性的制度设计。这一制度设计将监管资源集中于对基本权利和公共安全威胁最大的应用场景,同时为低风险应用保留发展空间,提升资源利用效率。

- **不可接受风险(Unacceptable Risk)**: 该类人工智能系统被认为对公民基本权利构成明显威胁,自 2025 年 2 月 2 日起在欧盟境内被全面禁止。具体包括: 利用人工智能实施具有危害性的操纵与欺诈行为;利用弱势群体弱点;社会信用评分;针对个人实施刑事犯罪风险评估或预测;无差别抓取互联网信息和监控用以扩充人脸识别数据库;工作场所和教育机构中的情绪识别;基于生物特征分类推断受法律保护的个人属性;执法机关在公共场所实施实时远程生物识别等。
- **高风险(High Risk)**: 会对民众健康、安全或基本权利造成严重风险,自 2027 年 12 月 2 日起,投放市场前须履行覆盖全生命周期的合规义务,包括建立风险评估与风险缓释机制、保障训练数据集质量、留存操作日志、编制并披露完备的技术文档、设置人工监督机制,并在欧盟数据库完成登记。适用场景包括: 关键基础设施及各类产品中的人工智能安全组件;可决定个人受教育机会与职业发展走向的评估系统;人员招聘与劳动者管理系统;信贷评分等关乎获取重要公私服务的应用;远程生物识别、情绪识别类人工智能系统;可能干预公民基本权利的执法应用;移民庇护与边境管控场景;司法审判及民主程序相关应用。
- **有限风险/透明度风险(Transparency Risk)**: 主要源于人工智能应用透明度缺失所带来的风险,法案设立信息披露义务,在必要情形下向人类作出提示告知,相关规则于 2026 年 8 月正式生效。要求用户使用聊天机器人等系统时,应当告知用户正在与机器交互;生成式人工智能提供者必须保障人工智能生成内容可识别,深度伪造、公共利益类人工智能生成文本还应当进行醒目标注。

- **最小风险/无风险 (Minimal or No risk)**：法案对此类人工智能不做专门监管要求，当前欧盟绝大多数人工智能系统都属于该类别，典型包括人工智能驱动的电子邮件、垃圾邮件过滤器等。

3.2.3. 监管对象：从高风险系统向通用人工智能拓展

欧盟人工智能治理的监管重心不仅是高风险应用，还重点关注具有系统性影响的通用基础模型，体现对人工智能技术底层风险的前瞻性监管。随着大语言模型和通用人工智能（GPAI）模型的崛起，为了防范技术扩散能力带来的系统性风险，除了下游技术应用外，更应该对上游基础模型进行专项监管。《人工智能法案》在规制高风险 AI 系统的同时，专门设立章节对通用人工智能模型实施分层监管。

- 2024 年 8 月 1 日生效的《人工智能法案》中，第五章内容对通用人工智能（GPAI）模型采取分层专项监管。1）针对全部 GPAI 模型设定普适义务，要求提供者编制并维护技术文档，向下游应用开发者提供必要合规信息；遵守欧盟版权相关规则；按照模板公开发布训练数据内容摘要；境外主体需在欧盟指定境内代表，部分开源模型可享受文档义务的有限豁免。2）针对具有系统性风险的 GPAI 模型进行额外强化监管，设立 10^{25} FLOPs 训练算力的可反驳推定阈值，超过该阈值则被推定为具有系统性风险。系统性风险模型还需完成对抗性测试等标准化模型评估、开展系统性风险识别与缓释、落实网络安全防护、跟踪并向监管机构上报重大安全事件，同时要及时向欧盟委员会完成模型通报。相关规则自 2025 年 8 月 2 日正式生效。
- 2025 年 7 月，欧盟委员会发布三份配套文件，落实《人工智能法案》第五章通用人工智能（GPAI）相关规则，以支持 GPAI 的负责任开发与部署：

1) 《通用人工智能模型提供者义务适用范围指南》：厘清了《人工智能法案》规定的 GPAI 相关义务边界，明确 GPAI 模型的判定标准、合规责任主体划分和投放欧盟市场的多种情形，同时界定开源模型部分豁免的使用条件，帮助产业链主体识别自身的合规义务。该文件不具备法律约束力，但将作为欧盟执法的重要依据。

2) 《通用人工智能实践准则》：由独立专家牵头、多利益相关方参与形成的自愿性合规工具，分为透明度、版权、安全与保障三大章节，前两章适用于全部 GPAI 模型提供者，安全与保障章节专门面向具有系统性风险的前沿模型。企业签署并落实该准则可简化合规举证，并提升法律确定性，为履行文档留存、版权合规、系统性风险评估、重大事件上报等法定义务提供可落实的实操指引。目前，亚马逊、谷歌、IBM、OpenAI 等 21 家企业已经自愿选择签署准则。

3) 《通用人工智能训练内容公开摘要模板》：作为强制性模版，要求所有投放欧盟市场的 GPAI 提供商（含满足豁免条件的开源模型）均须使用，以叙事方式概述训练数据来源与处理方式，在保护商业秘密的前提下保障版权方和数据主体等利益相关方行使法定权利。

3.3. 配套措施：推动监管要求落地，强化基本权利保护

《人工智能法案》确立统一的强制性监管框架后，欧盟通过发布一系列配套措施，推进法案实践落地的进度。一方面依托标准化研制、行业自愿倡议、数字化咨询工具与监管简化调整等措施，缩小合规要求与产业现实之间的差距，降低企业尤其是中小企业的合规负担；另一方面通过法案实体条款、国际公约及“数字一揽子法案”的程序性调整，将保护公民基本权利贯穿制度建设始终。但协调标准交付滞后、各成员国监管配套建设进度分化等问题，也给配套机制的实际实施效果带来不确定性。

3.3.1. 通过辅助合规措施与自愿倡议，降低统一监管框架的落地成本

面对《人工智能法案》复杂的合规义务以及中小企业资源有限的现实约束，欧盟通过自愿行业倡议、数字化服务平台和监管简化改革，推进统一监管框架的执行落地。这些支撑措施旨在通过指导企业合规流程，降低法案监管框架的落地成本。但是自愿性措施的覆盖范围，依旧是监管机构需要持续关注的重点。

- 2024 年 9 月 25 日，欧盟委员会公布《人工智能公约》（AIPact）首批 100 多家签署方名单。该自愿性倡议旨在帮助市场主体平稳适配全新监管体系，为法案后续落地实施提供支撑，并邀请人工智能服务提供方与部署方提前履行法案核心义务。公约要求参与企业承诺至少三项核心行动：采用人工智能治理策略以推进未来合规；识别和映射可能被归类为高风险的系统；提升员工人工智能意识与素养。超过半数签署方还额外承诺了人工监督、风险缓释和深度伪造内容透明度等附加义务。截至 2025 年 12 月，已有 230 多家公司签署公约，包括来自欧盟和其他地区的大公司及中小企业（SMEs），涵盖 IT、电信、医疗保健、银行、汽车和航空等领域。
- 2025 年 10 月，欧盟委员会正式推出人工智能法案服务台（AI Act Service Desk）和人工智能法规单一信息平台（AI Single Information Platform）。服务台由欧盟人工智能办公室专门团队运营，支持以所有欧盟官方语言提交问询，为利益相关方特别是中小型人工智能提供者和部署者提供免费信息与指导；单一信息平台提供合规检查器、人工智能法案浏览器等互动工具，帮助用户判断自身法律义务。
- 2025 年 11 月 19 日，欧盟委员会提出“数字一揽子”改革方案，其中包含对《人工智能法案》进行简化的《人工智能数字综合修订法案》（Digital Omnibus on AI，2026 年 7 月正式生效）。该提案推出一系列针对性简化监管措施，确保法案部分条款能够及时、顺畅、合乎比例原则地落地执行。核心内容包括：将高风险人工智能相关监管规则的实施时限与配套标准完备进度相绑定；将中小企业（SMEs）可享受的监管简化政策拓展至小型中型市值企业（SMCs），包括简化技术文件编制要求、处罚裁量时予以从轻考量；取消强制制定统一上市后监测方案的规定；扩大人工智能监管沙盒与实景测试的适用范围等。

3.3.2. 研制并采纳协调标准，构建技术合规基准

协调标准（Harmonised Standards）是《人工智能法案》合规推定机制的技术基石，根据法案第 40 条，符合协调标准的人工智能系统可被推定为满足相关基本权利与安全保障要求。欧盟委员会既依托标准化机构研制分领域合规技术标准，也通过专项行动计划培育前沿模型评估、安全测试等监管能力，为法案落地建立完整的技术合规体系。然而，受标准编制周期、多方共识协调难度的影响，截至 2026 年 7 月仍无协调标准在《欧盟官方公报》完成引用，技术层面的标准化制度供给略微滞后于立法文本。

- 2023 年 5 月 22 日，欧盟委员会通过 Commission Implementing Decision C(2023) 3215 号实施决定，向欧洲标准化委员会（CEN）和欧洲电工标准化委员会（CENELEC）下达 M/593 标准化委托请求。围绕风险管理、数据治理、记录保存、透明度、人工监督、准确性、鲁棒性、网络安全、质量管理和合格评定十大优先领域开展标准编制，以研制支撑《人工智能法案》落地的欧洲协调标准。原定交付期限为 2025 年 4 月 30 日，标准制定需兼顾安全保障、基本权利保护与产业创新。

- 2025 年 6 月 23 日，由于 CEN 和 CENELEC 未能按原定期限完成交付，欧盟委员会通过了修订版 Commission Implementing Decision C (2025) 3871，提出 M/613 标准化请求，废除原 C (2023) 3215，将交付期限延长至 2027 年 2 月 28 日，并要求制定机构每三个月提交进度报告。
- 2025 年 10 月 14 日至 16 日，CEN 与 CENELEC 技术委员会联系会议通过一揽子特别加速措施，以加快交付人工智能关键标准的制定，及标准化请求 M/593 及修正案 M/613 要求交付的成果。这项特殊决定反映了标准化机构在按时交付与包容性、透明度、共识原则之间寻求平衡的责任。
- 2026 年 3 月 13 日，CEN 正式批准 EN ISO/IEC 42001:2026，将国际标准化组织发布的《人工智能管理体系》(ISO/IEC 42001:2023) 未经修改地采纳为欧洲标准。但该标准并非在 M/593 或 M/613 框架下制定，因此不构成人工智能法案下的协调标准，无法产生合规推定效力，其主要功能在于为组织建立人工智能管理体系提供国际认可的方法论框架，而非直接服务于监管合规。
- 2026 年 6 月，CEN-CENELEC JTC 21 多项关键协调标准取得阶段性进展：EN 18286 (人工智能质量管理体系) 进入批准阶段，是十项优先交付物中进度最快的一项；prEN 18228 (人工智能风险管理)、prEN 18229-1 (记录保存) 以及 prEN 18282 (人工智能网络安全规范) 均进入公开征求意见阶段。然而，截至 2026 年 6 月，尚无协调标准在《欧盟官方公报》上被正式引用。
- 2026 年 7 月 7 日，欧盟委员会发布《网络安全与人工智能行动计划》，旨在应对前沿人工智能模型带来的新型网络安全挑战，利用人工智能技术增强欧盟数字防御能力。该计划围绕三大目标推进：保障先进人工智能模型进入欧盟市场前完成严格的风险与安全评估，强化成员国、企业与公共机构对人工智能驱动网络攻击的防御能力，培育欧洲本土人工智能网络安全技术能力以提升漏洞检测与威胁响应速度。计划 2027 年建立欧盟人工智能模型第三方独立评估能力，联合欧洲网络安全局制定先进人工智能系统安全访问的标准指南，并依托专业机构建设覆盖能源、交通、医疗、金融等关键领域的人工智能网络安全测试空间。

3.3.3. 强调保护基本权利，将隐私、公平和非歧视纳入人工智能治理

欧盟通过国内立法导向、算法偏见矫正、人工智能素养要求和国际公约约束等多层机制，将保护公民基本权利作为人工智能治理的价值导向。这些措施充分展示出欧盟对于保护公民基本权利的重视，努力寻求健康、安全发展与技术创新的平衡。

- 2024 年 5 月 17 日，欧洲委员会部长委员会正式通过《欧洲委员会人工智能与人权、民主及法治框架公约》(CETS No.225)，是全球首部在人工智能领域具有法律约束力的国际条约。核心目的在于确保人工智能系统全生命周期均需尊重人权、巩固民主并维护法治，涵盖了公共部门与私营部门，要求各签署国对人工智能风险进行管理，保护公民免受算法歧视和隐私侵犯。
- 2024 年 6 月 13 日，欧洲议会和欧盟理事会签署《人工智能法案》，法案序言及条款明确将“完善内部市场，发展以人为本的可信人工智能，对健康、安全、《欧盟基本权利宪章》所规定的基本权利（民主、法治、环境保护）提供高水平保护，抵御人工智能带来的损害，支持创新”作为立法宗旨之一。
- 2024 年 8 月 1 日，《人工智能法案》正式生效。法案第 4 条要求人工智能系统提供商与部署方确保其员工及代为操作系统的人员具备合格的人工智能素养，综合考量技术知识、从业经验、教育背景、培训经历及系统应用场景。第 27 条创设“基本权利影响评估”(FRIA) 制度，要求公共机构、提供公共

服务的私营实体以及部署信贷评分、人寿与健康保险风险评估等高风险系统的实体，在首次部署前评估系统对基本权利的影响，并将结果通知市场监督机构。

- 2025 年 11 月 19 日，“数字一揽子”提案在基本权利保护方面作出调整，委员会称其不改变原法案对基本权利的实质保护范围。修订第 4 条，将人工智能素养的强制性普及义务从私营主体转移至欧盟委员会与各成员国承担，但明确高风险人工智能适用主体仍需履行人员培训义务；新增 4a 条，人工智能系统与模型的提供商和部署者，在合成或匿名数据无法完成偏见检测与纠正的前提下，在落实访问管控、禁止对外传输、处理完毕删除数据等多重严格保障措施后，可处理种族、健康等特殊类别个人数据，目的仅限于检测与修正算法歧视风险，不能用于模型预训练或普通业务场景；修订第 77 条，强化各成员国基本权利保护主管机构调取人工智能监管材料、跨机构协作的权限。
- 2026 年 5 月 15 日，在欧洲委员会部长委员会第 135 届会议期间，欧盟交存《欧洲委员会人工智能与人权、民主及法治框架公约》批准书。欧盟成为该公约的重要缔约方，标志着欧盟将人工智能基本权利保护从区域立法进一步嵌入国际法框架，推动全球范围内形成尊重人权、民主和法治的人工智能治理共识。

图9：欧盟人工智能治理体系特征（1）



数据来源：欧盟委员会官网、EUR-Lex 欧盟法律数据库、欧盟委员会数字战略司、欧洲议会官网、欧盟委员会文件登记库、CEN-CENELEC 欧洲标准化组织，国泰海通证券研究

3.4. 监管与创新并重：通过监管沙盒和产业政策缓解高合规要求对创新的约束

欧盟在构建严格的人工智能强制性监管框架的同时，也逐步意识到高合规成本可能抑制技术创新及中小企业的发展，因此通过可控实验和产业政策来缓解法案落地带来的压力。欧盟通过监管沙盒与测试设施为企业提供低成本试错环境，依托一揽子产业政策强化技术主权与产业竞争力，并在算力、数据和人才等基础要素上加大战略投入，实现从监管优先到将创新纳入考量的转变。然而，目前欧盟的基础模型和算力依旧较大程度上依赖非欧盟企业，技术主权的进一步培育仍需欧盟及成员国的协调配合。

3.4.1. 设置监管沙盒与测试设施，以可控试验降低企业创新的合规成本

面对高强度监管对人工智能技术发展的影响，欧盟通过监管沙盒、测试与试验设施等机制，为企业提供受控的低成本试错环境，以平衡监管要求和技术创新。这些机制允许企业在有限时间内，在监管机构的监督下，对创新人工智能系统进行开发、训练、测试和验证，为企业在正式承担完整合规义务前提供缓冲空间。但从成员国落地情况看，受到准入门槛和资源协调效率影响，沙盒建设进度呈现明显分化，且完全自愿参与对企业激励有限，成为制约政策落地效果的关键因素。

- 2023 年 6 月 27 日，欧盟联合成员国及 128 个合作伙伴正式启动了第一批总投资达 2.2 亿欧元的四大领域行业人工智能测试与试验设施（Testing and Experimentation Facilities, TEF）项目。该项目是“数字欧洲计划”下最具落地价值的实体项目，为人工智能开发者，尤其是中小企业和初创公司，提供虚拟与物理相结合的真实世界模拟环境，覆盖医疗健康、农业与食品、先进制造、智慧城市与社区四大垂直行业，帮助企业省去昂贵的自建测试成本，并在接近真实应用的条件下验证系统性能与合规性。
- 2024 年 8 月生效的《人工智能法案》第 57 条最初明确要求各成员国主管机关在 2026 年 8 月 2 日之前建立至少一处人工智能监管沙盒（AI Regulatory Sandbox），也可与其他成员国联合建立。根据 Regulation (EU) 2026/1744 的修订，该截止日期已推迟至 2027 年 8 月 2 日。监管沙盒的作用是提供一个受控环境，企业在主管机关的监督下于有限时间内开展试验。在创新人工智能系统投放市场或提供服务之前，为其开发、训练、测试和验证提供便利。
- 2025 年 4 月 3 日，西班牙数字化转型与公共职能部正式宣布推出欧盟首个人工智能监管沙盒，由其下属的数字化与人工智能国务秘书处负责管理，遴选 12 套高风险人工智能系统进入沙盒试点。试点完成后将输出技术实施指南，为西班牙企业落实《人工智能法案》合规要求提供参考。
- 2025 年 5 月 9 日，德国联邦网络管理局（Bundesnetzagentur）联合黑森州数字化与创新部及联邦数据保护专员启动人工智能监管沙盒模拟试点项目。该项目旨在模拟《人工智能法案》框架下监管沙盒的关键要求、流程与挑战，其最终报告于 2026 年 3 月 17 日发布，为未来真实沙盒的建立与运营提供实践指导。同期，德国参与 EUSAIr（欧盟人工智能监管沙盒协调与支持行动）试点项目，该项目由“数字欧洲计划”资助，运行周期为 2025 年 10 月至 2026 年 6 月，支持欧盟范围内人工智能监管沙盒的推广与协调。
- 截至 2025 年 8 月，欧盟议会研究服务局（EPRS）宣布在 27 个成员国中，仅西班牙一国拥有运行中的人工智能监管沙盒；5 个国家正在积极实施，4 个国家已宣布意向，而多达 16 个国家尚未公开详细实施方案，成员国间就监管沙盒的实施进度出现显著分化。

- 2025 年 11 月 19 日，欧盟委员会提出“数字一揽子法案”提案，推出一系列针对性简化监管措施。明确扩大人工智能监管沙盒与实景测试的适用范围；AI Office 设立的欧盟级统一沙盒于 2028 年全面运营。该法案还将大量依托通用人工智能模型搭建或内嵌于超大型在线平台、超大型搜索引擎的人工智能系统，统一交由欧盟人工智能办公室集中监管，进一步降低企业在不同成员国间应对差异化沙盒规则的协调成本。

3.4.2. 产业政策导向从监管优先，转向兼顾强化技术主权和产业竞争力

欧盟人工智能的治理思路不仅强调监管优先，也逐步开始重视规则制定与产业竞争力的协同发展，发布产业政策以平衡监管与创新。目前全球人工智能竞赛加剧，而欧盟在基础模型和算力领域对外部的技术依赖加深。欧盟自 2024 年起密集推出一系列产业政策，通过公共资金引导、应用场景开放和科研生态建设，提升本土人工智能的创新活力。

- 2024 年 1 月 24 日，欧盟委员会推出《人工智能创新一揽子计划》（AI Innovation Package），其中包括欧洲生成式人工智能计划（GenAI4EU），为生成式人工智能领域的研发创新提供资金支持。该计划聚焦于应用研究，采用分行业推进模式，通过“地平线欧洲”（Horizon Europe）与“数字欧洲计划”（Digital Europe Programme）提供财政支持，用于支持各行业先进人工智能模型与解决方案的研发，覆盖制造业生产线优化、机器人自主能力与人机协作、网络防御及医学影像诊疗等领域，旨在打造强大的欧洲人工智能生态体系。
- 2025 年 2 月 11 日，欧盟委员会主席 Ursula von der Leyen 在巴黎举行的人工智能行动分会上正式启动 InvestAI 倡议，计划撬动 2000 亿欧元公私投资用于人工智能领域，其中包括 200 亿欧元专项投资用于建设 5 座人工智能超级工厂（AI Gigafactories）。InvestAI 被欧盟定位为人工智能领域的欧洲核子研究中心（CERN），旨在通过公共资金杠杆效应吸引私人资本，建设跨境共享基础设施。
- 2025 年 4 月 9 日，欧盟委员会发布《人工智能大陆行动计划》（AI Continent Action Plan），系统阐述如何将欧盟的人才储备与雄厚传统产业转化为人工智能发展助推引擎。该计划核心举措包括：推出 InvestAI 专项基金，计划撬动 2000 亿欧元欧盟境内人工智能产业投资；在欧洲布局人工智能工厂（AI Factories），利用 EuroHPC 超级计算中心建设欧洲人工智能训练基础设施，并建设至多 5 座人工智能超级算力工厂；在各人工智能工厂内设数据实验室，归集整理多来源高质量数据集；同时提出简化《人工智能法案》落地执行流程，编制禁止类人工智能应用指引并推动行业自律准则。
- 2025 年 10 月 8 日，欧盟委员会发布《应用人工智能战略》（Apply AI Strategy），旨在提升重点战略产业竞争力、强化欧盟技术自主可控能力，推动全欧洲普及人工智能应用与创新，重点帮扶中小微企业落地人工智能。该战略倡导“人工智能优先”（AI-first）政策，要求各类机构在制定战略和出台政策时将人工智能作为备选解决方案，同时审慎权衡收益与风险；提倡公共领域“采购欧洲本土技术”，并优先扶持开源人工智能方案。计划利用“地平线欧洲”和“数字欧洲计划”资助具体人工智能用例，预计 10 亿欧元公共投资将吸引成员国和私营部门共同融资。战略还将原有欧洲数字创新中心（EDIH）升级为“人工智能体验中心”（Experience Centres for AI），作为对接欧盟整体人工智能创新生态的统一入口。
- 2025 年 10 月 8 日，欧盟委员会同步发布《欧洲科学人工智能战略》（European Strategy for Artificial Intelligence in Science），旨在依托人工智能赋能欧洲科研体系，夯实前沿人工智能基础研究能力，兼顾“用人工智能发展科学、用科学支撑人工智能”的双向目标。该战略推出 RAISE（Resource for AI Science in

Europe) 欧洲人工智能科学资源平台试点项目，整合欧盟资金、超算算力、高质量数据集与科研人才四大战略资源，补齐此前人工智能产业政策在基础科研层面的短板。RAISE 于 2025 年 11 月在哥本哈根人工智能科学峰会上正式启动。

3.4.3. 支撑算力、数据、科研和人才发展，强化欧盟技术主权发展的产业基础

为了支撑产业政策落地并且缓解对外部技术的依赖，欧盟从算力基础设施、数据资源供给、科研平台建设和人才培养体系等维度，为人工智能产业发展提供基础保障。这些举措体现了欧盟对于人工智能技术发展的重视，试图通过政策支持和大规模公共投资弥补市场在算力基础和科研设施上的投入不足。InvestAI 与 EuroHPC 协同推进人工智能工厂网络的形成，《云与人工智能发展法案》和《芯片法案 2.0》等文件则从立法层面为基础设施扩张提供制度保障。

在算力基础设施方面，《人工智能创新一揽子计划》和《人工智能大陆行动计划》将建设人工智能工厂列为战略优先事项。目前，EuroHPC JU 正在筹建 19 座人工智能工厂以及 13 个人工智能工厂节点，面向初创企业、中小企业、科研机构开放资源，打造前沿人工智能模型与应用。欧盟将在全欧采购并部署至少 9 台面向人工智能优化的超级计算机，预计将使 EuroHPC 当前人工智能计算能力提升两倍以上。此外，欧盟在 InvestAI 框架下先提出最多 5 座人工智能超级工厂，2026 年 7 月启动的 EuroHPC 招标将目标扩大到最多 7 座，配置超十万个先进人工智能处理器，专门用于训练开发拥有万亿级参数的下一代人工智能模型，高度重视供电保障、稳定供应链、高速网络、能效优化与 AI 自动化技术。

- 2025 年 2 月，InvestAI 倡议承诺投入 200 亿欧元，用于在欧盟境内建设人工智能超级工厂。同年 12 月，欧洲投资银行集团（EIB）与欧盟签署谅解备忘录（MoU），将研究配套贷款方案，对倡议赠款形成补充，撬动私人资本，为欧洲初创企业、科研人员及产业界打造完备的人工智能基础设施。
- 2026 年 6 月 3 日，欧盟委员会推出“欧洲技术主权一揽子方案”，旨在提升欧盟在半导体、人工智能、云计算及开源技术领域的产业实力，助力欧洲在人工智能领域强化数字自主能力，其中包括《芯片法案 2.0》（Chips Act 2.0）和《云与人工智能发展法案》（Cloud and AI Development Act, CADA）。
- 面对在先进芯片制造和半导体设计等关键领域，欧盟仍依赖第三国的现实，《芯片法案 2.0》提出加强当前欧洲在主流芯片等领域的优势，并培育尖端半导体技术研发能力。2.0 版将重点转向危机应对能力，完善供应链监测预警与危机处置机制，并建立集中式采购系统，减少欧盟在芯片领域的战略依赖性和供应链脆弱性。
- 《云与人工智能发展法案》聚焦研究与创新、算力供给能力和技术自主权三大领域，目标在未来 5 至 7 年内将欧盟数据中心容量至少增加两倍，引入四级云主权框架，简化数据中心许可和部署流程，并要求改善能源、土地和融资等关键资源的获取。法案同时要求公共部门采购优先考虑符合主权标准的云服务，旨在减少对外部超大规模云服务商的依赖。

在数据资源方面，《人工智能大陆行动计划》提出以“数据联盟战略”（Data Union Strategy）构建一体化欧盟内部数据市场，支撑全欧盟人工智能研发规模化推进。作为该战略的具体落地举措，欧盟在各人工智能工厂内设立数据实验室（Data Labs），负责归集、整理多来源高质量数据集，并链接至共同欧洲数据空间（如欧洲健康数据空间），将碎片化数据源转化为可信且可访问的人工智能开发资源，降低企业在数据获取与预处理环节的成本。

在科研支持方面，RAISE 欧洲人工智能科学资源平台作为《欧洲科学人工智能战略》的核心载体，通过整合分散的科研资源，为欧洲科研人员提供统一的算力、数据、人才和研究资金申请入口。RAISE 试点由“地平线欧洲”专项资助，2025 年工作预算 3300 万欧元，2026 至 2027 年工作预算 1.07 亿欧元。此外，地平线欧洲计划向人工智能工厂投入 6 亿欧元，确保研究人员和初创企业获得专用计算资源。欧洲科学人工智能（SCIENCE）项目为 RAISE 的建设提供支撑，联合欧洲顶尖科研机构、科研基础设施、人工智能研究中心与电子基础设施，明确人工智能领域的优先创新方向，并试点建立 RAISE 的科学人工智能秘书处。

在人才培养方面，欧盟以《人工智能大陆行动计划》与《应用人工智能战略》为核心，配合《技能联盟》相关配套文件，将本土人才培养、人才留存与海外人才引进相结合，强调全民人工智能素养和劳动者相关技能的提升，为人工智能发展提供基础保障。

- 欧盟持续扩大本土人工智能专业人才供给，《人工智能大陆行动计划》提出扩大人工智能方向本科、硕士、博士学位招生规模，成立生成式人工智能学院（GenAI Academy），试点生成式人工智能学位项目与人工智能学徒项目，推出奖学金、重返职场计划，提升女性在人工智能领域的参与度；《应用人工智能战略》面向重点数字行业推出“商业人工智能”跨学科教育项目，培养能源、交通等行业的垂直领域人工智能专家。设立人工智能企业家实验室，为潜力毕业生匹配行业导师，促进知识交流并孵化人工智能创业项目。
- 为留住本土人才并吸引境外人工智能人才，《人工智能大陆行动计划》设立人工智能研究员资助项目，鼓励欧盟本土人才留在欧盟发展，同时支持境外高水平博士候选人和青年从业者到欧盟机构开展研究工作。综合运用人才储备计划、多用途法律通道办公室和玛丽·斯克洛多夫斯卡-居里行动（MSCA），面向全球引进科研人员等人工智能高端人才。
- 欧盟还面向全社会劳动者开展人工智能技能培训，《人工智能大陆行动计划》依托欧洲数字创新中心（EDIH）扩充人工智能技能培训服务，针对不同行业、岗位开发定制化人工智能素养培训。《应用人工智能战略》将培训资源打通生成式人工智能学院、人工智能工厂、欧洲数字创新中心、欧洲技术与创新共同体、互操作欧洲学院等多个平台。此外，欧盟还联合技能公约（Pact for Skills）发动产业力量参与劳动者技能升级，对受产业变革冲击、存在岗位流失风险的劳动者落实技能保障。

3.5. 国际影响：通过规则外溢和价值引领扩大欧盟治理话语权

欧盟在全球人工智能治理中的影响力主要来源于其规则制定能力和内部统一市场的规模优势，以权利保护为价值内核，并通过多边治理协作提升技术竞争力。欧盟依托《人工智能法案》的域外使用效力将内部监管规则转化为全球性合规基准，并通过主导国际公约、参与多边机制和开展数字外交，持续向外国传递“以人为本”的治理理念。在借助这种规则输出模式提升全球治理话语权的同时，欧盟也需要积极应对自身产业技术支撑不足带来的现实挑战。

3.5.1. 借助统一市场和规则输出，实现人工智能监管标准和价值观念外溢

欧盟凭借单一市场的规模优势提升在全球规则制定中的话语权，通过强制性立法监管、国际公约约束和自愿性承诺，推动治理标准和价值理念的跨境外溢。《人工智能法案》的域外适用效力，对全球产业链进行合规规则的传导。欧盟委员会人工智能框架公约和《欧盟国际数字战略》等国际性文件向全球推广以人权、民主与法治和“以人为本”为导向的治理观念，同时通过实践准则、行业公约等自愿性工具，在正式规则落地前提前引导全球企业适配欧盟治理逻辑。这一治理模式延续了欧盟数字治理的“布鲁塞尔效应”，但其合规要求也引发了对于平衡创

新和监管问题的关注。

- 2024 年 8 月 1 日,《人工智能法案》正式生效,其域外适用效力条款同步确立。法案规定,无论企业总部位于何处,只要将人工智能系统投放至欧盟市场、在欧盟境内使用或对欧盟用户产生实际影响,即须全面遵守法案规定的风险分级、透明度与合规义务。这一条款使得欧盟得以将其内部监管标准直接适用于全球人工智能产业链,迫使跨国企业为维持欧盟市场准入而在全球范围内统一调整产品设计、数据治理与合规流程。
- 2024 年 9 月 5 日,《欧洲委员会人工智能与人权、民主及法治框架公约》(CETS No.225)在维尔纽斯正式开放签署,美国、英国、以色列、欧盟等在当天签署。该公约是全球首部在人工智能领域具有法律约束力的国际条约,欧盟通过主导该公约的制定与开放签署,将其“以人为本”的人工智能治理价值观从区域立法升级为国际法框架,推动全球主要经济体在基本权利保护层面接受欧盟倡导的规则逻辑。
- 2025 年 6 月 5 日,欧盟委员会发布《欧盟国际数字战略》联合公报(JOIN(2025) 140 final),欧盟将依托规模庞大且一体化的数字单一市场,为伙伴国家提供技术解决方案。持续推广基于自身价值观的数字治理和数字标准路径,塑造以人为本、值得信赖、尊重人权与基本自由的全球数字转型。
- 2025 年 7 月 10 日,欧盟人工智能办公室发布《通用人工智能实践准则》,为 GPAI 模型提供者提供透明度、版权合规与安全治理的自愿性合规路径。该准则不具有法律约束力,但签署者可以获得合规推定以降低行政负担。截至 2025 年 8 月,亚马逊、谷歌、IBM、OpenAI 等 21 家主要前沿模型开发商已签署全部章节;xAI 仅签署安全与安保章节,拒绝接受透明度与版权义务。这一签署格局表明,欧盟正通过合规激励机制,使得全球科技企业在法定义务完全生效前主动接受欧盟规则约束。
- 2025 年 8 月 12 日,OpenAI 致信加利福尼亚州州长,呼吁在前沿人工智能监管中与联邦及全球标准协调一致,避免州级要求与联邦政府及民主盟友已推进的安全框架出现不一致。信中明确建议,州政府可将签署欧盟《通用人工智能实践准则》等框架的前沿模型开发者视为已满足州级合规要求,体现出欧洲自愿性准则向全球的扩散。

3.5.2. 依托多边机制和数字外交,拓展人工智能的技术竞争力和全球影响力

欧盟借助多边治理平台和双边伙伴合作机制,同时兼顾规则价值输出和技术能力合作,提升在全球治理体系中的技术自主权及产业竞争力。欧盟深度参与 G7 广岛进程、OECD 治理框架、人工智能安全研究所国际网络等多边合作机制,推动国际治理共识与自身规则体系对齐;并与多国建立数字伙伴关系,在供应链安全、开源技术、基础设施建设等领域开展务实合作。同时,依托“全球门户”战略对外提供算力基建与能力援助,减少对单一国家的技术依赖,弥补其在基础模型和算力产业上的相对短板。

- 2022 年 5 月至 2023 年 11 月期间,欧盟先后与日本、韩国、新加坡和加拿大建立数字伙伴关系,将人工智能治理、网络安全、半导体等领域纳入双边部长级合作框架,有助于欧盟解决外国技术依赖和网络安全威胁等漏洞。欧盟借助合作关系促进伙伴国家以人为本的数字化转型,保证隐私、问责制、透明度和包容性等基本权利。通过联合开展研究项目、互相访问基础设施、签署数字贸易协定等方式,支持人工智能和量子技术等领域的进步,提高合作双方竞争地位。

- 2024 年 11 月 20 日，欧盟人工智能办公室作为创始成员参与在旧金山举行的人工智能安全研究所国际网络（International Network of AI Safety Institutes）首次会议。该网络基于 2024 年 5 月首尔 AI 峰会期间发布的《首尔意向声明》建立，旨在通过研究、测试、指导和包容性合作推进 AI 安全科学，促进政府与社会更好地理解先进 AI 系统的风险并提出缓释方案。
- 2024 年 12 月 3 日，欧盟与智慧非洲（Smart Africa）联盟签署行政安排和赠款协议，在“全球门户”的战略框架下深化非洲数字转型合作。协议主要目标包括部署安全可负担的数字网络、推进电子政务服务以及探索人工智能潜力。双方共同推出了数据治理旗舰项目，旨在建立非洲自主的数据治理框架，并帮助非洲加强数字基础设施建设和技能发展。
- 2025 年 2 月 7 日，经 G7 委托、由 OECD 秘书处开发的《广岛人工智能进程报告框架》正式启动，以 G7 广岛进程《先进人工智能系统开发组织国际行为准则》为基础，旨在促进开发先进人工智能系统的透明度和问责制，欧盟作为 G7 核心成员深度参与了该框架的设计与推广。2025 年 4 月，首批 19 家前沿 AI 企业（包括 Anthropic、OpenAI、谷歌、微软等）通过 OECD 平台公开发布广岛 AI 进程报告，披露其在风险识别、信息安全、透明度治理、内容溯源等方面的实践，提出人权、法治、以人为本的价值观念，体现出欧盟监管范式与国际要求的协同，借助 G7 这一多边机制拓展其国际影响力。
- 2025 年 6 月 5 日，欧盟委员会发布《欧盟国际数字战略》联合公报，承认数字和人工智能革命需要与国际伙伴采取集体行动，旨在通过数字外交提升全球技术竞争力。提出借助联合国、G7 等现有多边平台开展人工智能治理磋商，与理念相同的伙伴国家开展务实合作，共同推进跨境人工智能研发项目合作及技术标准共同制定。不断夯实欧盟关键基础技术能力，吸引对欧盟投资，推动全球构建可信人工智能公平竞争环境。明确将对外输出欧盟人工智能相关技术工具与解决方案，联合伙伴开展人工智能能力建设项目，向发展中国家提供人工智能相关技术援助与能力培训。
- 2025 年 10 月 31 日，欧洲发展政策管理中心（ECDPM）举办首届欧盟—印度人工智能合作高级别对话会，对接欧盟-印度贸易与技术委员会（TTC）机制，邀请印度电子和信息技术部、欧盟委员会人工智能办公室、企业与智库代表参会，围绕“全栈式”技术合作开展交流，重点探讨半导体及人工智能供应链去风险、开源人工智能协作、数字主权规则协调，搭建多利益相关方专家对话社群，推动欧印技术政策的务实合作。

图10：欧盟人工智能治理体系特征（2）



数据来源：欧盟委员会官网、EUR-Lex 欧盟法律数据库、欧盟委员会数字战略司、欧盟研发创新总司、OpenAI 官网、欧洲议会智库 EPRS、欧洲发展政策管理中心 ECDPM，国泰海通证券研究

4. 中国人工智能治理体系

4.1. 整体统筹：政府主导、多部门协同推进人工智能治理体系建设

中国人工智能治理体系以国家整体战略规划为引领，通过中央层面的战略部署、国务院部门协调统筹以及各部门的具体执行落地，形成多层次的治理结构。中国通过中长期发展规划、专项政策文件和行政法规，建立起集中统一的国家战略体系。这有助于充分调动政府、产业和科研资源，在人工智能技术快速演进阶段提高政策响应速度和资源配置效率。

4.1.1. 强化顶层战略引导，推动人工智能发展方向的统一

中国通过制定国家战略规划和发布重点政策文件，将人工智能纳入科技创新和产业发展的重要方向，使其拥有“国之重器”的战略定位。我国坚持“以人为本、智能向善”的治理理念，《“十五五”规划纲要》将人工智能定位为驱动新质生产力发展的核心引擎。通过将人工智能纳入国家发展战略，稳定产业发展的未来预期，引导社会资源向人工智能领域聚集，推动技术创新与商业落地的协同发展。

- 2017 年 7 月 8 日，国务院印发《新一代人工智能发展规划》，提出“三步走”战略目标：到 2020 年与世界先进水平同步，到 2025 年部分领域领先，到 2030 年成为世界主要人工智能创新中心。《规划》系统部署了六大重点任务，包括构建开放协同的科技创新体系、培育高端高效的智能经济、建设安全便捷的智能社会、加强军民融合、构建智能化基础设施体系、前瞻布局重大科技项目。

- 2021年3月13日,《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》将人工智能列为前沿科技攻关领域,聚焦关键算法研发突破,培育壮大人工智能等新兴数字产业。同年12月12日,《“十四五”数字经济发展规划》进一步将人工智能作为数字经济重点产业,强调其与实体经济的深度融合。
- 2023年2月27日,中共中央、国务院印发《数字中国建设整体布局规划》,按照“2522”整体框架布局,包括夯实数字基础设施和数据资源体系“两大基础”,推进数字技术与经济、政治、文化、社会、生态文明建设“五位一体”融合,强化数字技术创新体系和数字安全屏障“两大能力”,优化国内国际“两个环境”,为人工智能发展提供底层架构支撑。
- 2025年4月25日,习近平总书记在二十届中央政治局第二十次集体学习时强调,要“全面推进人工智能科技创新、产业发展和赋能应用,完善人工智能监管体制机制,牢牢掌握人工智能发展和治理主动权”,提出要充分发挥新型举国体制优势,坚持自立自强,突出应用导向,推动我国人工智能朝着有益、安全、公平方向健康有序发展。本次讲话明确人工智能在国家战略中的核心地位,确立了发展与治理并重的总体方针。
- 2025年5月,《全国人大常委会2025年度立法工作计划》将“人工智能健康发展等方面的立法项目”列为预备审议项目,由有关方面抓紧开展调研和起草工作,视情安排审议。《国务院2025年度立法工作计划》强调“推进人工智能健康发展立法工作”。
- 2025年8月21日,国务院印发《关于深入实施“人工智能+”行动的意见》,作为人工智能领域战略性顶层设计文件,提出“三步走”发展目标:到2027年智能终端/智能体普及率超70%;到2030年超90%,智能经济成为重要增长极;到2035年全面步入智能经济和智能社会新阶段。文件部署“人工智能+”科学技术、产业发展、消费提质、民生福祉、治理能力、全球合作六大重点行动,涵盖从基础研发到产业应用再到全球治理的完整战略链条。
- 2026年3月13日,《中华人民共和国国民经济和社会发展第十五个五年规划纲要》将人工智能定位为驱动新质生产力发展的核心引擎,围绕技术创新、产业融合、安全治理等维度作出系统性部署,明确提出以“牢牢掌握发展主动权”为战略目标。纲要提出五大工作思路:要求健全高等教育学科专业设置调整机制、提升数智化发展水平、促进模型算法迭代创新、完善数据标准体系和质量管理体系等,标志着人工智能已从一般性技术议题,升级为战略性议题。
- 2026年5月11日,国务院办公厅印发《国务院2026年度立法工作计划》,明确提出“完善人工智能治理,加快推进人工智能健康发展综合性立法”,这是我国首次在年度立法计划中明确提出制定人工智能综合性法律。

4.1.2. 多部门协同治理,共同推动人工智能的政策协调与落地实施

中国人工智能治理呈现党中央集中领导、国务院统筹部署、科技部门牵头推动、行业主管部门分工监管、专业机构提供技术支持的多层治理架构。多部门参与覆盖了人工智能产业发展的不同环节,有助于实现从科技创新到产业应用、安全监管的综合治理。同时,也需重视部门间协调机制的建设,提高治理规则的衔接效率。

在顶层统筹层面，党中央负责总体把握人工智能发展的战略方向和治理原则，中央科技委员会和中央网络安全和信息化委员会参与战略协调工作，共同支撑人工智能发展的顶层决策。党中央主要通过重大会议部署、中央政治局集体学习等方式，对人工智能发展的战略定位、技术创新方向和安全治理要求进行顶层设计。其中，中央科技委员会负责加强党中央对科技工作的集中统一领导，统筹国家科技创新战略和重大科技任务；中央网络安全和信息化委员会办公室承担总体布局、统筹协调和督促落实等职责，在算法治理、数据安全、生成式人工智能服务监管等领域发挥重要作用。

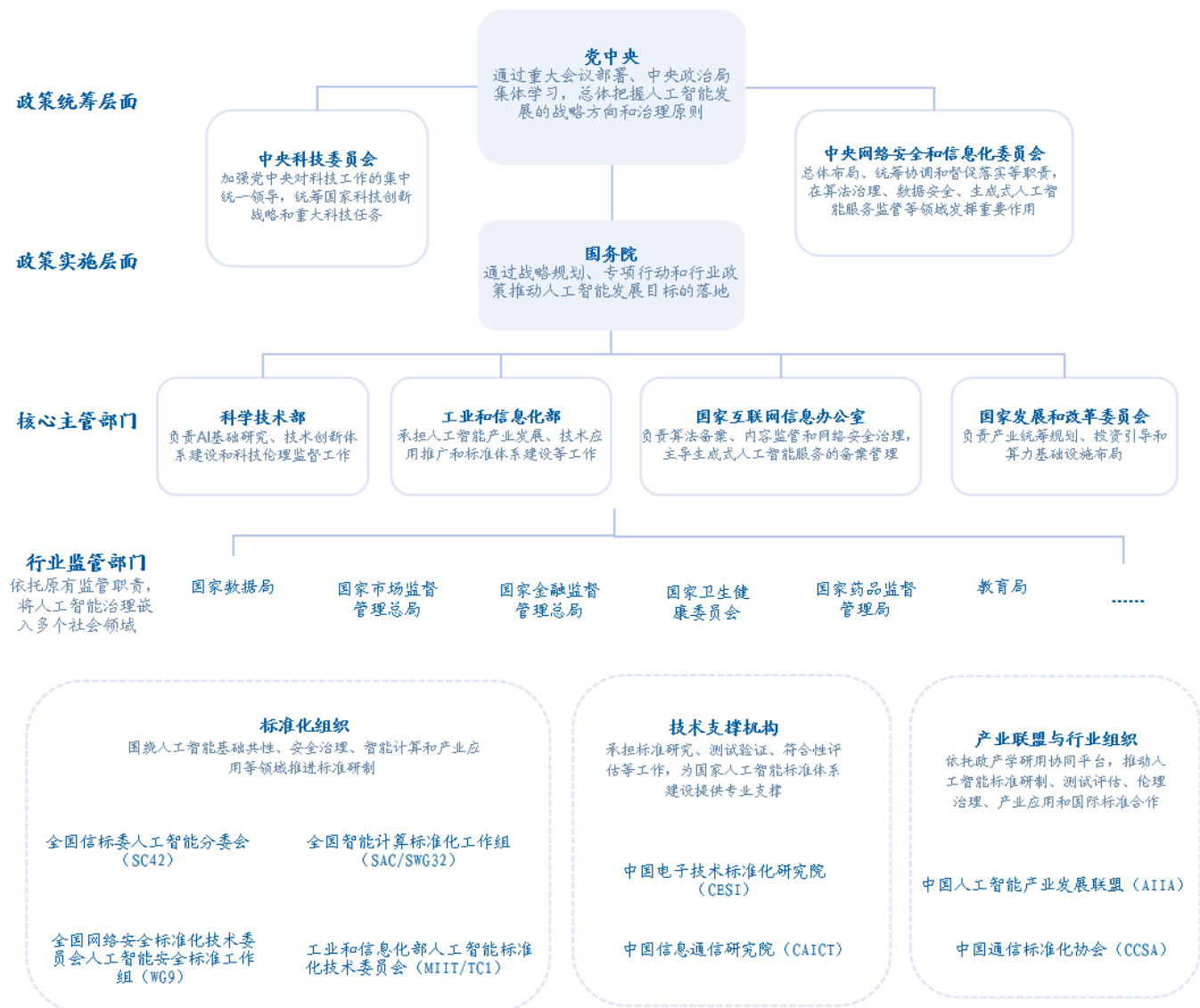
在政策实施层面，国务院及相关部门通过战略规划、专项行动和行业政策推动人工智能发展目标的落地。2017年国务院发布的《新一代人工智能发展规划》首次将人工智能发展上升为国家战略，围绕科技创新、产业应用、智能基础设施和安全治理等方面进行系统部署。此后，国务院发布的《“十四五”规划纲要》、《数字中国建设整体布局规划》、《关于深入实施“人工智能+”行动的意见》、《“十五五”规划纲要》等国家战略文件进一步将人工智能纳入科技创新和数字经济发展等核心领域，推动人工智能从单一技术领域上升为支撑现代化产业体系建设的重要战略能力。

在具体执行层面，科学技术部、工业和信息化部、国家互联网信息办公室和国家发展和改革委员会等部门分工协作，形成了覆盖科技创新、产业发展、安全监管和基础支撑的协同治理体系。科学技术部主要负责人工智能基础研究、技术创新体系建设和科技伦理监督工作；工业和信息化部重点承担人工智能产业发展、技术应用推广和标准体系建设等工作，成立了工业和信息化部人工智能标准化技术委员会，推动人工智能领域标准制定和产业规范发展；国家互联网信息办公室负责算法备案、内容监管和网络安全治理，主导生成式人工智能服务的备案管理工作；国家发展和改革委员会则负责产业统筹规划、投资引导和算力基础设施布局，将人工智能发展纳入国民经济和社会发展规划的顶层框架。

除核心主管部门外，各行业主管部门也依托原有监管职责，将人工智能治理嵌入金融、医疗、教育等多个领域。2023年成立的国家数据局统筹推进数字中国、数字经济、数字社会规划和建设，为人工智能发展提供数据基础支撑；国家市场监督管理总局主要从市场秩序、竞争监管等方面参与人工智能治理，其管理的国家标准化委员会负责统筹人工智能国家标准体系建设，推动国家人工智能标准化总体组、专家咨询组等机制运行；国家金融监督管理总局、国家卫生健康委员会、国家药品监督管理局、教育部等行业主管部门也在各自领域建立专项治理机制，分别针对金融业人工智能安全开发、医疗业人工智能应用场景、人工智能药品监管和人工智能教育体系发布专项治理政策，实现细分场景下的差异化监管落地。

标准化组织、技术支撑机构和产业联盟承担国家标准制定、技术评估和行业协同等功能，成为政府治理与产业实践之间的重要纽带。在国家人工智能标准化总体组的统筹下，全国信标委人工智能分委会（SC42）、全国智能计算标准化工作组（SAC/SWG32）、全国网络安全标准化技术委员会人工智能安全标准工作组（WG9）以及工业和信息化部人工智能标准化技术委员会（MIIT/TC1）等组织分别围绕人工智能基础共性、安全治理、智能计算和产业应用等领域推进标准研制。中国电子技术标准化研究院（CESI）和中国信息通信研究院（CAICT）作为核心技术支撑机构，承担标准研究、测试验证、符合性评估等工作，为国家人工智能标准体系建设提供专业支撑。此外，中国人工智能产业发展联盟（AIIA）和中国通信标准化协会（CCSA）分别依托政产学研用协同平台，推动人工智能标准研制、测试评估、伦理治理、产业应用和国际标准合作，促进治理规则向产业端传导。

图11：中国人工智能治理架构示意图



数据来源：中国政府网、中国电子技术标准化研究院 CESI、工信部人工智能标准化技术委员会、全国网络安全标准化技术委员会官网、中国人工智能产业发展联盟官网等，国泰海通证券研究

4.2. 创新要素：围绕算力、数据、技术、开源生态、资本和人才完善人工智能发展基础

中国人工智能产业发展更加注重基础能力建设，通过算力、数据、核心技术、开源生态、资本和人才六大关键要素，完善人工智能产业发展的基础。一方面，通过硬件、数据和算法补齐人工智能发展日益增长的生产要素缺口；另一方面，通过开源共享、稳定投资和全链条人才培养缓解创新发展的潜在难题，兼顾自主技术安全与产业创新活力，夯实我国人工智能长期稳定发展的根基。除了资源投入规模外，基础设施利用效率和产业调配能力也是提升未来人工智能竞争力的重要一环。

4.2.1. 算力基建：推进智能计算基础设施布局，提升人工智能发展支撑能力

算力基础设施是人工智能产业发展的物理支撑，面对人工智能发展带来的算力需求扩张，我国通过统筹布局智算集群、推动算力互联互通和优化能源配套保障等措施，构建起全域协同、低碳高效的算力供给体系。我国出台专项行动计划与指

导意见，围绕统筹智算网络、算力互通与基础设施建设等关键措施，持续完善国家算力供给体系，缓解人工智能发展中的算力供给压力。此外，算力基础设施建设具有投资规模大、建设周期长的特点，算力利用率的提升和资源的布局协调都是完善算力供给体系的关键任务。

- 2023 年 10 月 8 日，工业和信息化部等六部门印发《算力基础设施高质量发展行动计划》，从完善算力综合供给体系、提升算力高效运载能力、强化存力高效灵活保障、深化算力赋能行业应用、促进绿色低碳算力发展、加强安全保障能力建设等多方面发力，为算力安全发展提供了总体方向。
- 2023 年 12 月 25 日，国家发展改革委等五部门印发《关于深入实施“东数西算”工程 加快构建全国一体化算力网的实施意见》，统筹通用算力、智能算力、超级算力一体化布局，推动东中西部算力一体化协同、算力与数据算法一体化应用、算力与绿色电力一体化融合、算力发展与安全保障一体化推进。
- 2024 年 12 月 31 日，国家发展改革委、国家数据局、工业和信息化部印发《国家数据基础设施建设指引》，将国家数据基础设施作为人工智能发展的关键支撑底座，明确其通过“算力-数据-算法”协同，为人工智能技术研发、场景落地和产业融合提供底层保障，助力“人工智能+”行动落地。
- 2025 年 5 月 21 日，工业和信息化部印发《算力互联互通行动计划》，提出到 2026 年建立较为完备的算力互联互通标准、标识和规则体系，建成国家、区域、行业算力互联互通平台，实现全国头部算力企业的公共算力资源互联互通；到 2028 年基本实现全国公共算力标准化互联，逐步形成具备智能感知、实时发现、按需获取的算力互联网。文件同时提出优化算力设施互联、促进算力资源互用、创新算力业务互通、打造算力互联应用场景、夯实算力网络和数据安全保障等重点任务。
- 2026 年 6 月 18 日，国家金融监督管理总局发布《关于银行业保险业人工智能安全开发应用的指导意见》，明确提出加强智能算力设施建设，应用绿色低碳技术，支持同业探索基础设施共建共享。同时，支持金融机构在安全合规前提下，使用国家算力节点或行业基础设施降低人工智能研发应用成本、提高安全运行能力。该文件主要面向金融行业人工智能应用场景，其算力相关要求对金融领域具有直接指导意义。

为支撑人工智能算力设施的规模化扩张，我国同步推进能源体系配套改革，通过提升绿电直供比例，推动算力与电力系统双向协同，加快形成低碳高效的能源保障格局，为人工智能模型训练和产业应用提供基础支撑。

- 2024 年 7 月 3 日，国家发展改革委、工业和信息化部、国家能源局、国家数据局联合印发《数据中心绿色低碳发展专项行动计划》，部署完善数据中心建设布局、严格新上项目能效水效要求、推进存量项目节能降碳改造、提升可再生能源利用水平、加强资源节约集约利用、推广应用节能技术装备等重点任务。提出到 2030 年底，全国数据中心平均电能利用效率、单位算力能效和碳效达到国际先进水平。
- 2025 年 9 月 4 日，国家发展改革委和国家能源局发布《关于推进“人工智能+”能源高质量发展的实施意见》，提出持续开展能源算力需求监测，不断提高算力中心绿电比。针对人工智能计算耗能问题，加快突破人工智能绿色低碳技术瓶颈，研究柔性直流供电、模块化小型堆等能源供给技术。聚焦能源领域数据孤岛化、算力碎片化、算法黑盒化、算力高耗能等技术瓶颈，推动开展适用能源领域的的数据、算力、算法等共性关键技术攻关。

- 2026年4月8日，国家发展改革委、国家能源局、工业和信息化部、国家数据局印发《关于促进人工智能与能源双向赋能的行动方案》，明确统筹算力项目布局与能源资源规划，提高绿电直供比例，配套建设储能设施。依托人工智能优化电网调度和储能资源调配，布局能源行业边缘算力节点，分批开放能源领域数字化应用场景。目标为保障算力设施安全可靠的能源供给，推动算力设施绿色低碳转型，并实现算力电力高效经济协同。

4.2.2. 数据资源：推动数据资源开发利用，完善人工智能训练数据供给

中国通过完善数据基础设施建设和数据资源管理机制，推动数据从生产要素向人工智能训练与应用的重要创新资源转化。我国通过构建数据基础制度、设立专机构和建设高质量数据集，搭建数据资源的统筹框架，有助于提升人工智能模型训练效果。此外，数据开放程度和隐私安全保护之间的平衡，也是数据治理规则中需要尤其重视的问题。

- 2023年10月25日，国家数据局正式挂牌成立，负责协调推进数据基础制度建设，统筹数据资源整合共享和开发利用，统筹推进数字中国、数字经济、数字社会规划和建设，组织实施国家大数据战略，为人工智能数据资源统筹配置提供了组织保障。
- 2023年12月31日，国家数据局等十七部门联合印发《“数据要素×”三年行动计划（2024—2026年）》，在国家层面明确提出要建设高质量语料库和基础科学数据集，支持开展人工智能大模型开发和训练。聚焦工业制造、科技创新、医疗健康等十二个重点领域推动数据要素协同优化和融合创新。
- 2024年11月21日，国家数据局印发《可信数据空间发展行动计划（2024—2028年）》，提出到2028年建成100个以上可信数据空间，基本建成广泛互联、资源集聚、生态繁荣、价值共创、治理有序的可信数据空间网络；面向新药研制、新材料研发等前沿领域，推动基础科学数据集、高质量语料库汇聚，促进人工智能驱动的科研范式创新应用。
- 2024年12月26日，国家发展改革委等四部门印发《关于促进数据标注产业高质量发展的实施意见》，提出到2027年数据标注产业专业化、智能化及科技创新能力显著提升，产业规模年均复合增长率超过20%；明确加强交通、医疗、金融、科学、制造、农业等重点行业领域数据标注，建设行业高质量数据集，支撑人工智能在行业领域的应用赋能。
- 2024年12月28日，国家发展改革委等六部门联合印发《关于促进数据产业高质量发展的指导意见》，明确提出“高质量数据集”概念，将其作为人工智能与实体经济融合的核心载体；提出支持企业面向人工智能应用创新开发高质量数据集，强化数据标注、数据合成等核心技术攻关，加快可信数据空间、隐私计算等可信流通技术研发和应用推广。
- 2025年5月16日，国家数据局发布《数字中国建设2025年行动方案》，强调以数据要素市场化配置改革为主线，加快培育全国一体化数据市场。将“人工智能+”列为八大重大行动之一，提出持续深挖人工智能应用新场景，建设行业高质量数据集。
- 2025年12月30日，全国数据工作会议上，国家数据局部署2026年重点工作提出要强化数据赋能人工智能发展，进一步落实高质量数据集建设行动计划，深入实施强基扩容、应用赋能、提质增效、管理服务、价值释放等6大专项行动。力争到2026年底，形成一批满足人工智能就绪度要求、有效训练先进模型、切实解决行业难题、具有国际影响力的标杆型数据集，实现高质量数据集供给量质齐升，赋能效果更加显著。

- 2026年6月4日，国家数据局召开“完善数据相关规则 赋能人工智能创新发展”座谈会，各行业专家和人工智能企业代表围绕完善数据基础制度、推动数据赋能人工智能创新发展提出意见建议。会议提出，要坚持统筹数据发展和安全，深入推进数据要素市场化配置改革，更好地适应人工智能发展的需要，更大力度地完善数据相关制度规则，更好地建设面向人工智能发展的高质量数据供给体系，进一步释放数据要素价值，推进数据要素赋能人工智能创新发展取得新成效。

在总体部署的基础上，国家及各行业主管部门围绕制造、医疗、通信、金融等重点领域出台专项政策，以“模数共振”等专项行动为抓手，推动高质量数据集的构建与应用。

- 2025年10月20日，国家卫生健康委等五部门在《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》中提出，到2027年建立一批卫生健康行业高质量数据集和可信数据空间，形成一批临床专病专科垂直大模型和智能体应用。
- 2025年12月25日，工业和信息化部等八部门在《“人工智能+制造”专项行动实施意见》中部署开展“模数共振”行动，提出打造100个高质量工业数据集，强化人工智能赋能底座。
- 2026年3月13日，《中华人民共和国国民经济和社会发展第十五个五年规划纲要》明确提出完善数据标准体系和质量管理体系，加快建设人工智能语料库，面向能源、交通、制造、教育、健康、金融等领域建设高质量数据集，建立人工智能训练数据合理使用制度，为模型迭代创新提供持续的数据供给。
- 2026年4月28日，工业和信息化部办公厅和国家数据局综合司联合实施2026年“模数共振”行动，重点面向信息通信、网络安全、消费电子、石化化工、医疗装备等20个行业或领域，推动产出一批推广价值高、技术可行性强的人工智能应用场景，构建一批行业通识和行业专识高质量数据集，推动人工智能高水平赋能新型工业化。
- 2026年6月3日，国家数据局印发《关于推进行业高质量数据集建设行动的实施方案》，从国家层面对数据赋能人工智能发展作出系统性部署，围绕行业高质量数据集供给、流通、应用等关键环节，部署强基扩容、标注攻坚、提质增效、应用赋能、管理服务、价值释放六个专项行动。提出到2028年底建成一批覆盖重点领域、经过应用验证的行业高质量数据集。

4.2.3. 核心技术：加强关键技术攻关，提升人工智能产业自主创新能力

中国持续加强人工智能算法、芯片、基础软件等关键领域的技术突破，提升产业链自主创新能力。中国不断强化“根技术”攻关、构建自主可控的算法与芯片技术体系，通过打通国产模型与国产芯片之间的适配瓶颈，增强产业链韧性，降低外部技术限制带来的潜在威胁。

- 2024年1月18日，工业和信息化部等七部门印发《关于推动未来产业创新发展的实施意见》，提出加快突破GPU芯片、集群低时延互连网络、异构资源管理等技术，建设超大规模智算中心，满足大模型迭代训练和应用推理需求。
- 2025年2月19日，国务院国资委召开中央企业“AI+”专项行动深化部署会，聚焦“根技术”攻关，明确要求中央企业在AI芯片、算法框架等基础领域取得突破，打造行业大模型，推动人工智能在智能制造、智慧能源等领域的高

价值场景落地。

- 2025 年 7 月 15 日，工业和信息化部、中国人工智能产业发展联盟开展人工智能软硬件协同创新高级别研讨会，发布上海人工智能实验室跨域异构混训技术、AISHPerf 与中国电信研究院共建的“翼芯”智算测试与适配优化平台、中国移动研究院“芯合”异构超融合基础软件栈、北京智源研究院 FlagCX 统一通信库、AISHPerf 与百度飞桨共建的持续评测流水线五大成果。同期启动“公有云大模型服务推进计划”，推动国产软硬件适配与规范化发展。
- 2025 年 7 月 25 日，2025 世界人工智能大会开幕前夕，由阶跃星辰牵头，联合华为昇腾、沐曦、天数智芯、寒武纪等近 10 家国产芯片及基础设施企业共同发起“模芯生态创新联盟”，推动模型企业、芯片企业与平台企业加强协同合作，促进软硬件生态融合发展。
- 2025 年 8 月 22 日，工业和信息化部、市场监管总局发布《电子信息制造业 2025-2026 年稳增长行动方案》，明确提出加强 CPU、高性能人工智能服务器、软硬件协同等攻关力度，开展人工智能芯片与大模型适应性测试，适度超前部署新型基础设施建设，强化服务器、芯片和关键模块的兼容适配。
- 2025 年 10 月 28 日，全国人民代表大会常务委员会通过关于修改《中华人民共和国网络安全法》的决定，新增人工智能专门条款，明确“国家支持人工智能基础理论研究和算法等关键技术研发，推进训练数据资源、算力等基础设施建设”，从法律层面强化对核心技术研发的支持保障。
- 2025 年 12 月 25 日，工业和信息化部等八部门印发《“人工智能+制造”专项行动实施意见》，明确提出推动智能芯片软硬协同发展，支持突破高端训练芯片、端侧推理芯片、人工智能服务器、高速互联、智算云操作系统等关键技术；支持模型训练和推理方法创新，开发适应制造业实时性、可靠性、安全性特点的高性能算法模型；提升软件智能化水平，加快工业实时操作系统等核心软件智能化改造升级，优化基础软件内核，植入智能调度算法。
- 2026 年 3 月 13 日，《中华人民共和国国民经济和社会发展第十五个五年规划纲要》提出要促进模型算法迭代创新，加快突破人工智能基础理论和核心技术，推进人工智能模型架构改进、算法优化，强化“模芯云用”协同创新；构建任务导向、灵活授权、跨域协同的算法创新组织模式，加快研究更加高效的模型训练和推理方法；提出推动通用大模型和行业专用模型同步发展，依托高价值场景推动模型应用落地和迭代升级；鼓励多模态、智能体、具身智能、群体智能等前沿技术创新，探索通用人工智能发展路径，构建层次分明、协同演进的大模型技术体系，强化基础模型层面的自主创新能力。
- 2026 年 6 月 3 日，工业和信息化部办公厅发布《关于组织开展 6G 创新发展部省协同试点专项行动的通知》，超前布局 6G 原生智能通信网络体系，开展通感算智一体化试验验证，抢占下一代通信技术产业制高点。提出到 2029 年，通过实施 6G 创新发展部省协同试点专项行动，进一步激发地方和企业创新活力，形成一批自主创新的 6G 技术方案，培育一批前景可观的新型业务应用场景，涌现一批丰富多样的新型终端产品。

4.2.4. 开源生态：推动开源开放合作，促进人工智能技术扩散应用

开源模型有助于扩大技术扩散范围，提高中小企业参与人工智能应用的机会。开源模型将模型权重公开供开发者下载和修改，降低了人工智能技术的应用门槛，使资源有限的中小企业也能基于开源模型开发产品和服务。从实际使用情况看，OpenRouter 与 a16z 联合发布的《State of AI 2025》报告显示，国产开源模型的周 Token 使用量占比从 2024 年底的 1.2% 上升至 2025 年后期的近 30% 的峰值。从下

载规模来看，美国麻省理工学院与 Hugging Face 发布的联合报告显示，2024 年 8 月-2025 年 8 月，Hugging Face 平台上，中国研发的开源模型全球下载量占比达 17.1%，位居全球第一。

中国积极推动人工智能开源生态建设，通过开放模型、开发工具和应用平台降低技术应用门槛，促进产业主体共同参与创新。2025 年 8 月 21 日，《关于深入实施“人工智能+”行动的意见》将“促进开源生态繁荣”列为强化基础支撑能力的重要任务，明确提出推动大模型开源社区建设，促进模型、工具、数据集等资源汇聚开放，培育共享协作的开源社区生态。2026 年 3 月 5 日，《2026 年政府工作报告》首次将“支持人工智能开源社区建设，促进开源生态繁荣”写入国家年度工作部署，标志着开源从技术圈层的自发行为上升为国家层面的战略自觉。2026 年 7 月 7 日，工业和信息化部有关负责人在 2026 世界人工智能大会新闻发布会上指出，我国企业积极拥抱开源范式，AtomGit 等本土开源社区已汇聚超 1100 万注册用户。

4.2.5. 资本投入：发挥政府与市场作用，引导人工智能产业投资布局

国家聚集产业基金、政府投资和社会资本共同参与人工智能产业的发展，为技术研发和企业成长提供资金支持。政府通过设立专项基金、引导创业投资和布局并购基金，聚焦早期项目和硬核科技企业发展，撬动社会资本参与，以多元化资本投入加快创新资源聚集。

- 2025 年 4 月 18 日，国新办新闻发布会披露，工业和信息化部联合财政部设立了 600 亿元规模的国家人工智能基金，加快布局投资项目，目前已累计培育了 400 余家人工智能领域国家级专精特新“小巨人”企业。同时，工信部通过部省协同推进 11 个国家人工智能创新应用先导区建设，央地共建了具身智能机器人、人形机器人等制造业创新中心，推动产业集聚发展。
- 2025 年 10 月 29 日，由国务院国资委发起、委托中国国新设立和管理的中央企业战略性新兴产业发展专项基金正式启动，首期规模 510 亿元，出资合作单位包括中国移动、中国石化、中国电科、国家电投等。该基金将重点投向人工智能、航空航天、高端装备、量子科技等战略性新兴产业以及未来能源、未来信息、未来制造等未来产业，支持国资央企补齐产业短板弱项、布局前沿创新。
- 2025 年 12 月 26 日，国家发展改革委员会启动国家创业投资引导基金，形成万亿资金规模，坚持做早期基金、耐心基金、市场化基金、标杆基金，坚持投早、投小、投长期、投硬科技；加强对人工智能等领域的早期项目和种子企业的投资支持，加大基金投资力度，努力带动各类金融机构以及民间资本共同投资。
- 2026 年 3 月 6 日，十四届全国人大四次会议经济主题记者会上，国家发改委指出“十五五”末人工智能相关产业规模将超 10 万亿元，将会同财政部、人民银行等部门再设立国家级的并购基金，预计撬动各类资金规模超 1 万亿元，进一步畅通创业投资退出渠道，提高创业资本周转效率。

4.2.6. 人才培养：加强人工智能人才供给，支撑产业长期发展

中国通过高校培养、科研体系建设和人才引进等方式，加强人工智能领域的专业人才储备，为产业可持续发展提供支撑。依托系统化教育改革与顶层政策规划，中国搭建起了覆盖基础教育、高等教育、职业教育和全民通识的全学段人工智能培育体系，完善人才评价与科研保障机制，精准匹配产业创新需求。

- 2024 年 11 月 18 日，教育部办公厅发布《关于加强中小学人工智能教育的通知》，构建小初高分层次系统化课程体系，将人工智能纳入课后服务，推进任务式、项目式学习。开发普适化教学资源，建设泛在化教学环境：均衡配置 AI 实验室资源，推动规模化教师供给，并组织多样化交流活动。
- 2025 年 5 月 12 日，教育部基础教育教学指导委员会发布《中小生成式人工智能使用指南（2025 年版）》，明确了生成式人工智能在中小学的典型应用场景，包括促进学生成长、辅助教师教学和支撑教育管理。限制在创造性任务中滥用人工智能，建立生成式人工智能工具“白名单”制度，仅允许符合教育场景且数据安全合规的工具进入校园，构建覆盖数据安全、伦理审查、内容监管的全链条保障机制。
- 2025 年 5 月 12 日，教育部基础教育教学指导委员会发布《中小学人工智能通识教育指南（2025 年版）》，以素养培育为核心，通过螺旋式课程设计实现从认知启蒙到创新实践的素养发展：小学阶段注重兴趣培养与基础认知，初中阶段强化技术原理与基础应用，高中阶段注重系统思维与创新实践。通过“课程重构、资源融通、评价创新、师资赋能”的联动机制，推动人工智能教育从局部试点转向全域覆盖。
- 2026 年 4 月 2 日，教育部等五部门印发《“人工智能+教育”行动计划》，提出建设国家智能算力服务平台，构筑集约高效的智能教育基座，打造虚实融合的未来教育空间，加快普及中小学人工智能教育，培养面向智能时代的高层次人才，推动职业教育传统专业转型升级。
- 2026 年 3 月 13 日，“十五五”规划纲要就人才与科研体系建设作出专门安排，要求健全高等教育学科专业设置调整机制，超常规布局人工智能等新兴领域急需学科专业，深入实施基础学科和交叉学科突破计划。完善人工智能科研保障、职业支持和人才评价机制，推进人工智能全学段教育和全社会通识教育，为各类人才施展才华搭建平台、创造条件。

4.3. 应用落地：推动人工智能与实体经济深度融合

4.3.1. 政策部署：实施“人工智能+”行动，促进重点行业的智能化转型

中国将“人工智能+”上升为国家的战略行动，通过系统性政策部署推动人工智能从单纯的技术研发向千行百业融合，赋能包括制造业、服务业、民生保障和社会治理在内的全场景产业。2025 年 8 月 21 日，国务院印发《关于深入实施“人工智能+”行动的意见》，系统部署了“人工智能+”科学技术、产业发展、消费提质、民生福祉、治理能力、全球合作六大重点行动，明确要通过提升模型基础能力、加强数据供给创新、强化智能算力统筹、优化应用发展环境、促进开源生态繁荣、加强人才队伍建设、强化政策法规保障、提升安全能力水平来强化基础支撑能力。

4.3.2. 行业应用：推动人工智能在重点领域的加速落地

中国围绕制造业、医疗、交通、能源等重点领域推进人工智能应用，逐步出台细分领域专项指导政策，促进人工智能技术在各场景的落地实施。丰富的产业场景为人工智能的持续迭代进步提供基础，对技术能力与行业需求的匹配程度提出了更高要求。

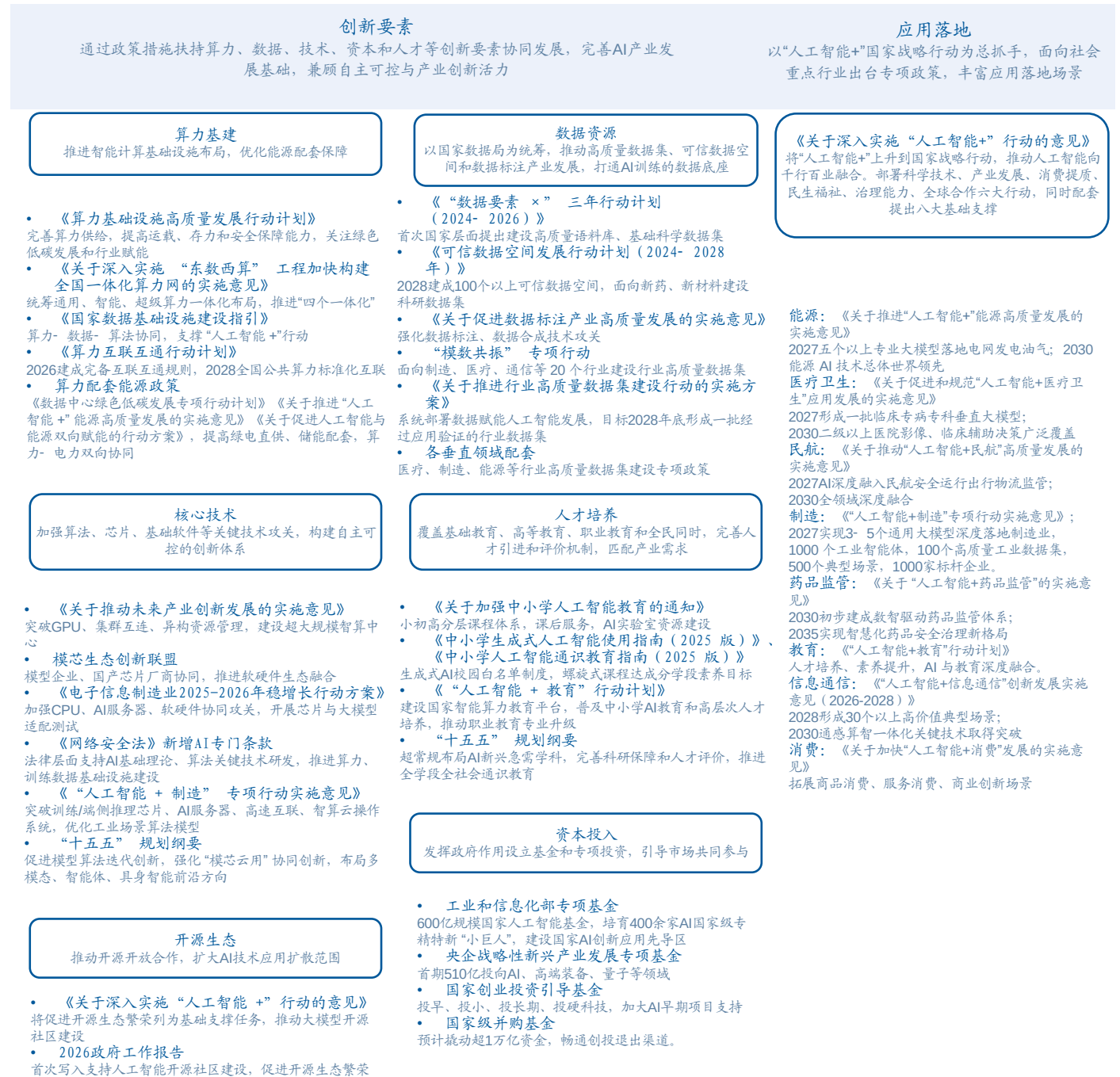
- 2025 年 9 月 4 日，国家发展改革委和国家能源局发布《关于推进“人工智能+”能源高质量发展的实施意见》，提出到 2027 年算力与电力协同发展根基不断夯实，人工智能赋能能源核心技术取得显著突破，推动五个以上专业大模型在电网、发电、煤炭、油气等行业深度应用；到 2030 年，能源领域人工智能专用技术与应用总体达到世界领先水平，算力电力协同机制进一步完善，

建立绿色、经济、安全、高效的算力用能模式。

- 2025 年 10 月 20 日，国家卫生健康委等五部门发布《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》，明确到 2027 年形成一批临床专科专科垂直大模型和智能体，基层诊疗智能辅助实现广泛应用；到 2030 年基本实现基层诊疗智能辅助应用全覆盖，推动二级以上医院普遍开展医学影像智能辅助诊断、临床诊疗智能辅助决策等人工智能技术应用全覆盖，系统部署了“人工智能+医疗卫生”发展的应用方向、重点任务、支撑体系与安全监管框架。
- 2025 年 11 月 19 日，中国民用航空局印发《关于推动“人工智能+民航”高质量发展的实施意见》，提出到 2027 年实现人工智能与民航安全、运行、出行、物流、监管等领域融合发展，形成一批具有行业引领性的示范场景和具有核心竞争力的智能化产品与应用；到 2030 年，实现人工智能与民航各领域广泛深度融合，形成一批高质量数据集、基础设施平台、行业模型算法，人工智能有效促进行业安全水平、运行效率、服务质量和资源配置能力提升。
- 2025 年 12 月 25 日，工业和信息化部等八部门印发《“人工智能+制造”专项行动实施意见》，提出到 2027 年推动 3-5 个通用大模型在制造业深度应用，推出 1000 个工业领域高水平工业智能体，打造 100 个高质量工业数据集，推广 500 个典型应用场景，选树 1000 家标杆企业。部署了夯实人工智能赋能底座、拓展推广高价值应用场景、构建智能新产品新业态等具体措施，发布人工智能赋能原材料、装备制造、电子信息等制造业重点行业的转型指引，以及制造业企业的人工智能应用指南。
- 2026 年 3 月 11 日，国家药监局印发《关于“人工智能+药品监管”的实施意见》，提出到 2030 年初步构建药品监管与人工智能融合创新体系，推进人工智能在审评审批、监督检查、检验监测、政务服务等场景中有效应用；到 2035 年，基本形成数智驱动、智能敏捷、自主可控、生态协同的智慧化药品安全治理新格局。文件紧扣智慧监管重点场景，把握人工智能发展新趋势，以数智赋能推动监管改革落地生效，筑牢“人工智能+药品监管”基础支撑。
- 2026 年 4 月 2 日，教育部等五部门印发《“人工智能+教育”行动计划》，推动人工智能人才培养与素养提升，促进人工智能与教育深度广泛融合，同时通过构筑智能教育基座、开展研究创新和条件保障，优化“人工智能+教育”的基础环境和发展生态。
- 2026 年 4 月 17 日，国家发改委在新闻发布会上指出，2026 年将重点在“人工智能+”基础设施、城市更新、国家水网、新型能源体系等领域开展一系列扩大有效投资行动。
- 2026 年 6 月 3 日，工业和信息化部印发《“人工智能+信息通信”创新发展实施意见（2026—2028 年）》，提出到 2028 年信息通信智能运营和服务能力达到国际先进水平，形成 30 个以上高价值典型场景，打造一批典型应用和特色智能体；到 2030 年人工智能与信息通信网络融合关键核心技术取得显著突破，通感算智一体化服务能力大幅提升，形成完备的协同创新和产业生态体系。部署了推动信息通信行业智能化升级、夯实人工智能发展底座、深化融合应用创新推广和增强信息通信行业治理能力等具体措施。
- 2026 年 6 月 9 日，商务部等八部门印发《关于加快“人工智能+消费”发展的实施意见》，从商品消费、服务消费和商业创新三个维度拓展人工智能的应用场景，加强人工智能在消费领域的推广，优化“人工智能+消费”的环境。

- 2026年6月22日，人社部等四部门联合印发《关于加快推进“人工智能+人社”应用发展的实施意见》，分阶段推进人社领域人工智能落地，围绕数智就业、智慧社保、人才培养、劳动关系、人力资源服务、智慧治理打造六大类应用场景。同步布局行业数据集、人社行业大模型、算力底座与配套标准，健全数据、算法伦理安全保障，推进人社领域数智化转型。

图12：中国人工智能治理体系特征（1）



数据来源：中国政府网、国家发改委官网、工业和信息化部、中央网信办、教育部官网、北京国际科技创新中心等，国泰海通证券研究

4.4. 风险监管：构建覆盖人工智能全生命周期的动态治理机制

中国的人工智能治理逐步形成覆盖算法、深度合成和生成式人工智能的动态监管机制，通过备案管理、分类分级监管、内容标识和伦理审查等方式，提升人工智能安全保障能力。上述举措逐步构建涵盖人工智能全生命周期的敏捷治理体系，在规范技术应用和降低社会风险方面发挥积极作用，同时也需要着力提升风险识别能力以适配快速迭代的技术与应用场景。

4.4.1. 监管对象持续演进，从算法管理逐步拓展至生成式人工智能治理

中国人工智能安全治理经历了从算法推荐管理、深度合成内容监管到生成式人工智能服务治理的逐步演进，监管对象覆盖了技术全生命周期，呈现出敏捷和动态监管的特征。中国人工智能监管体系根据技术发展阶段不断调整，治理的重心从单一技术环节扩展至覆盖模型研发、训练、部署和应用全过程，在事前阶段介入，以主动规范人工智能的使用流程。

- 2021年12月31日，国家互联网信息办公室等四部门正式颁布《互联网信息服务算法推荐管理规定》。该规定适用各类算法推荐服务提供者，涵盖生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等。核心要求包括：算法服务提供者须在国家公共算法备案系统完成模型备案；向用户推送算法推荐内容时，需作出明确提示；严禁利用算法从事违法违规、危害公共利益的行为；保障用户权利，可一键关闭个性化推荐、删除用于画像的用户标签；若算法对用户权益造成重大影响，平台需向用户作出合理解释。
- 2022年11月25日，国家网信办等三部门发布《互联网信息服务深度合成管理规定》，将监管对象从算法推荐逻辑扩展至深度合成内容本身，覆盖所有人工智能生成音视频图像等合成内容。该规定明确所有深度合成内容必须坚持正确政治导向，不得危害国家安全、损害国家形象、侵害社会公共利益、扰乱经济与社会秩序。若合成内容容易造成公众混淆、误导大众，必须清晰标注为人工智能生成。
- 2023年7月10日，国家互联网信息办公室等七部门发布《生成式人工智能服务管理暂行办法》，对生成式人工智能进行专项监管。该办法确立“发展与安全并重、促进创新和依法治理相结合”的原则，明确服务提供者需使用合法来源数据和基础模型，建立健全的投诉举报机制，不得生成虚假新闻和信息。
- 2024年3月1日，全国网络安全标准化技术委员会发布《生成式人工智能服务安全基本要求》（TC260-003），系统规定了语料安全、模型安全、安全措施、风险词库建设及安全评估等全流程技术要求；要求服务提供者在备案时提交安全评估报告，重点防范人工智能被用于编写恶意软件、制造生物武器或化学武器等长期潜在风险；2025年4月，该文件正式升级为国家标准 GB/T 45654-2025《网络安全技术 生成式人工智能服务安全基本要求》，2025年11月1日起实施，为监管执法提供了标准化技术依据。
- 2024年7月18日，党的二十届三中全会通过的《中共中央关于进一步全面深化改革、推进中国式现代化的决定》明确提出“加强网络安全体制建设，建立人工智能安全监管制度”，从顶层设计上确立了人工智能安全治理的制度化方向。
- 2024年9月，全国网络安全标准化技术委员会（TC260）发布《人工智能安全治理框架》1.0版，提出“包容审慎、确保安全”“风险导向、敏捷治理”“技管结合、协同应对”“开放合作、共治共享”四项原则，将人工智能安全风险分为内生安全风险（模型算法、数据、系统）和应用安全风险（网络

域、现实域、认知域、伦理域），指导研发者、服务提供者、系统使用者从技术措施和综合治理两方面防范风险。框架提出实施人工智能应用分类分级管理，建立风险等级评估体系，同时构建人工智能服务可追溯管理制度。

- 2025 年 9 月 15 日，全国网络安全标准化技术委员会发布《人工智能安全治理框架》2.0 版，在 1.0 基础上实现显著升级。在人工智能安全治理原则中增加“可信应用、防范失控”，制定《可信人工智能基本准则》，明确人类最终控制、尊重主权、价值观对齐、提升透明度、可验证、全生命周期防护、主动应急、全球共治等核心要求。在安全风险分级方面，特别增加“模型开源风险”和“低质有害信息扩散污染内容生态”等新型风险识别。强调在设计、开发、测试、部署、运行等全生命周期各环节全面嵌入安全考量，例如在研发环节强调提升算法可靠性、公平性和隐私保护等内生安全能力，在部署环节要求对软硬件工具进行安全检测和漏洞修补，在运行环节则强调完善应用安全管理和人工复核机制等。由此建立治理措施动态调整机制和分级管理，标志着治理思路从“静态治理”向“动态治理”转变。

4.4.2. 通过备案制提升治理精准度，实现高风险模型的持续跟踪

中国实行算法和大模型的备案登记制度，通过事前备案识别重点监管对象，将人工智能治理原则转化为具体可操作的监管流程。目前，备案制监管范围主要聚焦于具有舆论属性或者社会动员能力的公开服务，对于非公众场景模型的监管仍侧重于创新与风险的平衡。

- 2022 年 3 月 1 日起施行的《互联网信息服务算法推荐管理规定》确立了算法备案制度，要求具有舆论属性或社会动员能力的算法推荐服务提供者在国家公共算法备案系统完成模型备案并开展安全评估，明确了我国人工智能的事前监管程序。
- 2023 年 7 月 10 日发布的《生成式人工智能服务管理暂行办法》进一步建立了“备案”与“登记”两级管理制度：备案面向自研大模型且向境内公众提供服务的提供者，须向国家网信办提交材料，审核周期通常为六至八个月；登记面向仅通过 API 调用已备案模型能力、未进行训练优化的应用，由属地省级网信部门负责，采用简化流程，登记办理周期通常为三至四个月。
- 2024 年 4 月，国家互联网信息办公室发布生成式人工智能服务已备案信息，公开披露通过备案的大模型名单，要求提供具有舆论属性或者社会动员能力的生成式人工智能服务的，通过网信部门履行备案程序。截至 2026 年 6 月 30 日，全国已有 988 款生成式人工智能服务完成备案、598 款完成登记。
- 中国气象局于 2025 年 4 月 23 日发布的《人工智能气象应用服务办法》，及国家卫生健康委等五部门于 2025 年 10 月 20 日发布的《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》等行业专项文件，均强调投入服务应用的行业专用人工智能模型需满足备案要求。

4.4.3. 强化全链条内容标识，通过强制溯源提升人工智能应用的透明度

中国通过立法与技术配套，落实合成内容标识标准规范，将人工智能治理要求落实到具体应用环节，逐步构建起覆盖生成、传播和使用全过程的强制性内容标识体系，提高人工智能应用的透明度和责任可追溯性，并持续向垂直行业与新兴场景延伸。

- 2022 年 11 月 25 日，《互联网信息服务深度合成管理规定》率先建立深度合成内容标识要求，明确若合成内容容易造成公众混淆、误导大众，必须清晰标注为 AI 生成，并须取得用户单独同意，为内容标识制度奠定规则基础。

- 2023 年 7 月 10 日，国家互联网信息办公室等七部门发布的《生成式人工智能服务管理暂行办法》将内容标识义务扩展至所有生成式人工智能服务，要求服务提供者对图片、视频等生成内容进行标识，确保生成内容可追溯、可问责。
- 2024 年 6 月 5 日，工业和信息化部等四部门发布《国家人工智能产业综合标准化体系建设指南（2024 版）》，提出到 2026 年新制定国家标准和行业标准 50 项以上，推动超过 1000 家企业实施标准。围绕基础共性、基础支撑、关键技术、智能产品与服务、赋能新型工业化、行业应用、安全/治理 7 个模块构建标准体系。
- 2024 年 9 月，全国网络安全标准化技术委员会制定《人工智能安全治理框架》1.0 版，建立人工智能服务可追溯管理制度。对面向公众服务的人工智能系统，通过数字证书技术进行标识管理。制定出台人工智能生成合成内容标识标准规范，全面覆盖制作源头、传播路径、分发渠道等关键环节。
- 2025 年 3 月 7 日，国家网信办等四部门发布《人工智能生成合成内容标识办法》，建立显式标识与隐式标识相结合的全链条标识制度。显式标识要求在文本、音频、图片、视频、虚拟场景等内容的显著位置添加用户可感知的标识；隐式标识要求在文件元数据中添加机器可读的标识信息（如生成属性、服务提供者编码等），并将传播平台纳入监管范围，实现从生成到传播的全链条治理。
- 2025 年 3 月 15 日，浙江大学区块链与数据安全全国重点实验室联合中国电子技术标准化研究院，研发人工智能生成合成内容标识服务平台 GCmark，作为我国首个人工智能领域强制性国家标准《网络安全技术 人工智能生成合成内容标识方法》的配套服务平台，为生成合成服务提供者与内容传播平台提供一站式的标识生成、检测能力支持，从技术层面保障标识制度落地。
- 2025 年 4 月 23 日，中国气象局、国家网信办发布《人工智能气象应用服务办法》，明确要求人工智能气象应用服务须落实合成内容标识、建立安全风险机制，体现内容标识制度从通用规范向气象等垂直行业延伸落地。
- 2025 年 9 月发布的《人工智能安全治理框架》2.0 版强调人工智能生成内容的溯源标识，完善对合成内容的显式、隐式标记和追溯机制，覆盖内容生产源头、传播路径和分发渠道，方便用户识别信息来源真伪。对模型安全漏洞扩散、人工智能自主性风险以及生成内容可信度等问题进行了更高的关注和更积极的防护。

4.4.4. 全面推进科技伦理审查，落实研发周期前置防控措施

中国逐步建立起覆盖研发前、中、后期的伦理审查机制，将伦理考量前置、制度化，并持续拓展至社会伦理风险防范等新兴领域。通过不断完善人工智能专项科技伦理审查体系，降低虚假信息、隐私泄露等问题可能带来的社会风险。

- 2023 年 9 月 7 日，科技部等十部门发布《科技伦理审查办法（试行）》，将“具有舆论社会动员能力、高度自主决策以及涉及人机融合的 AI 研发”纳入科技伦理审查敏感清单，要求从事人工智能等科技活动或研究内容涉及科技伦理敏感领域的单位，设立科技伦理（审查）委员会，并在研发前进行严格的伦理审查与风险评估，弥补了事后监管的时滞性，为全生命周期治理提供了前置性审查机制。
- 2025 年 11 月 12 日，全球首个系统级伦理垂域大模型“问道”在东南大学发布，标志着我国在人工智能伦理治理领域实现从被动合规向主动共建的重要

转变，通过技术手段赋能伦理审查。

- 2026 年 3 月 20 日，工业和信息化部等十部门印发《人工智能科技伦理审查与服务办法（试行）》，明确从事人工智能科技活动的高等学校、科研机构、医疗卫生机构、企业应设立人工智能科技伦理委员会，审查要点涵盖人类福祉、公平公正、可控可信、透明可解释、责任可溯源、隐私保护等维度，建立一般程序、简易程序、专家复核程序和应急程序。地方、相关主管部门可结合实际情况依托相关单位建立专业性人工智能科技伦理审查与服务中心。服务中心接受其他单位委托，提供人工智能科技活动伦理审查、复核、培训、咨询等服务。
- 2026 年 5 月 19 日，全国网络安全标准化技术委员会发布了《人工智能应用伦理安全指引 1.0》，给出了人工智能应用开发、服务提供和应用使用等环节的伦理安全指引，进一步夯实可信 AI 建设的规则基础。

4.5. 安全保障：强化数据与关键领域管控，筑牢人工智能发展边界

在推动人工智能产业发展的同时，中国持续强化对数据安全、模型安全和重点行业应用风险的监管，通过法律法规、技术标准和行业准入机制提升安全保障能力。中国人工智能治理体系建立了以《数据安全法》《个人信息保护法》为法律基础的数据安全制度，推进实施面向重点行业的刚性准入门槛，完善国家安全审查与常态化执法，确保技术发展在可控、可信、可追责的轨道之上。

4.5.1. 建立人工智能风险分类管理机制和技术安全基准

中国以《人工智能安全治理框架》为核心技术基准，对人工智能风险类型、风险等级和治理措施进行系统划分，推动监管从统一约束向分类分级管理转变。有助于提高安全治理的精准度，使得监管资源更加集中于高风险领域，为低风险领域保留一定创新空间，同时风险等级划分和评估标准也需要随着技术演进不断调整。

- 2024 年 9 月发布的《人工智能安全治理框架》1.0 版提出实施人工智能应用分类分级管理，建立风险等级测试评估体系，对算力、推理能力达到一定阈值或应用在特定行业领域的人工智能系统进行登记备案，要求其具备在设计、研发、测试、部署、使用、维护等全生命周期的安全防护能力。

表5：《人工智能安全治理框架》风险类别

风险类别		技术应对措施
内生安全风险	模型算法安全风险	建议提供明确说明、建立并实施安全开发规范。
	数据安全风险	主要着眼于个人信息保护和知识产权保护，以及数据处理（包括跨境数据处理）的规则和措施。
	系统安全风险	包括适当公开风险、标识输出内容、加强风险识别检测与防护、加强运维能力、关注供应链安全。
应用安全风险		包括建立安全防护机制、数据护栏、设置服务提供边界、筛选判别不真实或有偏见歧视的数据等措施。

数据来源：中央网信办、全国网安标委 TC260，国泰海通证券研究

- 2025 年 9 月发布的《人工智能安全治理框架》2.0 版完善优化了风险分类标准，提出技术内生、技术应用、应用衍生三类核心安全风险。框架从应用场景、智能化水平、应用规模等维度科学评估风险等级，建立“低、一般、较大、重大、特别重大”五级风险分级标准，采取相应的差异化分类分级管理，以做到高风险严管、低风险松绑，提升治理的精准性和有效性。具体措施包括提升模型可解释性与鲁棒性、健全法律法规等技术跨维度综合治理举措。

- 2026 年 3 月，全国网络安全标准化技术委员会正式组建“人工智能安全标准工作组”（WG9），专门负责人工智能安全标准现状调研、标准体系研究与标准制定工作，标志着我国人工智能安全标准建设进入专业化、系统化新阶段。2026 年 6 月，工业和信息化部进一步成立工业智能体标准工作组与人工智能功能安全标准工作组，分别聚焦工业智能体标准与人工智能功能安全标准研制，推动安全标准从通用要求向工业场景、功能安全等纵深领域延伸，形成覆盖通用与垂直领域的标准组织保障体系。
- 2026 年 4 月 30 日，国家市场监督管理总局、国家标准化管理委员会发布推荐性国家标准《人工智能 可信赖 通则》（GB/T 47507-2026），由中国电子技术标准化研究院等单位牵头起草。作为国内可信赖人工智能领域基础性通用标准，搭建可信赖人工智能技术框架，明确安全性、公平性、可解释、可问责等核心要求，并具体罗列出可信赖应用场景。该标准于 2026 年 8 月 1 日起实施。
- 2026 年 5 月 8 日，工业和信息化部、商务部、市场监管总局等部门联合启动实施《人工智能终端智能化分级》（GB/Z 177-2026）系列国家标准，采用“2+N”架构，明确终端智能化的概念、等级划分和测试方法，建立从 L1 响应级、L2 工具级、L3 辅助级到 L4 协同级的智能化等级体系，为手机、电脑、电视、眼镜、汽车座舱、音箱、耳机等不同终端产品的安全能力和智能化水平评估提供统一标准，推动人工智能终端产品安全准入的制度化。

4.5.2. 以基础法律强化数据安全治理和个人信息保护

中国设立了数据安全法和个人信息保护特殊规则，对人工智能训练过程中的数据和个人信息要素建立刚性约束，划定安全治理的法律边界。以两部基础性法律为根基构建起制度框架，严格管控人工智能训练数据和个人信息处理方面的安全风险，为人工智能设定刚性的数据安全规则。

- 2021 年 9 月 1 日，《中华人民共和国数据安全法》正式施行，明确国家对数据实行分类分级保护，建立数据安全审查制度，对影响或者可能影响国家安全的数据处理活动进行国家安全审查。要求数据处理者建立健全全流程数据安全管理制度，组织开展数据安全教育培训，采取相应的技术措施和其他必要措施保障数据安全，为人工智能训练数据的安全管理提供了顶层法律依据。
- 2021 年 11 月 1 日，《中华人民共和国个人信息保护法》正式施行，建立个人信息处理全链条规则，对生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等敏感个人信息实行严格保护，明确处理敏感个人信息应当取得个人的单独同意。要求处理个人信息应当具有明确、合理的目的，并应当与处理目的直接相关，采取对个人权益影响最小的方式，为人工智能训练中涉及的个人信息处理划定了法律边界。
- 2024 年 8 月 30 日，国务院第 40 次常务会议通过《网络数据安全管理条例》，进一步细化生成式人工智能服务训练数据的安全管理要求，明确提供生成式人工智能服务的网络数据处理者应当加强对训练数据和训练数据处理活动的安全管理，采取有效措施防范和处置网络数据安全风险。同时，条例明确大型网络平台服务提供者不得利用网络数据、算法以及平台规则等从事误导欺诈骗、不合理差别待遇或限制用户访问其在平台上产生的网络数据等活动，从数据治理层面防范算法滥用风险。
- 2024 年 9 月，《人工智能安全治理框架》1.0 版提出了完善人工智能数据安全和个人信息保护规范，强化人工智能供应链安全保障，推进人工智能可解释性研究、人工智能安全风险威胁信息共享和应急处置机制等综合治理措施。

4.5.3. 构筑国家安全审查体系，将人工智能纳入安全部署

中国将人工智能纳入国家安全层面，通过企业安全承诺、专项治理行动和完善法律制度等方式健全国家安全审查机制，提升人工智能的产业安全和战略风险管理能力。国家通过常态化执法和专项行动，提升人工智能发展过程中的风险识别和应对能力，推动人工智能从可用向可信过渡。

- 2024 年 12 月 23 日，在 2025 中国信通院深度观察报告会主论坛上，首批 17 家行业领军企业正式签署《人工智能安全承诺》，通过产业自律方式承诺设置安全团队或组织架构、构建安全风险管理机制、开展模型安全测试、保障训练数据和业务数据安全、提升基础设施安全、增强模型透明度、积极开展前沿安全研究。
- 2025 年 2 月 21 日，中央网信办将“整治 AI 技术滥用乱象”列为 2025 八大重点整治任务之一，突出人工智能技术管理和信息内容管理，强化生成合成内容标识，打击借 AI 技术生成发布虚假信息、实施网络水军行为等问题，规范 AI 类应用网络生态。
- 2025 年 4 月启动“清朗·整治 AI 技术滥用”专项行动，第一阶段强化 AI 技术源头治理，重点整治违规 AI 产品、传授售卖违规 AI 教程、训练语料管理不严、安全管理措施薄弱、未落实内容标识要求、重点领域安全风险等问题；第二阶段聚焦利用 AI 制作发布谣言、不实信息、色情低俗内容，假冒他人、从事网络水军活动等突出问题。
- 2025 年 10 月 28 日，全国人民代表大会常务委员会通过关于修改《中华人民共和国网络安全法》的决定，新增人工智能专门条款，明确“完善人工智能伦理规范，加强风险监测评估和安全监管，促进人工智能应用和健康发展。国家支持创新网络安全管理方式，运用人工智能等新技术，提升网络安全保护水平”，从法律层面为人工智能安全监管提供了上位法支撑。
- 2026 年 3 月 13 日，《中华人民共和国国民经济和社会发展第十五个五年规划纲要》就人工智能监管作出系统性部署，要求完善人工智能领域法律法规、政策制度、应用规范、伦理准则，健全算法备案、透明度管理、安全评估等制度，推动建立人工智能全生命周期风险管理制度，健全覆盖安全监测、风险预警、应急响应的风险防控体系，加强平台企业数据、算法、流量和规则监管，依法打击数据滥用、深度伪造、泄露隐私等行为，将敏捷治理纳入国家战略规划。
- 2026 年 4 月 27 日，国家发展改革委外商投资安全审查工作机制办公室发布公告，依法禁止 Meta(原 Facebook)以约 20 亿美元收购 AI 智能体公司 Manus 的交易。这是《外商投资安全审查办法》2021 年实施以来，中国首次在人工智能领域以“国家安全”为由公开否决外资并购案，明确将智能体等前沿技术纳入外商投资安全审查核心范畴。
- 2026 年 4 月，中央网信办启动为期 4 个月的“清朗·整治 AI 应用乱象”专项行动。第一阶段为“清朗·AI 应用服务典型违规问题”专项治理，重点整治未按规定履行大模型备案登记义务、安全和审核过滤能力不足、大模型训练语料安全、AI 数据投毒、生成合成内容标识落实不到位、开源模型安全管理不到位等问题；第二阶段为“清朗·整治 AI 信息内容乱象”，聚焦利用 AI 生成低质有害信息、制作发布虚假不实信息、散播暴力低俗内容、假冒仿冒他人、侵害未成年人权益、从事网络水军活动等问题。
- 2026 年 4 月 14 日，中国信息通信研究院发布可信 AI 智能体评估体系 2.0，覆盖智能体基础设施、数据资源、核心组件、平台支撑、关键能力、典型应

用、运维管理、价值评价等八大维度，构建了面向企业级智能体的全链路评估框架，为智能体在复杂开放环境下的可信运行、安全可控提供标准化支撑。

- 2026 年 6 月 29 日，国务院常务会议听取人工智能发展情况汇报，明确提出“要守牢人工智能安全底线，完善科技伦理、测试认证等制度规则，构建动态适应、分级分类的安全监管体系”，将安全底线作为人工智能发展的前置条件。

4.5.4. 完善重点领域的准入机制

中国通过发布技术规范和行业监管规则，完善重点行业的安全准入规制与产品安全认证机制，防止高风险应用的无序扩散。在通用治理框架之外，中国针对金融、社交互动、终端产品等重点领域建立差异化安全准入与产品认证机制，通过垂直行业的刚性约束和终端产品的标准化分级，确保人工智能产品和服务的安全底线。

- 2022 年 6 月 30 日，《人工智能 深度合成图像系统技术规范》（T/CESA 1197-2022）作为团体标准由中国电子工业标准化技术协会发布并实施，确立了深度合成图像（含视频）系统的框架和技术要求。2026 年 4 月 28 日，工业和信息化部正式批准该标准为行业标准（SJ/T 12108-2026），完成了从团体标准到行业标准的升级，适用于深度合成图像系统的设计、开发、测试、评估、管理等，为深度合成产品的安全准入提供了技术基准。
- 2025 年 3 月 13 日，国家网信办、公安部发布《人脸识别技术应用安全管理规定》，针对生物识别信息这一高风险个人数据建立特殊保护规则。办法明确，人脸信息应当存储于本地设备，不得通过互联网对外传输；禁止将人脸识别技术作为唯一身份验证方式；人脸识别技术应用于人脸验证、人脸辨识、人脸分析等场景，存储人脸信息数量达到 10 万人的，应当向所在地省级网信部门备案。该办法体现了对人工智能应用中最为敏感的生物识别数据的底线保护思维。
- 2026 年 4 月 10 日，国家网信办等五部门公布《人工智能拟人化互动服务管理暂行办法》，针对模拟自然人人格特征的持续性情感互动服务（如 AI 陪伴、虚拟伴侣等）设立安全红线。办法要求，不得向未成年人提供虚拟亲属、虚拟伴侣等虚拟亲密关系服务；用户连续使用每超过 2 小时需提醒注意使用时长；发现用户过度依赖或沉迷需动态提醒；提供便捷退出途径，不得阻碍用户退出。该办法自 2026 年 7 月 15 日起施行。
- 2026 年 5 月 8 日，中央网信办等三部门发布的《智能体规范应用与创新发展实施意见》将防范未成年人、老年人沉迷成瘾、情感依赖等风险作为重要治理目标，要求智能体服务提供者建立防沉迷机制和用户保护机制，强化社会伦理风险防范。
- 2026 年 6 月 18 日，国家金融监督管理总局发布《关于银行业保险业人工智能安全开发应用的指导意见》，系统建立金融行业人工智能风险治理框架。意见要求银行业保险业金融机构完善人工智能治理架构，建立人工智能应用管理体系；实施风险分类分级管理，强化高风险应用准入管理，加强高风险应用监测和干预；加强外包风险管理与供应链风险、开源技术管理；从增强稳健性、提高透明度、促进可解释性、保障道德伦理与公平性、加强数据安全与个人信息保护、提升网络安全防御能力等维度，全面提升人工智能安全开发应用能力。
- 2026 年 7 月，《智能体个人信息保护自律公约》正式发布，腾讯、百度、美团等 31 家企业首批签署，该公约填补了智能体领域专项规范空白，通过头部企业发挥表率带头作用，引领人工智能行业生态规范发展。签署企业将严格

对表公约要求，在数据采集、存储、加工等关键环节建立全链条风险管控机制，从源头织密个人信息保护网。

4.6. 国际参与：提出人工智能治理的中国方案

中国积极参与全球人工智能治理进程，通过发布国际倡议、参与多边机制和推动建设合作平台，加强人工智能治理规则交流，提升在国际人工智能议题中的参与度。

4.6.1. 积极发布国际倡议，提出人工智能治理理念和合作方案

中国以“以人为本、智能向善”为价值内核，通过发布国际倡议、合作计划和行动方案参与全球人工智能治理讨论，促进形成公正合理的国际人工智能治理秩序。从系统阐述全球治理立场，到通过世界人工智能大会提出完整的合作行动计划，中国在国际人工智能治理上实现了从原则倡议到计划落地的深化过程。

- 2023 年 10 月 18 日，习近平主席在第三届“一带一路”国际合作高峰论坛开幕式上宣布提出《全球人工智能治理倡议》，倡议提出“以人为本、智能向善”的理念，倡导相互尊重、平等互利，各国无论大小、强弱，都有平等发展和利用人工智能的权利，反对以意识形态划线或构建排他性集团，为全球人工智能治理贡献中国方案。
- 2024 年 9 月 25 日，中方在纽约联合国总部举行的“加强人工智能能力建设国际合作高级别会议”上，宣布《人工智能能力建设普惠计划》，聚焦弥合数字和智能鸿沟。该计划坚持联合国在国际发展合作中的统筹协调作用，倡导真正多边主义，基于主权平等、发展导向、以人为本、普惠包容、协同合作原则，通过南北合作、南南合作和三方合作等形式，切实落实联大加强人工智能能力建设国际合作决议，推动落实联合国 2030 年可持续发展议程。
- 2025 年 7 月 26 日，2025 世界人工智能大会暨人工智能全球治理高级别会议发表《人工智能全球治理行动计划》，以“向善为民、尊重主权、发展导向、安全可控、公平普惠、开放合作”为目标，提出 13 项行动呼吁，涵盖促进创新发展、推动赋能千行百业、加快数字基础设施建设、营造多元开放创新生态、推进优质数据供给、有效应对能源环境问题、促进标准及规范共识、开展安全治理、构建包容治理模式等方面。
- 2025 年 9 月，中国国务院总理李强在纽约联合国总部全球发展倡议高级别会议上提出《“人工智能+”国际合作倡议》，围绕人工智能赋能民生福祉、科技进步、产业应用、文化繁荣、人才培养五大重点方向，呼吁各国结合自身国情积极推进“人工智能+”行动，加强政策交流和务实合作，支持联合国落实《全球数字契约》，推进全球人工智能治理对话和国际人工智能科学小组相关工作。
- 2025 年 10 月 31 日，习近平主席在亚太经合组织（APEC）第三十二次领导人非正式会议上，倡导人工智能成为全球公共产品，倡议成立世界人工智能合作组织，推动亚太地区技术开源、提升人工智能素养，促进人工智能与绿色技术融合，助力数字化绿色化转型。
- 2026 年 7 月 17 日，2026 世界人工智能大会暨人工智能全球治理高级别会议上，国家主席习近平发表题为《携手构建公正合理的全球人工智能治理体系》的主旨讲话，提出“坚持开放共赢，驱动创新发展”、“强化风险意识，确保安全可控”、“鼓励包容并蓄，促进文明互鉴”、“倡导和衷共济，完善全球治理”四点意见，系统阐述了中国对全球人工智能治理的核心主张。

- 2026 年 7 月 17 日，工业和信息化部会同有关方面共同指导发布《国际人工智能伦理治理行动计划》，在联合国《未来契约》及《全球数字契约》的框架下，围绕人工智能全生命周期伦理治理、风险分类分级防控、敏捷治理机制建设、产业链协同生态培育、社会氛围营造等方向，呼吁各国本着共商共建共享的原则加强政策协调与实践合作。
- 2026 年 7 月 17 日，国家发展改革委会同有关部门共同发布了《人工智能合作发展行动计划》，旨在贯彻落实习近平总书记有关“人工智能应该是造福全人类的国际公共产品”的重要指示精神，从数据、算力、生态、赋能、人才、规则、治理和伦理等方面，提出优质数据供给、智能算力普惠、开源生态共享、人工智能深度赋能、数智人才共育、规则标准共建、安全治理协作、人工智能向善等八项行动，以务实举措响应联合国关于加强人工智能国际合作、弥合数字鸿沟、推动人工智能赋能可持续发展等方面的倡议。

4.6.2. 倡导多边合作，依托国际机制拓展人工智能治理网络

中国通过联合国、金砖国家、二十国集团等多边平台以及国际组织，加强人工智能能力建设，搭建国际合作网络，推动人工智能治理从理念共识到平台化合作的发展。中国强调在联合国框架下的包容性多边合作，持续扩大人工智能发展成果的共享范围，为全球人工智能治理提供制度性公共产品。

- 2024 年 7 月，联合国大会通过《加强人工智能能力建设国际合作》决议（A/RES/78/311）后，中国积极推动相关倡议落地，成立人工智能能力建设国际合作之友小组，并与联合国经济和社会事务部合作开展能力建设研讨活动，帮助全球南方国家提升人工智能发展能力，推动弥合全球智能鸿沟。
- 2024 年 10 月 23 日，金砖国家领导人第十六次会晤通过《喀山宣言》，提出加强人工智能领域合作，推动人工智能技术普惠发展。中国积极依托金砖合作机制，推动发展中国家在人工智能能力建设、技术应用和治理规则方面加强协调，提升全球南方国家参与人工智能发展的能力。
- 2024 年 11 月 18 日，在二十国集团里约峰会上，各方围绕人工智能发展与治理展开讨论，并在《二十国集团里约热内卢领导人宣言》中强调推动包容性人工智能发展。中国积极参与 G20 人工智能治理议题协调，推动各国加强人工智能发展战略和治理理念交流。同期，中国参与亚太经合组织数字经济与人工智能合作进程，促进区域人工智能能力建设。
- 2025 年 7 月 26 日，世界人工智能大会开幕式上“全球人工智能创新治理中心”正式成立，该中心在国家和上海市大力支持下由复旦大学牵头建设，依托复旦大学在人工智能治理、跨学科研究、政策咨询和国际合作网络方面的综合优势，致力于建设面向全球的人工智能学术研究、政策咨询与能力建设开放平台。2026 年 7 月 5 日，中心作为中方人工智能机构代表和首批创始成员，在瑞士日内瓦签约加入联合国支持下的全球人工智能能力发展网络建设。
- 2026 年 7 月 16 日，由中国倡议发起的世界人工智能合作组织（WAICO）协定签署仪式在上海举行，哈萨克斯坦、老挝、巴基斯坦、俄罗斯、印度尼西亚等首批 29 个国家代表签署协定，成为创始成员国，联合国秘书长古特雷斯出席了签字仪式。该组织是独立的政府间国际组织，总部设在上海，是一个共商共建、开放包容的多边合作平台，推进各成员国人工智能能力建设，加强发展战略、治理规则、技术标准的对接协调，促进应用赋能和务实合作，确保人工智能朝着有益、安全、公平方向健康有序发展。
- 2026 年 9 月 23 日至 25 日，习近平主席对美国进行国事访问。人工智能是中美元首会晤重要议题。习近平指出，中美两国都是人工智能大国，有能力也

有责任发展好、管控好人工智能，确保人工智能发展始终受控于人、造福于民；特朗普表示，人工智能关乎人类未来，美中应就此保持对话，加强合作。在9月20日至23日举行的第八轮中美经贸磋商中，双方已在经贸磋商机制下建立人工智能对话机制，并举行首次对话。此访八点成果共识第七条将该对话纳入元首共识：双方同意建立中美人工智能对话，交流相关风险和惠益，下次对话于2026年11月举行；同时建立人工智能事件沟通渠道。此举标志着中美自2026年5月两国元首同意开展人工智能政府间对话以来，首次落实为定期对话安排。

图13：中国人工智能治理体系特征（2）



数据来源：中国政府网、中央网信办、工信部官网、中国互联网络信息中心、外交部官网、全国网安标委 TC260、全国人大网、商务部、中国信息产业网、中国电子工业标准化技术协会、国家信息中心、新华网、各高校官网等，国泰海通证券研究

5. 中美欧人工智能治理体系比较分析

5.1. 治理逻辑：从发展目标出发形成差异化的治理模式

中美欧均将人工智能治理纳入本国中长期战略，在产业基础、制度环境与战略诉求差异下，对发展、安全的优先级安排有所区别，逐步形成创新驱动、规则约束、统筹发展三类差异化治理模式。近年来，人工智能技术快速迭代并向社会各领域渗透，各国面临着技术发展速度超过传统监管能力和应用场景趋于复杂的现实挑战，数据安全、隐私保护、算法歧视也逐渐成为了影响经济社会的潜在问题。人工智能治理的目的并非单纯限制技术发展，而是在防范风险的同时为产业创新和规模化应用提供稳定的制度保障。对此，中美欧的人工智能治理体系开始兼顾产业发展、安全保障和社会治理，通过战略规划、法律法规、技术标准等明确的制度建设，降低市场主体对监管边界的不确定性，减少重复性合规成本，也有助于形成可复制的产业规则体系，提高本国人工智能企业参与全球竞争的能力。

图14：中美欧人工智能治理体系比较分析



数据来源：美国白宫、美国商务部、美国国家人工智能倡议办公室官网、欧盟委员会官网、欧盟人工智能办公室官网、全国人大官网、中国政府网、中央网信办官网、工信部官网、国家发改委官网、科技部官网、全国网安标委 TC260 官网等，国泰海通证券研究

5.1.1. 美国：以创新驱动发展的治理模式

美国以保持技术领先和国家竞争优势为核心目标，形成由创新驱动发展的治理模式。美国政府将人工智能视为影响科技竞争和国家战略优势的重要战略资源，治理政策主要围绕保持技术领先、释放产业创新活力和维护国家安全展开。拜登政府时期发布的第 14110 号行政令《关于安全、可靠和可信任地开发和使用人工智能的行政命令》强调建立风险管理和安全保障体系。而特朗普第二任期则明显提高创新在政策体系中的优先级，通过第 14179 号行政令《消除美国在人工智能领域领导地位的壁垒》以及后续《赢得竞赛：美国人工智能行动计划》等政策，强调放松监管约束和降低合规负担，释放企业创新活力。同时，美国还通过先进技术的出口管制限制潜在竞争对手获取前沿人工智能能力，并借助盟友和伙伴体系扩大美国技术和标准的国际影响力，维护国家人工智能技术优势和产业竞争力。

这种治理逻辑进一步形成了美国的分层治理特征：民用场景合规要求相对宽松，战略重点领域提高监管和风险防控要求。一方面，美国依托既有政府部门和行业监管体系，根据金融、医疗、消费者保护等不同应用场景开展治理，并通过 NIST 等机构提供风险管理框架和技术指导，完善企业自律承诺机制，减少统一监管对企业创新的约束。另一方面，在军事国防、前沿模型、网络安全等涉及到国家安全的领域，显著提升管控强度。例如，《国防授权法案（FY26 NDAA）》将人工智能安全要求纳入国防采购体系，第 14409 号行政令《促进先进人工智能创新与安全》设立前沿模型自愿前置测试机制，同时明确不设立强制性政府审批要求。

因此，美国的治理模式并非整体降低监管约束，而是将监管资源更多集中于可能影响技术优势和国家安全的关键领域。这种模式能够为市场主体留足创新空间，但政策方向随政府更替发生调整。由于行政令可被后续总统撤销、国会尚未出台覆盖全领域的综合性联邦 AI 成文法，长期监管预期存在一定不确定性。此外，在人工智能快速扩散的背景下，统一性联邦监管框架的缺位有可能导致模型滥用、算法歧视、归责困境等系统性安全风险，这也是美国政府面临的现实问题。

5.1.2. 欧盟：以规则约束和统一立法为核心的治理模式

欧盟通过统一立法框架实施强制性监管约束，形成规则约束型的治理模式，其制度基础是《人工智能法案》的风险分级机制，价值内核是基本权利保护，外部效应则通过布鲁塞尔效应向域外延伸。2024 年 8 月 1 日生效的《人工智能法案》是全球首部具有法律约束力的综合性人工智能监管法规，以风险分级为核心对人工智能系统实施差异化合规要求，并配套以最高 3500 万欧元或全球年营业额 7% 的罚款机制。欧盟人工智能办公室被赋予相应执法权限，各成员国亦须设立市场监督机制，以统一强制性规则替代成员国碎片化监管。这一模式能够在提升法律确定性的同时，将监管资源集中于高风险领域。

此外，欧盟将基本权利保护作为这一治理体系的价值内核。《人工智能法案》明确将发展以人为本、可信的人工智能，以及保护健康、安全和《欧盟基本权利宪章》所规定的基本权利作为核心立法目标，并通过基本权利影响评估等制度，将隐私保护、非歧视等价值要求纳入高风险人工智能系统的部署过程，同时推动《欧洲委员会人工智能与人权、民主及法治框架公约》向国际层面延伸。

在此基础上，欧盟还将统一规则的影响力延伸至域外，通过布鲁塞尔效应将内部监管优势转化为国际规则影响力。《人工智能法案》对在欧盟市场投放、在欧盟境内使用及部分对欧盟产生实际影响的人工智能系统设置域外使用要求，使跨国企业需要根据欧盟规则调整产品设计和合规流程。由此，欧盟试图依托统一市场和规则优势，将监管影响力向外传导，在一定程度上提升国际规则话语权，有望部分弥补其在算力和基础模型等关键产业环节上的相对不足。

欧盟治理体系的优势在于规则确定性，但高合规成本也构成产业竞争压力，因此近年来在风险管理和产业创新之间寻找新的平衡点。其规则体系具有较强确定性和统一性，有助于向外输出治理规则，提升国际话语权并简化本土企业出海的合规流程。但较高的合规要求也可能增加企业制度成本，对技术创新动力、本土人工智能产业竞争力形成一定约束。因此，欧盟近年来同步推动监管简化和创新支持政策，比如发布《人工智能大陆行动计划》和《应用人工智能战略》，提升技术自主创新能力，推动重点产业应用人工智能并支持中小企业创新。同时，通过简化合规流程、扩大监管沙盒适用范围、推出人工智能法案服务台等一系列措施降低企业合规负担。

5.1.3. 中国：形成以统筹发展为核心的治理模式

中国将人工智能纳入国家科技创新和现代化产业体系建设的总体布局，通过国家统一部署中长期规划明确总体发展方向，引领产业发展与场景化落地，形成统筹发展型的治理模式。从2017年《新一代人工智能发展规划》确立“三步走”战略目标，到《“十四五”规划纲要》将其列为前沿科技攻关领域，再到2025年国务院《关于深入实施“人工智能+”行动的意见》系统部署科学技术、产业发展、消费提质、民生福祉、治理能力和全球合作六大重点行动，以及2026年《“十五五”规划纲要》将其定位为驱动新质生产力发展的核心引擎，人工智能的战略定位不断提升。这一战略迅速下沉至具体产业场景，通过“人工智能+制造”、“人工智能+医疗”、“人工智能+能源”等专项行动，形成场景化、分领域的精细监管模式，体现出以产业应用为导向的治理思路。与美国突出市场主体驱动不同，中国更强调通过国家战略目标集中引导资源配置，形成覆盖科技创新、产业培育、基础设施建设和安全监管的治理体系。

在治理机制上，中国采取动态监管模式，随着技术演进持续调整治理工具，并积极倡导开放共建的全球治理路径。监管对象从2021年《互联网信息服务算法推荐管理规定》到2022年《互联网信息服务深度合成管理规定》，再到2023年《生成式人工智能服务管理暂行办法》和2024年《人工智能安全治理框架》，逐步形成覆盖算法备案、内容标识、伦理审查和分类分级管理的全生命周期动态治理体系，并根据不同场景实施差异化治理。与此同时，中国在全球人工智能治理中积极倡导开放共建和多边治理，通过《全球人工智能治理倡议》、世界人工智能合作组织（WAICO）和人工智能能力建设普惠计划，坚持“以人为本、智能向善”的治理理念，将发展中国家纳入治理体系，强调各国平等发展和利用人工智能的权利，推动构建平等互惠的全球人工智能治理体系。

中国由国家统一部署的治理模式在资源调配和政策落地方面具有较强执行力，但制度运行效果对多部门协同衔接和治理工具与实践的适配程度提出了较高要求。这种模式能够集中调配跨部门关键资源，实现政策目标落地，推动人工智能与实体经济的深度融合。敏捷迭代的监管规则能够跟进技术应用的新变化，实现产业发展与风险防控的同步推进。同时，治理体系落地效果仍依赖多部门协同衔接、跨领域政策规则互通，全生命周期治理工具也需要持续与产业实践适配，属于制度实践中需要持续探索完善的方向。

5.2. 治理体系：治理主体与制度工具的不同特征

5.2.1. 治理主体：分散协同、超国家协调与集中统筹并存

中美欧均已形成政府部门、专业机构和产业主体共同参与的人工智能治理体系，但三方在治理权力的配置上存在明显差异，分别呈现出分散协同、超国家协调与集中统筹的特点。人工智能发展至今已远超单纯的技术范畴，而是成为跨行业和跨领域的复合型战略资源，单一监管机构难以覆盖从基础研究和模型开发，到具体行业应用的全部环节，因此三方均不同程度地将人工智能治理嵌入既有政府部门和行业监管体系，同时引入标准化组织、技术机构与企业参与治理。不过，不同国家治理体系中的政府统筹范围、行业监管机构参与程度和专业机构管理模式均存在差异，形成与自身制度环境相匹配的权责配置方式。

美国采取白宫战略统筹、部门分工执行、行业机构嵌入的分散式治理模式，依托既有监管体系覆盖不同的人工智能治理场景。美国目前没有设立统一的综合性人工智能监管机构，白宫主要承担国家层面的战略统筹，商务部、国防部、能源部、国土安全部等部门根据各自职责开展相关治理，FTC、SEC、FDA、EEOC等传统行业监管机构将人工智能纳入消费者保护、金融、医疗和就业等领域的既有监管框架。美国还设立人工智能标准与创新中心（CAISI）、国家安全AI风险测试工作组（TRAINS）等跨部门机构开展标准研究、风险测试和政策协调，在一定程度上缓解分散治理带来的跨部门协调压力。此外，企业等市场主体也通过行业自律

承诺等方式在人工智能治理中具有较高参与度。这种模式能够有效利用既有行政体系资源和专业监管能力，使不同领域可以根据自身风险特征进行差异化管理，政策具有灵活性和调整空间。但由于不同部门之间的监管目标和政策工具存在差异，治理权责的相对分散和跨部门协调成本仍会对治理效率产生影响，联邦与州政府之间的监管关系也进一步增加了制度协调的复杂性。

欧盟形成了欧盟层面集中监管与成员国配合执行的超国家治理模式，试图在成员国分散监管的基础上建立统一治理框架。《人工智能法案》实施后，欧盟人工智能办公室承担了人工智能领域的治理和协调职能，欧洲人工智能委员会负责成员国之间的协调工作，科学专家小组、咨询论坛等机构则为监管提供技术和利益相关方支持。同时，成员国仍需设立相应的市场监督机构负责具体执行。这种制度设计将人工智能治理职能上收至欧盟层面，能够以统一规则框架减少单一市场内部的监管碎片化，缓解监管规则的不确定性。这种治理模式下，欧盟需要持续处理欧盟机构和成员国监管机构之间的职能衔接，成员国之间的监管资源和执行能力差距则成为治理体系面临的重要挑战。

中国由中央负责战略统筹和方向引领，政府部门和行业组织积极参与治理，强调通过层级分明的纵向统筹和部门协同落实治理目标。顶层治理层面，党中央通过重大会议和中央政治局集体学习，对人工智能发展和治理进行战略部署；中央科技委员会主要负责统筹党中央对科技工作的集中统一领导，承担总体布局职责；中央网信办则具体负责网络与算法领域治理的统筹协调；国务院通过发布中长期战略规划引导政策目标落地。政策执行层面，科技部、工业和信息化部、国家发展改革委等部门承担起科技创新、产业发展、算法与内容治理等职责，国家数据局以及金融、医疗、教育行业的主管部门进一步将人工智能治理嵌入具体应用领域。此外，中国电子技术标准化研究院、中国信息通信研究院以及全国信标委人工智能分委会、人工智能安全标准工作组等专业机构和标准化组织承担起技术研究、标准制定和测试评估等功能，包括中国人工智能产业发展联盟在内的产业组织则进一步连接起政府和企业。因此，中国以国家战略统筹为引领，将政府部门、行业主管部门、标准化组织和产业主体全面纳入治理体系。这一结构有利于提高跨部门政策协调效率和政策执行的统一性，但随着人工智能应用规模不断扩大，也对于治理机制的动态调试能力和市场主体的充分参与提出更高要求。

5.2.2. 制度工具：硬法监管、软法治理与复合平衡之间的不同选择

中美欧在制度工具上呈现出不同组合，其在硬法监管、软法治理与复合平衡之间的不同选择，实则也是在监管规则确定性、强制性和治理灵活性、动态性之间进行权衡。人工智能技术迭代速度迅猛，传统法律能够为风险治理提供强制性保障和治理预期的确定性，但从制定到实施的复杂过程往往存在一定时滞。为了兼顾人工智能技术的创新发展空间以及政策与实践的适配度，各方均综合使用法律法规、标准体系、实践指南、评估测试、备案制度以及行业自律的工具，只是在治理体系中给予不同工具差异化的地位和适用范围。随着人工智能从模型开发向智能体、跨行业应用和复杂人机协作等场景拓展，未来治理体系对于根据风险等级和应用场景动态组合制度工具提出更高要求。

欧盟以具有法律约束力的统一立法作为制度基础，建立风险分级监管体系，并设定了强制性的合规义务。《人工智能法案》通过禁止不可接受风险、高风险、透明度风险和低风险的风险差异化要求，为不同风险等级设立不同的监管义务，并通过法律责任和处罚机制确保规则执行。此外，欧盟还通过协调标准、合格评定、基本权利影响评估以及实践准则等工具，将框架性的法律要求具体转化为企业可执行的合规要求。欧盟制度工具的特点在于首先通过法律确定强制性监管边界，再借助标准体系和评估机制实现框架的落地，近年来也根据技术发展要求适当简化监管要求和合规流程。这种模式能有效降低成员国之间的制度差异，提升监管规则的确定性和统一性，从而减轻差异化标准带来的重复合规成本。但对于技术快速迭代的人工智能而言，法律规则制定成本和配套标准制定速度能否与技术演进

相匹配成为无法忽视的重大挑战。

美国更强调通过软法、标准和行业自律形成灵活的治理模式，在保留创新空间的同时通过行政命令和国家备忘录等顶层文件实现事后约束。2023 年 NIST 发布的《人工智能风险管理框架》（AI RMF 1.0）为政府和企业提供风险识别、评估和管理的自愿性框架。美国政府还通过企业安全承诺、行业标准和指南等方式，引导企业培养人工智能安全治理能力。然而，美国并非完全依赖自愿性规则，而是通过既有行业监管体系在消费者保护、金融、医疗等具体场景中开展事后执法。对于涉及军事、前沿模型、网络安全和关键技术出口管制等涉及国家安全的领域，则通过行政命令、法律法规和许可制度形成较强的强制性约束。美国关注监管工具和具体风险之间的适配，形成了软法整体引导、既有法律事后监管和战略领域强化约束相结合的制度工具组合。这种治理模式在为企业技术创新提供制度环境的同时，也需要警惕放松监管可能导致的内生与应用风险。

中国形成了法律法规、行政监管、标准规范和专项制度相互配合的多层次治理组合，在基本规则和法律边界明确的基础上，敏捷调整具体要求。中国目前尚未形成覆盖人工智能全领域的综合性法律，而是通过既有法律法规形成底线约束，再针对算法、深度合成、生成式人工智能等不同阶段的技术形态制定动态调整的专项管理规定。例如，《网络安全法》《数据安全法》《个人信息保护法》构成数据和网络安全领域的基础性法律框架，《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》则针对不同技术应用建立具体监管要求。在此基础上，中国进一步通过行业标准和技术规范将法律原则性要求转化为具有实操性的规则，并建立算法备案、内容标识、伦理审查等制度，将宏观监管要求落实至具体责任主体。这种治理模式通过法律基础为企业提供基础的合规边界，避免治理规则完全依赖企业自主判断。但这并非以单一法律文件确定所有监管要求，而是根据技术发展和风险变化不断调整治理工具，尝试在规则确定性和制度灵活性之间寻求动态平衡。这种治理模式一定程度上降低了单纯依赖刚性法律可能产生的监管滞后，也有效避免完全依靠软法所导致的约束不足。在未来治理过程中，不同层级和不同应用场景下多重政策之间的衔接配合，也是提升国家整体治理能力的重要方向。

5.3. 风险治理：风险关注点从技术安全拓展至社会安全

5.3.1. 风险治理范围：从技术安全向国家、社会与个人安全扩展

中美欧在人工智能风险治理中，不仅重点关注模型幻觉、数据安全、算法偏见、模型漏洞等技术内生风险，还将隐私侵害、信息操纵、就业冲击和人工智能武器化等应用外溢风险纳入治理范围。但是，由于对人工智能潜在风险的判断标准和治理观念不同，各国在风险治理的重点上存在明显差异。美国侧重国家安全领域的前沿风险，欧盟侧重个人基本权利和社会秩序风险，中国兼顾技术安全与社会治理。

美国的风险治理重点关注前沿模型对国家安全造成的影响，强调防范人工智能能力向网络攻击、生物风险和军事用途等领域外溢。拜登政府时期，第 14110 号行政令已经要求前沿模型开发者开展安全测试并向政府报告测试结果，同时 NIST 针对双用途基础模型发布了风险管理指南，将生物武器、网络攻击等潜在滥用行为纳入模型风险评估范围。特朗普第二任期虽然整体放松了监管要求，但并未取消对高风险人工智能领域的关注，而是将监管资源集中至前沿模型、网络安全和国家安全领域。第 14409 号行政令提出建立前沿模型自愿安全评估机制，并要求政府在模型发布前获得最长 30 天的测试访问权限。此外，美国持续通过网络漏洞管理、关键基础设施保护和军事人工智能政府采购等措施降低人工智能对国家安全的潜在威胁。

欧盟更加重视人工智能对个人基本权利和社会运行秩序的潜在影响，安全治理边界由技术安全进一步向公平、隐私和消费者保护等领域延伸。《人工智能法案》把生物识别、医疗、教育、招聘、信贷、执法、司法等与个人权益密切相关的应用场景归入重点治理范围，考察人工智能应用是否会对个人和社会造成不可接受的影响。高风险系统和通用人工智能模型的系统性风险均为欧盟重点关注的监管对象，在治理过程中不仅关注模型本身的技术可靠性，还要求开展一系列合规工作，并通过“基本权利影响评估”事前预判人工智能对个人权利的实际影响。此外，还通过《欧洲委员会人工智能与人权、民主及法治框架公约》和《人工智能法案》中的人工智能素养要求等机制，将算法偏见矫正、隐私保护和非歧视要求纳入监管体系之中。

中国在风险治理中兼顾技术安全和社会治理，风险识别范围由模型和数据安全逐步拓展至网络、现实、认知和伦理等应用风险，并重点关注人工智能对内容生态和社会秩序的外溢影响。中国人工智能风险治理经历了从算法推荐、深度合成到生成式人工智能服务的演进，监管对象由算法运行发展至生成内容和模型服务，并通过算法备案、内容标识等制度加强事前管理。2024 年《人工智能安全治理框架》1.0 版将风险划分为内生安全风险和应用安全风险，覆盖模型算法、数据、系统以及网络域、现实域、认知域和伦理域，推动风险治理从单纯的技术安全向综合应用安全扩展。2025 年 2.0 版增加模型开源、低质有害信息扩散等新型风险，并提出技术内生、技术应用、应用衍生三类核心安全风险和五级风险分级机制。同时，中国通过《数据安全法》《个人信息保护法》等基础性法律，以及生成合成内容标识、科技伦理审查和专项治理等措施，将隐私保护、内容真实和可溯源性、算法偏见和伦理风险纳入治理范围。2026 年“清朗·整治 AI 应用乱象”将进一步将训练语料、数据投毒、生成内容标识、开源模型安全以及虚假信息、低俗内容、身份冒用等一系列问题纳入治理。

5.3.2. 风险治理工具：覆盖事前介入、过程监管、事后处置的全链条差异化治理

随着智能体和自主系统的持续发展，以及人工智能与生物、网络、机器人等技术的进一步融合，人工智能风险形态正在逐步扩大，仅靠技术标准和风险评估难以覆盖全部风险。中美欧的风险治理手段逐步延伸至对事前准入的干预、运行过程中的实时监管和事后的执法处置，形成贯通全生命周期的治理工具。但是，三方实际上在风险干预方面的治理路径上形成了不同的选择。

美国在民用领域整体仍以事后执法为基础，但针对前沿模型和国家安全领域则强调前置安全测试和风险评估。在金融和消费者保护等民用场景，美国依托传统监管机构在既有法律框架下开展执法，人工智能风险通过企业信息披露、虚假宣传、算法歧视等问题在事后进行追责。医疗领域虽已将人工智能医疗器械的风险管理延伸至设计、验证和上市后检测，但主要仍沿用传统行业监管体系。相比之下，针对前沿模型和国家安全领域的潜在风险，美国则持续强化事前安全测试和风险评估。拜登政府曾要求达到一定算力阈值的前沿模型在公开发布前向政府报告安全测试结果，特朗普政府虽然取消了该强制要求，但又通过 CAISI 与企业开展模型发布前的独立评估，并通过第 14409 号行政令强化前沿模型的自愿性前置测试。NIST 针对生成式人工智能和双用途基础模型，进一步补充幻觉、隐私泄露、对抗攻击、生物和网络滥用等风险的具体评估方法。同时，美国通过前沿模型企业安全承诺、人工智能安全研究所联盟等机制推动企业自愿开展红队测试、第三方评估和安全信息共享。通过事后执法和自愿性合规的模式，能够降低普遍性合规要求对技术创新的约束，但安全治理较多依靠企业的主动参与，制度的稳定性和风险治理效果相对有限。

欧盟以高风险系统上市前评估与运行阶段持续监管为治理重心，通过风险分级对重点领域实施全流程治理。《人工智能法案》要求高风险人工智能系统在进入市场或投入使用前履行风险管理、数据治理、技术文档、日志记录、人类监督以及准确性、鲁棒性和网络安全义务，并通过合格评定机制确认其符合监管要求。对于公

共机构、公共服务以及信贷评分、人寿和健康保险等特定高风险应用，还要求在首次公开前开展基本权利影响评估。进入运行阶段后，高风险系统需要持续记录和检测相关风险，通用人工智能系统性风险还需要开展对抗性测试、风险缓释并报告重大安全事件。此外，成员国市场监管机构需要对模型运行开展持续性的监督和执法工作。欧盟也通过监管沙盒和实景测试为新技术提供测试环境，并尝试推出协调标准以降低企业合规的不确定性。这种模式能够提前预防高风险系统规模化应用的潜在风险，但是其治理效果高度依赖技术标准的制定速度和成员国的监管执行能力。

中国在风险治理中侧重事前准入环节监管，同时兼顾事中动态监管和事后处置，通过分类分级、技术安全标准、备案制度、内容标识和伦理审查形成覆盖全生命周期治理机制。在事前阶段，中国针对不同技术形态分别建立算法备案、生成式人工智能安全评估和科技伦理审查等机制，对达到一定风险等级或涉及敏感领域的人工智能活动提前设置准入门槛。在事中阶段，《人工智能安全治理框架》2.0明确将安全要求嵌入设计、开发、测试、部署和运行的全过程，在部署环节开展软硬件安全检测和漏洞修补，在运行环节强化应用安全管理和人工复核，并根据实践需要动态调整治理措施。在事后阶段，则通过风险监测、应急响应、内容标识、投诉举报和专项执法等方式，对数据滥用、深度伪造和虚假信息等风险进行持续性处置。《“十五五”规划纲要》进一步提出健全算法备案、透明度管理、安全评估以及覆盖安全检测、风险预警和应急响应的全生命周期风险管理制度。这一系列措施均表明中国正在向覆盖全生命周期的动态化风险治理模式转型。

5.4. 产业发展：通过完善创新要素配置建设人工智能产业生态

人工智能治理的终极目标是要推动产业健康、可持续地发展，而其中算力、数据、资本、科研、人才等创新要素的合理配置起到关键作用。中美欧的人工智能治理政策不仅规范技术发展，更旨在为产业生态建设提供基础保障，三方均在资源配置和产业组织上选择了契合自己市场条件的路径。

美国人工智能产业发展以科技企业和私人资本为主要投入主体，联邦政府通过公共资源补充科研预算、算力资源和能源基础设施等方面的建设，强化在前沿技术和基础设施领域的优势。美国的大型科技企业是算力采购、前沿模型研发和数据中心建设的核心力量，包括英伟达、亚马逊、OpenAI 在内的多家企业已宣布投入数千亿美元级别资金用于人工智能基础设施建设。其中“星际之门”等项目获得政府政策支持，但主要资金仍来源于企业资本及相关融资安排。政府的投入则更侧重于降低约束创新的外部成本：一方面通过 NSF 启动国家人工智能研究资源（NAIRR）试点，为科研人员提供共享算力与数据资源，并推出“创世纪计划”整合国家实验室与高校科研数据，搭建公共科研算力平台；另一方面围绕算力需求促进能源体系配套改革，简化大型数据中心审批流程，推动电网适应性改造，协调算力扩张与居民用电权益。在资金方面，联邦政府通过年度预算向基础研究、国防应用和关键部门能力建设倾斜，依托国家科学基金会与产业界共建研究设施，希望通过公私合作撬动社会资本参与基础研究。这种产业发展模式能够快速集中优势资源助力技术发展，但同样较为依赖头部科技企业、私人资本的持续投入以及算力和能源供给的可持续性。

欧盟产业政策借助公共资金和跨国协调的基础设施投资，弥补在算力、科研资源和技术能力方面的不足，扩大人工智能应用能力，尤其重视中小企业的发展。在前沿算力方面，欧盟通过 EuroHPC 建设人工智能工厂网络，将超级计算资源、数据、人才和产业用户连接起来，面向初创企业、中小企业和科研机构开放计算资源。2025 年启动的 InvestAI 计划撬动约 2000 亿欧元人工智能投资，其中包括用于建设人工智能超级工厂的专项资金，并通过欧洲投资银行等机构吸引私人资本参与。此外，欧盟通过数据联盟、RAISE 等平台整合分散的数据、算力和科研资源，降低企业和科研机构获取人工智能研发要素的成本。《应用人工智能战略》将政策重心由前沿技术研发延伸至制造、医疗、能源等重点产业应用，重点支持中小企业

采用人工智能，并推动公共领域采购欧洲本土技术与开源方案。《芯片法案 2.0》和《云与人工智能发展法案》则试图在半导体供应和云基础设施领域减少对外部技术的依赖。目前，欧盟在基础模型和先进算力领域仍较大程度依赖外部企业，正试图通过公共投资补足基础能力，并借助跨国平台整合创新资源，以期在算力和基础模型等关键环节缩小与中美差距。

中国由国家统一政策统筹和公共资源配置将算力、数据、技术、资本和应用场景连接起来，推动人工智能产业从基础要素到应用落地的系统性建设。在基础能力方面，中国持续推进智算中心和全国一体化算力网络建设，实施算力互联互通行动，并通过绿电直供、储能配套和电网协同等措施优化能源保障。在数据资源方面，政策要求加快可信数据空间建设，推动数据标注产业发展并实施“模数共振”行动，面向制造、医疗、能源等重点领域构建高质量行业数据集。在技术生态方面，聚焦芯片、算法和基础软件等根技术攻关，实现模型、芯片与云平台企业的软硬件适配和协同创新。同时积极支持开源社区建设和国产大模型开放共享，促进人工智能由单点技术竞争向开放共享的公共资源发展。在资金投入方面，设立国家人工智能产业基金、央企战略性新兴产业基金，启动万亿级国家创业投资引导基金，并设立国家并购基金，形成覆盖初创项目到产业整合的多层次资金支持体系。在具体落地方面，《关于深入实施“人工智能+”行动的意见》围绕制造、医疗、能源、民航、教育、信息通信等重点行业出台专项政策，推动人工智能与实体经济深度融合。中国产业发展路径具有系统性部署特征，能够利用完整的支撑链条和丰富的应用场景推进技术迭代，而政策资源的多层次配置，也对算力利用率、资源配置效率和市场主体创新活力提出了更高要求。

5.5. 全球人工智能治理趋势

5.5.1. 国际竞争：形成技术优势、规则输出和产业生态建设的综合竞争模式

人工智能现已成为影响产业竞争和国家安全的重要技术，治理规则逐步成为国际竞争的全新要素，各国试图通过规则制定、标准输出和产业生态建设提升全球影响力。人工智能治理已经不再局限于国内技术监管，各国既希望通过治理降低潜在风险、促进技术研发并平衡合规成本，也试图借助规则制定和治理模式输出提升国际话语权，提升本国产业在全球布局中的优势。未来人工智能国际竞争不会再是单一的技术能力竞争，而是涉及技术能力、治理规则和产业生态相互协同的综合竞争。

美国依托领先的技术能力、标准体系和产业链部署优势塑造全球人工智能治理影响力，通过掌控关键技术节点和全栈技术部署来巩固自身技术优势。美国的全球影响力并非来自于统一监管规则，而是其在先进芯片和基础模型等全栈技术领域的领先地位。美国以 NIST 人工智能风险管理框架、双用途基础模型风险管理指南等标准体系为基础，推动盟友采用相近的风险评估方法。在产业链的关键技术环节，美国对先进 AI 芯片和模型权重实施分级出口管制，并且限制本国主体对特定国家高性能人工智能系统和尖端半导体领域的投资。此外，美国还积极推进全栈技术出口，建立人工智能出口计划，向伙伴国家输出涵盖芯片、数据中心和模型的完整技术体系，借助芯片供给和基础设施运营来强化与盟友的技术绑定。通过 G7 广岛人工智能进程、人工智能安全研究所国际网络等多边机制，美国试图构建以自身技术体系和规则理念为基础的国际治理生态。

欧盟主要通过规则外溢推动治理模式输出，以《人工智能法案》建立全球范围内具有影响力的风险治理框架。欧盟在基础模型研发和算力资源方面相对弱于美国，因此更多依靠市场规模优势输出监管规则以形成全球影响力。《人工智能法案》提出风险分级、透明度要求、高风险系统义务以及通用人工智能监管要求，使得人工智能企业进入欧洲市场面临制度化的合规要求。由于欧盟单一市场规模较大，相关规则可能促使跨国企业在全业务中遵循欧盟更高标准的治理规则，以减少重复合规要求带来的额外成本，进而产生类似于数据保护领域“布鲁塞尔效应”的规则外溢效应。

中国不断完善国内治理体系以提升国际适应能力，并主动参与全球人工智能治理规则讨论，推动形成开放共建的国际秩序。中国并非单纯输出某一套监管模式，而是综合形成涵盖算法治理、生成式人工智能监管、安全治理框架和产业标准体系建设等多维度的治理体系，为在国际规则讨论中提供中国方案打下制度基础。中国主动提出《全球人工智能治理倡议》和《人工智能能力建设普惠计划》，聚焦弥合数字和智能鸿沟，通过南北合作、南南合作和三方合作帮助发展中国家提升人工智能发展能力。中国还积极推动开源开放与共建共享，支持人工智能开源社区建设，将开源生态当作降低技术门槛和扩大全球参与的重要路径。此外，中国发起成立世界人工智能合作组织（WAICO），打造共商共建的多边合作平台，深度参与联合国、G20、金砖国家、亚太经合组织等多边机制，推动人工智能成为全球公共产品，促进技术开源和能力进步。由此，中国尝试以国内治理经验为基础，通过包容的多边合作和开源共享，推动构建公正合理的全球人工智能治理体系，提升自身国际规则制定中的话语权。

5.5.2. 全球治理存在规则碎片化特征，增加跨境协调成本

全球层面人工智能监管规则的碎片化问题依然突出。中美欧各方在人工智能需要安全、可靠和可控地发展这一目标上已经达成普遍共识，也尝试在模型测试、风险评估、术语标准制定等领域开展务实合作，但各国在治理优先级、监管约束强度和竞争目标上存在一定差异。

治理理念的深层差异是造成规则碎片化的根本原因。美国借助行业标准、自愿承诺和事后执法机制，形成创新优先的轻合规治理模式以释放技术发展活力；欧盟建立强制性风险分级监管体系，将基本权利保护嵌入治理全过程；中国采取以分类分级和动态调整为特征的治理思路，根据技术演进和应用场景实施差异化监管，统筹发展与安全。

这种规则差异直接导致了企业跨境合规成本的上升。对于跨国运营的人工智能企业而言，同一模型需要同时适应不同法域的市场准入条件、数据处理规则、内容标识要求和风险评估程序。监管要求差异不仅抬高了企业重复的合规投入，也增加了技术产品全球部署在法律层面的不确定性。目前在风险分类术语、风险缓释机制等技术操作层面，各国相继通过盟友体系和多边机制加强协调，但在产业发展策略和国家安全边界等涉及深层利益的议题上，长期差距难以弥合。因此，如何在尊重各国治理主体性的前提下降低跨境制度摩擦，将成为全球人工智能治理面临的重大挑战。

5.5.3. 治理模式由静态规则制定向敏捷动态监管转变，以适应技术的快速迭代

人工智能技术的快速迭代使得传统监管体系中的事前立法和固定标准面临巨大挑战，未来治理体系需要更加重视规则的动态调整和全过程的持续评估能力。相比于传统产业，人工智能的技术模型、应用场景乃至风险类型都经历快速发展，监管规则的制定流程可能导致时滞问题。因此，未来人工智能治理重点将从制定静态规则转向完善持续监测、动态评估和快速响应机制。

目前，各国正在探索更加动态灵活的监管工具。美国通过 NIST 风险管理框架、红队测试和第三方评估形成持续性的风险识别机制；欧盟建立监管沙盒和实景测试，为企业提供在受控环境下验证技术应用的机会；中国则随着技术迭代更新监管对象，通过备案管理、分类分级监管、内容标识和伦理审查形成覆盖全生命周期的敏捷治理体系。这些机制并非在确定风险后再制定相应规则，而是伴随着技术发展持续调整监管规则。此外，由于人工智能模型的复杂性和跨领域程度不断提升，仅靠政府监管部门难以全面评估风险，未来需要依靠更多的独立评估机构、科研机构 and 产业联盟配合开展模型安全测试和风险评估，形成更加专业可靠的监管模式。

5.5.4. 人工智能治理将持续关注产业竞争与社会影响的平衡

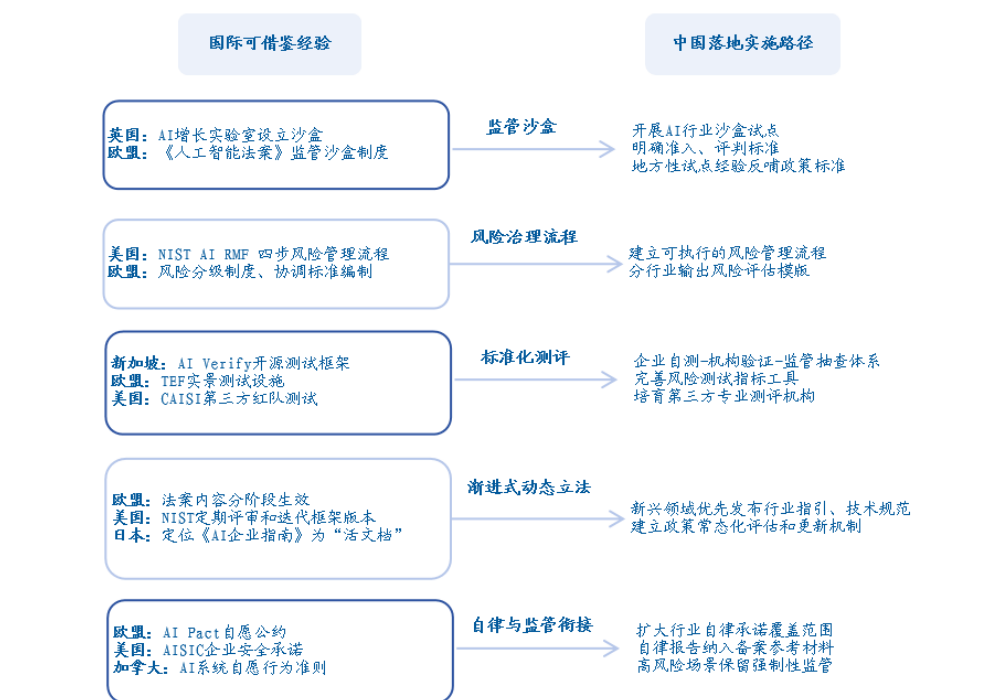
安全监管与创新激励的动态平衡，是全球人工智能治理长期演进的核心主线。早期人工智能治理针对模型安全、数据保护、内容幻觉等技术内生风险建立了完整的风险评估和缓释体系，但随着人工智能渗透至教育、就业、医疗和公共服务等领域，技术应用对社会公平和经济发展的外溢影响逐步显现。目前全球人工智能治理正进入新的阶段，安全仍是治理的基础要求，但更加重要的是实现技术创新与产业可持续发展，以及社会价值之间的平衡。过度强调风险控制可能增加企业研发成本，抑制技术发展速度；缺乏有效治理可能导致数据滥用、内容幻觉等技术问题，甚至蔓延至就业冲击、社会信任下降等社会风险。因此，未来监管需要根据应用场景和风险等级动态调整干预程度，建立与风险程度相匹配的合规义务，提升监管资源利用效率。在风险监管中，需要尤其关注技术发展造成的社会影响。例如，人工智能在提高生产效率的同时，也可能重塑就业结构、加剧数字鸿沟；对于公共服务的优化，也可能带来算法偏见并扩大机会不平等。因此，人工智能治理的评价标准应不再局限于系统安全，而是将公平正义、民生福祉和公共利益纳入政策考量之中，聚焦人工智能技术发展所带来的社会价值。

6. 治理启示：国际经验与国内展望

6.1. 借鉴国际经验：完善人工智能治理工具体系

从全球人工智能治理实践来看，美国、欧盟、新加坡、英国、日本、加拿大等经济体虽然在监管理念和制度安排上存在一定差异，但均在风险管理、技术测试和政策迭代等方面形成了一批可供借鉴的治理工具。我国目前已经在分类分级、算法备案、安全评估、内容标识、伦理审查和行业监管等方面建立起较为成熟的治理体系，未来政策或可在现有制度基础上吸收国际经验，进一步强化风险量化评估、技术测试、政策动态调整和行业自律等环节，推动人工智能治理由制度框架向更加细化的治理工具转变。

图15：国际 AI 治理经验借鉴与我国路径对照图



数据来源：英国 FCA 金融行为局、英国 DSIT 科学创新技术部、欧盟委员会官网、美国 NIST、新加坡 AI Verify 基金会、日本 METI 经济产业省、加拿大 ISDE 创新科学与经济发展部、中国国家认监委等，国泰海通证券研究

6.1.1. 建立监管沙盒机制，为人工智能监管规则提供试错空间

面对人工智能技术迭代快于监管规则调整的情况，监管沙盒能够为企业提供低成本、低风险的试错空间，通过真实应用验证监管规则的适用性。英国于 2015 年在金融科技领域提出监管沙盒的概念，后续不断拓展适用范围，英国信息专员办公室（ICO）尝试在人工智能治理方面运用监管沙盒。2025 年 10 月英国科学、创新与技术部（DSIT）就人工智能增长实验室（AI Growth Lab）发布政策征求意见，拟在医疗、专业服务、交通、先进制造等领域设立国家跨行业人工智能沙盒，允许在严守不可突破红线的前提下限时调整部分监管规则，于真实环境中测试人工智能产品。欧盟在《人工智能法案》基础上进一步完善沙盒制度，既要求成员国设立本国人工智能监管沙盒，还计划 2028 年设立欧盟层级统一的人工智能监管沙盒，依托通用人工智能模型搭建的系统由欧盟人工智能办公室集中监管，降低不同成员国间的协调成本。

结合我国在金融科技领域借鉴英国沙盒模式完成本土化改造的经验，将该制度迁移至人工智能领域具备现实可行性。央行于 2019 年披露的《金融科技（FinTech）发展规划（2019-2021 年）》提出建立健全试错容错机制，并支持北京率先开展金融科技监管创新试点落地，立足现实国情探索我国“监管沙盒”落地的相关制度设计。面向人工智能领域，一种可能的方向是由国家数据局协调推进，在金融、医疗、消费等行业监管体系相对成熟的领域开展地方性试点，明确企业参加沙盒测试的准入标准和测试过程中的评判标准。在控制风险的前提下，可考虑允许企业与监管部门共同验证数据使用、责任划分和安全要求等规则，并将试点结果反馈至后续政策和标准的制定过程。2026 年 4 月 23 日，市场监管总局正式发布《关于对检验检测机构资质认定实施“一单一库”管理的公告》，明确提出建立“沙盒监管”制度，在服务战略性新兴产业、未来发展需求的领域构建包容审慎治理模式，为监管沙盒向人工智能领域的拓展提供制度基础。

6.1.2. 将原则性风险治理要求转化为可落地的风险管理流程

原则性的风险治理规则需要转化为企业可量化、可执行的具体操作标准，才能够有效推进治理目标落地。欧美通过标准化风险管理框架与风险分级体系，逐步实现治理原则向企业合规行动的传导。美国 NIST 发布的人工智能风险管理框架（AI RMF 1.0），以“治理、识别、测量、管理”四项功能，为企业提供风险识别、评估与缓释的完整流程。同时，配套操作手册、生成式 AI 专项指南、Dioptra 开源测试平台，以自愿性工具帮助各类企业落实全周期风险管控要求。欧盟依托《人工智能法案》确立四级风险分级制度，以应用场景为核心划分风险等级，对高风险系统设置差异化合规义务。此外，推动 CEN/CENELEC 协调标准编制，搭建 AI Act 服务台、合规检查工具，将法律条文转化为可操作的合规流程。

未来或可在现有人工智能安全治理框架基础上，建立可量化、可复用的风险管理工具，连接起顶层原则到企业实操的全过程。对于模型幻觉、数据泄露、算法偏差、隐私保护、内容安全等重点风险，可探索形成统一的执行步骤和操作标准，在政策的指引下，将内生、应用和衍生风险的原则性要求拆解为可执行的具体风险管理流程。针对金融、医疗、制造、教育等重点行业，可研究制定差异化风险评估模板，指导企业基于应用场景开展风险识别和持续检测。同时，进一步有效衔接国家标准、行业标准和企业内部风险管理工具，降低企业理解和执行监管要求的成本。此外，可参考欧盟以应用场景为核心的分级逻辑，逐步完善我国五级风险分级的判定细则与对应义务清单，实现风险分级形成可核查的实操制度，在安全可控的前提下为创新留出弹性空间。

6.1.3. 建立健全人工智能标准化测试评价体系

国际上通过标准化测试工具、第三方合格评定和行业场景化测试设施，弥合制度文本及落地执行之间的差距，解决仅依靠企业自评容易出现的公信力问题。新加坡资讯通信媒体发展局（IMDA）推出的 AI Verify 是全球首个由政府主导并开源的人工智能治理测试框架，企业可在本地进行部署，围绕公平性、可解释性、稳健性和数据治理等维度开展标准化测试。2025 年 2 月，启动 Global AI Assurance Pilot，将 16 家专业 AI 测试机构与 17 家生成式 AI 应用部署企业配对，覆盖医疗、金融、人力资源等 10 个行业，在真实业务场景中验证模型可靠性并提供标准化测试指南。欧盟依托《人工智能法案》，针对高风险人工智能系统设置合格评定机制。同时建设 TEF 行业测试设施，提供贴近真实业务的测试环境，该设施不授予企业规则豁免，而是仅用于完成技术性能与风险验证。美国人工智能标准与创新中心（CAISI）推行前沿模型部署前的第三方安全评估，配套红队测试等技术手段，强化模型安全风险的客观验证。

未来可考虑立足现有算法备案和安全评估制度，建设企业自测、机构验证、监管抽查相结合的治理格局，补齐技术保证和场景化测试能力。针对基础模型、生成式人工智能、智能体和行业应用，围绕模型幻觉、数据泄露、智能体自主决策等典型风险，分别制定测试指标和测试工具，将原则性安全要求转化为可复现的测试项目。可借鉴国际第三方测评实践经验，鼓励具有资质的科研机构、检测认证机构和行业组织承接具体的测试技术服务，提升安全测试的专业化水准。在此过程中，政府承担测试标准的统一制定、服务机构的资质管理和测试结果的监督抽查等职能，提升安全评估的客观性和公信力。

6.1.4. 采用渐进式立法和敏捷监管的治理逻辑

人工智能技术的发展速度显著快于传统立法周期，仅依靠单次立法容易出现规则滞后的问题，此外治理规则的一次性大幅调整也容易导致过高合规成本。国际上形成了渐进立法和柔性文件动态迭代的敏捷治理实践，在法律稳定性和技术适配性之间寻求平衡。英国采取“先协调、后法定”的渐进路径，先以跨部门非法定原则框架推动治理，设置一定评估期，随后根据实施效果再决定是否将特定要求上升为法定义务。欧盟《人工智能法案》采用分阶段生效实施的制度设计，针对禁止类规则、高风险系统合规义务和通用大模型义务设置差异化的生效时间点，配套授权立法和定期评估修订机制，允许依托实践反馈来调整高风险场景清单和技术配套规则，避免全部规则一次性落地带来的适配压力。美国 NIST《人工智能风险管理框架》建立了定期评审和版本迭代机制，依托产业反馈持续优化柔性指引文件。日本则通过经济产业省（METI）与总务省（MIC）联合发布《AI 企业指南》（AI Guidelines for Business），通过专家讨论和公众意见征集形成统一的企业指导框架，明确将其定位为“活文档”（Living Document），后续根据技术发展和实践反馈持续更新。

我国已经初步建立起较为灵活的人工智能治理机制，下一阶段可考虑进一步利用政策文件及行业指引，根据技术成熟度和风险程度动态调整监管模式。我国始终遵循敏捷治理思路，《人工智能安全治理框架》完成从 1.0 到 2.0 版本迭代，紧跟智能体、开源模型等新兴技术形态更新风险分类与管控要求。在安全评估和备案制度方面，实现从算法推荐、深度合成内容到生成式人工智能的监管对象转变。针对智能体、具身智能等风险特征尚未明确的新兴领域，可优先通过行业指引和技术规范明确基本治理要求，并依托创新试点和产业实践积累现实案例和风险数据，逐步将经实践检验的治理要求纳入部门规章或法律法规。同时，可研究建立对于已有制度的常态化政策评估和效果检测机制，在涉及法律基本制度和重大权利义务的事项保持稳定的基础上，对于部门规章、规范性文件和行业指引动态调整，提升监管细则对于技术迭代的适应力。

6.1.5. 拓展企业自律机制，推进与政府监管的衔接

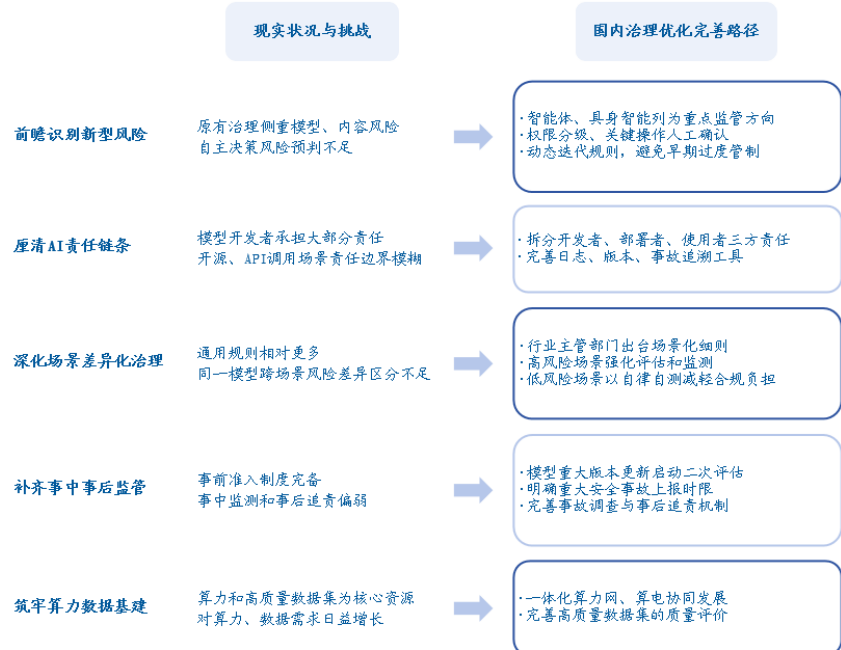
在人工智能新业态持续涌现的背景下，行业自律和企业自愿承诺可作为强制性监管的重要过渡和补充。国际实践将自愿承诺、第三方评估和监管体系相结合，既发挥市场主体能动性，又避免企业自律完全脱离监管约束。欧盟推出《人工智能公约》（AI Pact）自愿倡议，鼓励企业提前对标法案义务开展内部治理，企业自律实践可以作为后续合规举证的参考依据。美国依托 AISIC 行业联盟和白宫召集签署的前沿模型自愿安全承诺机制，引导头部企业开展红队测试和第三方安全评估，通过市场化自律手段弥补成文监管规则的时滞，并为企业保留一定创新空间。加拿大于 2023 年发布《负责任开发和管理先进生成式人工智能系统自愿行为准则》，围绕问责、安全、公平、人类监督等提出企业行动要求，配套红队测试和基准测试等实操举措，定位为正式立法落地前的过渡性制度。

我国已经初步构建政府引导下的行业自律制度，未来或可进一步扩展行业自律承诺的覆盖范围，打通和第三方测试、政府监管的有效衔接。国内已在行业自律方面展开部分实践，例如行业联盟发布《人工智能安全承诺》、智能体领域专项自律公约。立足现有制度基础，可根据风险程度实施差异化制度安排。针对一般风险系统，继续鼓励行业协会、产业联盟出台自律公约及安全承诺，引导企业完善内部风险管控机制。针对高风险场景，不弱化强制性监管义务，而是考虑将企业提交的自律风险报告、第三方测试材料作为备案制度和安全评估的重要参考。

6.2. 展望国内治理：立足中国发展实际，补齐新型风险与治理能力短板

在吸收国际治理工具经验的基础上，我国人工智能治理可考虑立足本国制度与产业实际，面向新兴技术补齐治理能力和制度短板，形成具有我国国情特色的治理体系。我国现已建成统筹发展与安全、覆盖范围全面、事前准入机制完善的治理框架，但在新型风险预判、产业链主体责任划分、场景化差异监管、事中事后闭环监管等方面仍可进一步推进。下一阶段，我国可逐步推动已有制度框架走向更加适配人工智能全方位应用的完备形态，建设具有时代特色、中国特色的治理体系。

图16：国内 AI 治理短板与优化路径图



数据来源：中国政府网、中央网信办、国家网信办、国家发改委、工业和信息化部、商务部等，国泰海通证券研究

6.2.1. 前瞻识别新型风险，向智能体和具身智能等自主行为安全延伸

目前，人工智能正从“生成内容”向“自主决策和执行任务”演进，治理重点将逐步由模型输出本身转向人工智能自主行为及其现实后果。当前生成式人工智能的主要风险仍然集中于内容幻觉、算法歧视和隐私泄露等模型和内容层面风险，但智能体能够自主感知、记忆、决策、交互和执行，具身智能进一步将模型决策延伸至物理环境，相关风险将升级为错误决策产生的实际后果。我国 2026 年发布的《智能体规范应用与创新发展实施意见》已经明确提出，对智能体按照应用场景和潜在影响实施分类分级治理，并关注自动化攻击、隐私侵犯、网络诈骗等应用衍生风险。

我国后续可考虑延续将风险随技术演进逐步纳入监管体系的治理思路，将智能体和具身智能作为下一阶段的重点监管对象，实现对于模型行动能力的前瞻性安全约束。对于具有较强自主决策和执行能力的智能体，可侧重自主调用工具、权限管理和人机交互等风险，设置权限分级、关键操作人工确认和异常行为阻断等安全机制，并根据其影响现实世界的能力设置差异化安全标准。总体上，可根据技术演进和风险程度动态调整监管标准，避免在技术尚处于早期阶段时设置过度刚性的统一准入门槛。

6.2.2. 厘清人工智能责任链条，明确开发者、部署者和使用者责任分担机制

人工智能应用链条被日益拉长，单纯将治理责任集中于模型开发企业可能难以覆盖实际风险，未来治理需要进一步明确开发者、部署者和使用者之间的责任边界。传统生成式人工智能治理主要围绕模型开发者和提供者展开，需完成科技伦理审查、算法备案、安全评估报告、内容标识等多项合规义务。但随着开源模型、API 调用和应用场景快速发展，同一基础模型可能被不同企业部署在金融、医疗、教育等多个应用场景中。如果仍由基础模型提供者承担全部责任，可能弱化实际应用企业的风险管理责任，无法覆盖风险产生的全流程。

我国可以在“谁提供、谁负责”的治理基础上，进一步形成与应用场景相匹配的责任分配机制，将模型开发、系统部署和具体使用三个环节的责任在不同主体之间进行拆分。基础模型开发者主要承担训练数据合规、漏洞响应和内生风险管理等责任；应用开发和部署企业则根据实际场景承担权限管理和场景适配等责任，并继续履行包括算法备案、安全评估和内容标识在内的法定合规义务；最终使用主体对实际决策结果承担主体责任，避免将人工智能生成结果直接等同于具有法律效力的最终判断。2026 年 2 月 6 日，国家发改委等部门印发《关于加快招标投标领域人工智能推广应用的实施意见》明确提出，人工智能生成结论不得替代相关主体自主判断，也不改变使用主体的法定责任，这可视为未来划分人工智能应用责任边界的一种实践方向。同时，可进一步探索与责任划分相匹配的日志留存、版本记录、操作记录和风险事件追溯机制，使得发生问题后能识别到具体的责任环节。

6.2.3. 聚焦高风险应用场景，进一步深化行业和场景差异化治理

随着“人工智能+”行动向金融、医疗、教育、制造等领域的深度渗透，落地应用场景在更大程度上决定了人工智能的实际风险，治理不仅需要评判模型安全，还需要关注技术在不同场景下产生的现实影响。相同的基础模型应用于不同场景，其风险等级和后果危害都会存在显著差异。例如，金融领域风险集中于自动交易和信贷评估带来的财产权益风险，医疗更加关注诊疗安全和责任认定，公共决策类应用关乎公共资源分配与公民利益。仅仅依靠通用化的模型风险评估，难以充分反映人工智能进入不同产业后的真实风险。

我国已经具备分行业专项治理的政策实践基础，未来可考虑依托现有行业监管体系，由各行业主管部门细化适配业务特征的场景化监管细则。我国《“十五五”规划纲要》提出统筹通用大模型与行业专用模型发展，面向能源、交通、制造、教育、医疗、金融等领域布局行业高质量数据集。这些领域也紧随“人工智能+”行动部署要求，相继出台专项政策落实针对应用场景的原则性监管要求。未来，可逐步发挥行业传统监管机构的现有监管资源和执法经验，将人工智能治理嵌入既有的行业监管框架，实现监管强度与具体应用场景相匹配。在保障国家整体规则统一的基础上，由各个行业监管部门牵头，根据应用行业的风险特征制定针对性、可落实的监管细则。例如，面向信贷、医疗、就业等高影响场景，强化事前评估、人工复核、操作日志等硬性规则；面向制造、交通等涉及物理安全的场景，完善产品准入测试、运行状态监测和应急处置机制；面向普通办公、日常消费等低风险场景，则以企业自测、行业自律和信息披露为主要抓手，避免导致过高合规成本。

6.2.4. 事前监管基础上，同步关注持续监测和事故追责等事中、事后监管措施

我国人工智能在事前准入环节会经历全面的审查措施，但是运行过程中的风险监测和事后的责任追溯也十分重要。我国已经通过算法与大模型备案制度、科技伦理前置审查、安全评估报告、数据与个人信息保护的前置约束等制度解决人工智能系统上线前的合规问题，形成覆盖领域全面、考量维度多样的准入门槛体系。但是模型在运行过程中可能随着应用场景和实际情况的变化进行模型更新或数据调整，仅依靠上线前的评估和审查难以覆盖人工智能运行期间不断变化的风险。

我国可结合现有事前监管体系的治理经验，进一步补齐事中和事后环节的监管要求。事中环节我国已经建立了内容标识、分类分级、日志留存和投诉举报等监管制度，事后环节则主要依托“清朗”系列专项整治和国家安全审查，但是在持续监测、事故报告和应急处置方面相比于事前监管仍然存在明显短板。对于投入运行的人工智能系统，要进行覆盖全过程的持续监测，在版本发生重大更新时及时启动二次评估义务。针对造成人身伤害、重大财产损失或大规模社会影响的重大事件，要明确企业上报时限和事件溯源材料报送要求。同时，可逐步完善事后责任链条机制，将日志留存和版本溯源记录作为事故调查与责任认定的重要依据，实现与行业处罚、事后追责的有效衔接。

6.2.5. 统筹算力与数据资源，筑牢人工智能发展的基础设施根基

算力和数据正在成为人工智能规模化应用的重要基础资源，其供给规模和使用方式也成为治理需要密切关注的问题。算力是大模型训练与运行的硬件底座，决定了模型迭代速度和推理响应效率。高质量数据则是模型学习规律并形成能力的核心原料，直接影响人工智能模型的准确性与场景适配水平。我国人工智能发展具有市场规模较大、应用场景丰富的特征，未来随着模型研发和训练需求持续扩张，对于算力布局和数据供给的要求也会日益增加。因此，治理的重点不能仅停留在模型和应用本身，还需要关注支撑人工智能发展的基础设施是否高效、安全和可持续，以确保技术发展有充足的保障条件。

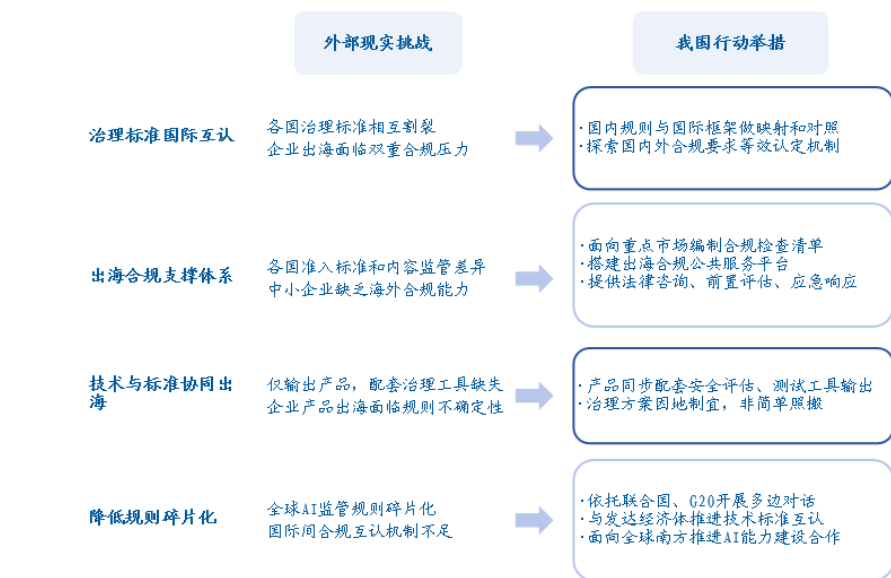
未来可考虑在现有算力基础设施布局的基础上，进一步完善算力资源的监测、调度和安全管理机制，推动算力设施一体化、集约化和绿色化发展。《“十五五”规划纲要》已经提出推进全国一体化算力网、发展标准化可拓展智算云服务、推动绿色电力与算力协同布局，国务院《关于深入实施“人工智能+”行动的意见》也明确强化智能算力统筹调度。未来的算力发展可将政府顶层设计和市场驱动相结合，推进算力设施全国一体化发展，将“算电协同”上升为战略高度的同时，引导企业积极探索算力的高效实用和有效配置方法。数据治理则可继续强调高质量供给与合理利用兼顾的发展模式。《“十五五”规划纲要》提出建设国家数据资源体系，完善数据标准和质量管理体系，并面向能源、交通、金融等领域建设高质量数据集。中国互联网规模庞大、工业门类齐全，随着人工智能在各个领域的

应用落地，有机会将数据资源发展为全新优势。未来可以在数据分类分级和安全保护之外，进一步完善公共数据权限管理、行业数据开发利用和数据质量评价机制，通过高质量数据集的供给赋能行业模型发展。

6.3. 展望全球治理：推动治理规则国际化，构建企业出海支撑体系

当前全球人工智能治理呈现出规则碎片化特征，美欧依托各自技术与市场优势向外输出治理范式，跨国企业跨境重复合规成本突出。我国可考虑立足国内成熟的治理实践，统筹在规则体系中的话语权和出海现实需求。一方面推动国内经过实践检验的规则标准对接国际规范，探索合规等效互认；另一方面搭建出海合规公共服务支撑体系，推动技术产品与治理工具协同出海，坚持治理方案与各国国情的适配。依托多边平台开展能力建设合作，弥合全球智能鸿沟，构建包容普惠的全球人工智能治理合作路径。

图17：全球 AI 治理行动参与方向



数据来源：中国政府网、外交部官网、工信部官网、中央网信办官网、中国—东盟人工智能产业创新中心、国家互联网信息办公室等，国泰海通证券研究

6.3.1. 推动国内治理标准向国际标准转化

中国已经建立了较为全面的国内治理规则和技术标准，这些治理工具不仅应作为国内监管手段，还可通过标准映射和互认实现向国际标准的转化。这一问题已经成为欧美推动治理规则国际化的重要着力点，美国通过《人工智能标准全球参与计划》、人工智能安全研究所国际网络等机制，推动风险评估和安全测试方法向盟友及国际组织扩散，欧盟通过协调标准、数字伙伴等方式推动内部规则与国际标准体系衔接。当前我国已经在算法、数据、模型、伦理等多领域有了标准化基础，也在持续推进国内治理要求向技术规范转化。例如，《人工智能可信赖通则》和《人工智能终端智能化分级》均是在国家标准方面的重要实践，制定过程充分参考 ISO/IEC 发布的国际标准，为技术标准的国际衔接提供基础。然而，目前我国企业出海还面临着双重合规要求的困局，同一套模型需要先后满足国内和国际标准会带来重复合规成本。

未来中国参与国际规则制定可考虑将国内已经经过产业实践验证的治理要求转化为国际社会共同认可的技术规则。《人工智能全球治理行动计划》提出加强与 ISO、IEC 等国际标准组织合作，推动安全、产业、伦理等领域标准的制定，增强标准体系的包容性和互操作性。可以推动中国《人工智能安全治理框架》的风险分级逻辑与欧盟《人工智能法案》风险分类、美国 NIST AIRMF 进行技术性映射，

或者加强国内标准与 ISO/IEC、ITU、IEEE 等国际标准体系之间的对照。此外，可探索建立人工智能系统的等效性认定机制，即通过国内合规要求与安全评估的企业，在申请欧盟或其他国家市场准入时，可凭国内合规记录简化部分程序，降低企业出海成本并缩短市场进入周期。

6.3.2. 建立人工智能企业出海合规支撑体系

面对全球人工智能治理的碎片化问题与合规成本的持续上升，国家层面或可为企业提供包括法律咨询、技术认证在内的出海合规支持。中国人工智能企业出海除了技术竞争之外，还需要应对不同国家在数据治理、内容安全、算法透明度、模型测试等不同维度的多重合规要求，对于中小企业会导致合规成本过高的问题。因此，国内不仅可通过政策要求企业自身满足境外规则，更可尝试为企业提供跨境合规的基础设施。其中，可借鉴欧盟较为成熟的企业合规公共服务思路，依托 AIAct 服务台和 AI 单一信息平台，为企业提供免费咨询、合规检查工具和多语言支持。

未来可依托行业主管部门、标准化机构或专业服务机构，建立面向海外市场的人工智能合规公共服务体系。《“人工智能+制造”专项行动实施意见》明确提出支持行业组织、专业机构为企业提供出海配套服务，引导企业高效开展各类技术验证和合规认证，更好服务产业有序出海发展。可围绕欧盟、中东、日韩、北美等重点市场形成实时动态更新的合规指引，将当地关于人工智能产品的准入要求转化为企业可以直接执行的检查清单。还可以建设综合服务平台，为企业提供出海法律咨询和前置评估，帮助企业提前规避出海后被动调整产品的情况，并在企业面临数据跨境限制、模型权重出口审查时启动应急响应。

6.3.3. 推动技术、产品与治理标准配套协同出海

中国企业可考虑在促进技术与产品出海的同时，尝试寻找推动相关治理规则协同输出的新可能。中国人工智能治理的国际竞争优势不仅来源于模型和技术，也来自产业应用过程中积累的治理经验和场景解决方案。若仅输出技术而不同步输出配套治理规则，中国企业可能面临海外国家的规则不确定性。如果能推动技术产品、产业服务和治理标准协同进入海外市场，则能够为中国企业营造熟悉且可预期的治理环境，帮助企业更平稳地落地并适应海外市场。

具体而言，中国企业对外输出智能制造、医疗、金融、机器人和智能终端等产品之外，可同步输出与产品配套的安全评估标准、数据治理方案和风险管理制度，在海外项目落地过程中形成“技术+标准”的组合模式。特别是在全球南方国家，部分国家人工智能治理能力和算力基础设施相对不足，中国可通过技术合作和能力建设帮助合作方建立基础建设能力，在提供技术产品的同时形成相应的治理标准和实践经验。可支持头部企业将分类分级模板、GCmark 标识生成与检测平台、模型安全测试工具、科技伦理审查系统等治理基础设施，以开源软件或云服务的形式向海外部署，使东南亚、中东、非洲等新兴市场监管机构和企业习惯使用中国技术路径开展合规管理。

需要注意的是，治理规则的输出并不等同于简单复制国内的监管模式。由于不同国家在法律制度、产业基础和社会环境等方面存在较大差异，可考虑在安全测试、风险评估等技术性较强领域底层标准共通的基础上，将内容治理、行业准入等具有较强地域属性的治理方案本土化，不断提升中国治理方案的国际兼容性。

6.3.4. 通过标准互认和多边合作降低规则碎片化

面对全球人工智能治理规则碎片化的问题，中国可考虑在参与全球治理的过程中尝试推动不同治理体系之间的规则衔接。《2026 世界人工智能大会暨人工智能全球治理高级别会议主席声明》提出，建立具有广泛共识、科学、透明、包容的标准规范框架，促进各国间标准互认，降低人工智能国际合作的成本和门槛。对此，可继续依托联合国、G20、亚太经合组织、金砖国家以及区域合作机制，推动基础治理规则的对接，针对不同发展阶段国家建立差异化的合作路径。对于发达经济体，可重点推动技术标准和合规认证体系之间的互认；对于全球南方国家，则可结合产业和基础设施条件，提供人工智能能力建设，帮助其参与到全球治理的规则制定之中。

7. 风险提示

7.1. 人工智能产业增长不及预期

当前全球人工智能产业仍处在快速发展阶段，算力、数据、能源等基础设施投入与模型研发成本持续走高，商业化落地存在投入产出周期偏长等挑战。若技术迭代、场景落地或资本投入不及预期，或将拖累全球 AI 产业发展节奏，并影响相关企业盈利预期。

7.2. 人工智能治理规则分化加剧

美国、欧盟和中国在监管理念、治理工具和重点领域上存在差异，随着人工智能技术快速发展，各经济体可能进一步结合自身产业基础和诉求调整治理规则。若不同国家和地区的规则衔接不足，可能加剧跨境合规要求差异，提高人工智能企业国际化经营的合规成本。

7.3. 人工智能技术风险演变超预期

随着人工智能由内容生成向智能体、具身智能等更高自主性形态演进，人工智能风险可能由模型输出风险进一步延伸至自主决策、工具调用和现实环境交互等新兴风险。若现有安全测试、权限控制和持续监测机制无法及时覆盖新型应用场景，可能增加人工智能技术应用中的安全风险。

7.4. 中国人工智能治理政策落地不及预期

我国已形成覆盖算法、生成式人工智能、安全治理、伦理审查等领域的治理框架，但在多主体责任划分、高风险场景精细化监管、持续监测及事后追责等方面仍需进一步完善。若相关政策、标准和治理机制的执行效果不及预期，可能影响人工智能产业规范发展及中国企业参与国际市场竞争。

本公司是本报告所述寒武纪(688256)的做市券商。本报告系本公司分析师根据寒武纪(688256)公开信息所做的独立判断。

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

免责声明

国泰海通证券股份有限公司（以下简称“本公司”）授权客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

本报告版权仅为本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制、发表或引用。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“国泰海通证券研究”，且不得对本报告进行任何有悖原意的引用、删节和修改。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

国泰海通证券研究所

地址 上海市黄浦区中山南路 888 号

邮编 200011

电话 (021) 38676666