DOI: 10.20180/j.qjem.2025.03.04

家长式防控、守夜人看护与伴侣型监管
——全球 AI 治理模式的一个比较分析框架

凌 斌 马润艺

摘 要：全球人工智能（AI）治理的多元实践呈现为三种基本模式："家长式防控"（以欧盟为代表）、"守夜人看护"（以美国为代表）与"伴侣型监管"（在中国和英国均有体现）。通过法律工具与政策导向、行政监管结构、司法制衡机制以及地方实验与权力分配四个维度的比较分析，可以揭示出不同模式在价值取向、制度逻辑及创新－风险平衡策略上的差异。"家长式防控"强调通过强规制保障基本权利，"守夜人看护"侧重宽松环境以激励创新，而"伴侣型监管"则在政府－企业协作中寻求战略引导与有效规制的紧密结合。全球 AI 治理的实践尽管复杂多样，但许多国家和地区都表现出上述三种基本模式的类似特征，因而可以纳入这一分析框架。在 AI 深度嵌入经济社会发展的时代背景下，具备协同导向的"伴侣型监管"正成为全球治理的重要路径，并对政府的专业化能力提出了更高要求。

关键词：人工智能治理；监管机制；创新－风险平衡；家长式防控；守夜人看护；伴侣型监管

中图分类号：D996；D922.1；TP18

JEL 分类号：K23；K20；O33

一、引 言

随着人工智能技术和应用的高速发展，不同国家和地区呈现为各具特色的 AI 治理实践。这些 AI 治理实践可以概括为三种模式："家长式防控""守夜人看护"和"伴侣型监管"。

[**收稿日期**] 2025－05－28

[**作者简介**] 凌斌（通信作者），北京大学法学院，E-mail: lingbin@pku.edu.cn；马润艺，北京大学法学院，E-mail: marunyi@stu.pku.edu.cn。本文初稿由凌斌完成，曾在 2025 年北京大学光华管理学院、宾夕法尼亚大学沃顿中国中心与沃顿负责任 AI 实验室共同主办的"全球视野下的人工智能治理研讨会（Global Perspectives on AI Governance Workshop）"上介绍。感谢与会者的提问和批评，感谢耶鲁中国中心的卡曼·陆凯（Karman Lucero）研究员关于美国 AI 治理特征提供的建议。马润艺补充了重要资料，并参与了全文修改，最终由凌斌定稿。作者感谢苏仁心同学的助研工作，感谢匿名评审专家和编辑部的宝贵意见，当然文责自负。

家长式防控模式以欧盟为代表，其特点是采取预防性的、自上而下的硬性规则，旨在于危害发生前将其消除。这里的“家长式”是指政府像保护性父母一样，通过预先设定硬性规则来防范可能的风险，即使这些规则可能限制个人或企业的投资热情而导致产业发展受到抑制。在 AI 治理语境下，这种模式体现为通过全面而严格的事前规制来防范 AI 可能带来的社会风险。

相比之下，以美国为代表的守夜人看护模式则依赖于一种警惕但放任的立场，主要通过既有规则和机构进行事后监督，而非制定全面而严格的预防性规则。这一监管哲学强调政府作为“守夜人”的角色，在维持基本秩序的同时允许创新自由发展，仅在必要时进行干预。这种模式源于古典自由主义的“守夜人国家”概念，在技术治理领域体现为响应式而非预防式的监管策略。

伴侣型监管模式——以中国和英国为两个各具特色的代表——则意味着一种更为贴近式的、既积极扶持又紧密管控的治理路径。政府将支持性政策和产业合作与有针对性的监管策略相结合，在人工智能发展中扮演着一类特殊的合作伙伴角色。因此，伴侣型监管模式既不同于自上而下的中心化的家长式防控模式，也不同于自下而上的去中心化的守夜人看护模式，而是以一种多管齐下、央地协同的方式实现治理目标。

“伴侣型监管”是本文提出的一个新概念。通过类比“伴侣型婚姻”（Companion Marriage）、“伴侣型动物”（Companion Animal），这一概念描述的是一种特定类型的治理模式：政府部门的角色不仅是“看住”企业，而且给予企业规则引导以及政策乃至资金支持，推动企业致力于经济社会协同发展的国家战略。这正如在伴侣型婚姻中，配偶之间不是一种单向的服从与被服从关系，而是既相互支持又彼此“盯紧”。当然，这并不意味着企业可以和政府平起平坐，所有的“监管”都意味着政府和企业之间的某种不平等乃至支配性关系，伴侣型监管也是如此。正如伴侣型婚姻和伴侣型动物这些概念的着重点不在于双方的平等关系，而是紧密的伴随性特征，“伴侣型监管”在这里强调的也不是某种平等的政企关系，而是监管部门通过专门化的规则（硬法）和政策（软法）制定、现行部门的敏捷治理和事中监管、有效的司法补充以及中央推动的地方实验来规范、引导企业实现某种政企合作关系。

伴侣型动物有很多种类，如从虎皮鹦鹉到拉布拉多。伴侣型婚姻也会有不同类型。就文化而言，英国式配偶对于另一方的“监管”程度通常和中国有所不同。中国式配偶对彼此生活和事业的介入通常更强，“监管”的意愿、范围和力度都会更大。但是就伴侣型婚姻的共同特征而言，两者都表现为通过相互提醒和彼此呵护来实现家庭生活的持续稳定发展。中英两国伴侣型监管的异同也是如此。我们将在后文中看到，两者在硬法与软法的选择以及行政与司法的互补方式方面都有所不同，如同不同家庭中夫妻的性格组合也各具特点，有的是妻子更为活泼，有的是丈夫更

为热情。中国和英国在政治体制、经济发展模式、社会结构、历史文化等各个方面都有着显著差异，存在这样的差异并不奇怪，但耐人寻味、值得研究和深入思考的是，两者竟然在治理模式上呈现出如此多的共同之处。并且，中英两国在与欧美的对比中呈现出的共同点，并不妨碍我们进一步阐释两国之间的彼此差异，进而提炼出伴侣型监管的两个亚型。实际上，我们将在后文中看到，每种 AI 治理模式下各国的实践都有一定差异，但是同样会表现出共同的实践特征。

本文希望在随后的论述中向读者表明：全球 AI 治理模式呈现出某种与最初印象迥异的结构性特征。欧盟尽管一直强调自由、平等为核心的启蒙价值，但在治理模式上反而更为中心化和家长主义；美国尽管由于其联邦体制，州层面的碎片化众所周知，其实联邦层面包括司法判决的碎片化同样严重，正是源自其“守夜人”式的监管哲学；而英国尽管在地理上属于欧洲、文化上与美国同源，但是在监管模式上却与中国更为相近——这或许是源于两个国家都实行单一制，并且政府都具有敏捷而高效的治理能力。因此，尽管多数国家很容易受到“布鲁塞尔效应”的影响而采取家长式防控的 AI 治理模式，但是那些希望在人工智能产业发展中有所作为的国家，更倾向于以中国和英国为代表的伴侣型监管模式——这些国家不仅与中国和英国，而且彼此之间，在政治、经济、社会、文化特别是政企关系方面，都有很多显著的不同。伴侣型监管正在成为一种极具包容性同时又具有鲜明特征的 AI 治理模式。

限于篇幅和主旨，本文无意评价每种模式的优劣。本文的目的在于提供一个理解全球 AI 治理模式的一般性的比较分析框架。这一框架下的每种模式都反映了特定的价值取向：欧盟强调防范潜在风险和保护基本权利，美国偏爱鼓励企业和地方创新并进行选择性干预，而中国和英国则在促进人工智能发展的同时寻求一种政府可控的发展路径。家长式防控、守夜人看护和伴侣型监管作为 AI 治理模式的三个“理想类型”，能够将尽可能多的国家纳入一个统一的分析框架。尽管各国之间的差异数不胜数，类型之间也会相互转化，但是这三种模式各自具有的基本特征仍然有助于我们对纷繁复杂的治理现象做出相对清晰的理论界分。

为了系统地比较这些模式，本文借鉴了比较政治学中的制度分析方法，将从四个治理维度展开分析：①法律工具与政策导向；②监管结构与实施方式；③司法审查与制衡机制；④地方实验与权力分配（March and Olsen，1989；Pierson，2004）。尽管版权等私法规则也具有广义上的 AI 治理作用，但限于篇幅，本文主要聚焦于公法层面。此外，文中将结合一些重点行业案例——包括面部识别、自动驾驶汽车、医疗保健、算法决策以及其他新兴领域——来说明每种模式在具体应用场景中的表现形式。在此基础上，本文把参与全球 AI 治理实践的主要国家和地区纳入特定类型的模式之中，进而提出一个一般性的关于 AI 治理的比较分析框架。

二、法律工具与政策导向

AI 治理模式的典型特征，通常体现为相关的规范来源。这些 AI 治理规范既可能是具有约束力和强制性的硬法，也可能是类似于指南、战略、计划、标准这样的指导性和倡议性的软法（Snyder，1994；Abbott and Snidal，2000）。这些硬法和软法既可能是全面的、统一的规则体系，也可能是专门化的、多领域的规则组合（Baldwin et al.，2011）。正是这些立法和政策层面的差异，将三种 AI 治理模式明确区分开来。

（一）欧盟：基于风险制定全面统一硬法的“家长式防控”

欧盟采取了积极的硬法方案，提出并颁布了全球首部全面的人工智能法规——《人工智能法案》（*AI Act 2024*）（European Union，2024）。欧盟《人工智能法案》是一个统一的横向框架，而非零散的行业法律，旨在统一适用于所有行业和欧盟成员国。

这部具有强制约束力的法规建立了一个统一的、基于风险的分类系统（Hutter，2005；Black，2010），将人工智能应用分为不同等级——从不可接受的风险（完全禁止）到高风险（严格监管）、有限风险（透明度要求）和最小风险（自由使用）。例如，根据该法案第 5 条，被视为对安全或基本权利构成威胁的人工智能系统（如政府进行的社会评分）是被禁止的；高风险人工智能（如在医疗诊断或关键基础设施中）将面临严格的事前合规义务，提供商必须进行合格评定，确保人工监督，并在投放市场前获得认证。更不用说，当欧盟《人工智能法案》与其《通用数据保护条例》（GDPR）和《数字服务法》（DSA）结合起来，形成的是一个高度预防性的法律网络。这种预防性导向反映了一种“家长式”哲学——欧盟试图通过法律来预测和预防损害，即使这会给行业带来负担（De Sadeleer，2002；Sunstein，2005）。

在防控式监管的同时，欧盟也为符合其价值观的人工智能提供大量的政策支持和激励措施。欧盟委员会的《人工智能协调计划》（*Coordinated Plan on AI*）（2018 年，2021 年更新）（European Commission，2021a），作为欧盟成员国在 AI 研发、部署和治理方面的协调机制，为人工智能研究调集资金和进行协调。还有一些项目重点投资于人工智能创新中心和测试设施，比如“地平线欧洲”（European Commission，2021b）（欧盟 2021—2027 年科研创新框架计划，其中 AI 研究是重点领域之一）和“数字欧洲”（European Commission，2025）（欧盟 2025—2027 年的一项旗舰资助计划，旨在加速欧洲企业、公民和公共行政部门的数字化转型）。欧盟机构已承诺投入数十亿欧元用于“可持续人工智能”，旨在通过强有力的法律保障推动可信赖人工智能系统的发展。这种“胡萝卜加大棒”的组合体现了家长式防控的两个目的——国家优先考虑公众信任和安全，同时通过资金投入和政策支持来引导人工智能发展方

向。欧洲的人工智能开发者面临着高昂的合规成本，但也有机会受益于明确的规则和对负责任人工智能的公共资助。

并且，AI 治理领域也有潜在的“布鲁塞尔效应”（Bradford，2020）：全球公司可能会为了维持欧洲市场的准入而预先采用欧盟标准，从而将欧盟的严格方法输出到国外。事实上，在 2025 年初，《人工智能法案》的首批条款生效，开始禁止欧盟范围内某些“不可接受的”人工智能实践（如公共场所的实时面部识别和无差别的面部图像抓取）（European Union，2024）。

（二）美国：通过去中心化的软法和行业措施实现“守夜人看护”

与欧盟形成鲜明对比的是，截至 2025 年，美国尚未颁布任何专门的或全面的人工智能法律，而是依赖于一个去中心化的、由针对特定行业的原有法律和新近指南形成的组合方案。政府主要通过现有的而非新定的法律来监督和指导人工智能的发展。

美国具有约束力的人工智能法律工具仍然是零散的，没有统一的“人工智能法案”的等价物。相反，各种既有的联邦法规和规章被适用于特定领域的人工智能应用。例如，食品药品监督管理局（FDA）——作为美国联邦政府负责监管食品、药品、医疗器械等产品安全的独立机构——对人工智能驱动的医疗设备做出规定，国家公路交通安全管理局（NHTSA）——作为美国联邦政府负责机动车辆和道路交通安全标准制定与执行的机构，制定关于自动驾驶汽车的安全标准（National Highway Traffic Safety Administration，2017），而像反歧视法规这样长期存在的法律正在通过解释的方式涵盖算法决策。联邦机构在其原本的授权范围内各行其是：联邦贸易委员会（FTC）负责监管商业中的欺骗性人工智能行为，平等就业机会委员会（EEOC）负责处理招聘中的人工智能偏见，等等。

与此同时，正如我们在后文中将会进一步介绍的，美国州政府在某些人工智能应用的立法方面一直很活跃（如面部识别或算法招聘工具），形成了州一级人工智能法律的拼凑局面。这种多层次的法律格局反映了美国对多元化的、自下而上治理的偏好，但也造成了碎片化和潜在的保护漏洞。

为了填补这些漏洞并协调政策，美国联邦政府一直依赖不具约束力的框架和行政指导。拜登政府时期，美国国家标准与技术研究院（NIST）于 2023 年发布了一份自愿性的《人工智能风险管理框架》（National Institute of Standards and Technology，2023），为行业提供了最佳实践。白宫发布了《人工智能权利法案蓝图》（White House Office of Science and Technology Policy，2022）和一项内容广泛的关于“安全、可靠和可信赖人工智能”的行政命令（2023 年 10 月），阐述了高层原则（如安全、公平、隐私）并指示联邦机构制定人工智能标准（Executive Office of the President，2023）——但这些措施都是软法，不同于欧盟式的硬法，直接的法律效力有限。

美国的政策支持往往倾向于创新激励而非严格控制。近年来，美国联邦政府在人工智能研发方面的投资激增。例如，2022 年的《芯片与科学法案》（Congress，2022）授权数十亿美元用于人工智能研究和半导体产能。但是，美国政府的基本姿态，依然是“守护”公共利益，通过拨款、挑战赛和指南来引导人工智能行业走向负责任的行为，同时基本上避免了先发制人的禁令或许可。这种“轻触式监管”（Light-Touch Regulation）（Black，2001）姿态可以迅速适应并利用行业专业知识，但它在很大程度上依赖于行业自律和事后执法。

尽管特朗普政府宣称对美国 AI 监管体系做出重大调整，但实际的变化极为有限。特朗普于 2025 年 1 月 20 日颁布的《初步撤销有害行政令和行动》（*Executive Order 14148* “*Initial Rescission of Harmful Executive Orders and Actions*”）撤销了拜登政府关于 AI 安全的行政令，23 日的《消除美国人工智能领导地位障碍行政令》（*Executive Order 14179* “*Removing Barriers to American Leadership in Artificial Intelligence*”）要求在 180 天内制订 AI 行动计划，强调去监管和创新驱动，但是基本的监管规则并没有发生重大变化。美国国会正在审议的《算法问责法案》（*Algorithmic Accountability Act*）（要求对自动化决策系统进行影响评估）和《禁止 AI 欺诈法案》（*No AI Fraud Act*）（针对深度伪造技术，旨在保护个人肖像权免受 AI 滥用）仍然毫无进展。美国 AI 监管的立法格局仍然是中央“软”（联邦监管部门主要通过软法治理）而地方“硬”（各州制定具有约束力的法律规则进行地方实验）。

正如观察家普遍指出的那样，美国路径产生了一个“碎片化的格局”，合规性复杂，但一些高风险应用依然可能漏网。总体而言，美国模式体现了守夜人的警惕和最小干预：它旨在避免为人工智能的演变制造障碍，并仅在需要时进行干预，而不是预先确定所有规则。

（三）英国：制定专门立法的“伴侣型监管”

英国新兴的 AI 治理战略体现为一种“伴侣型”的监管路径——政府将自己定位为人工智能创新的务实的合作伙伴，拒绝制定统一的人工智能法案，而是在特定领域制定具有全英适用效力的专门立法。这种总体上基于原则而非规则，但在特定 AI 产业领域出台专门立法，特别是以硬法形式确立强制性规则的治理路径，既区别于美国式的地方探索或软法指引，亦不同于欧盟式的统一覆盖所有领域的全面立法。

2021 年英国发布的《国家 AI 战略》（*National AI Strategy*）（UK Government，2021）确立了未来十年的愿景，目标是使英国成为“全球 AI 超级大国”。该战略强调了在人才、数据、算力、资金等关键驱动因素上的投入，并明确提出要建立“世界上最具创新性的监管环境”。作为该战略的组成部分，通过《AI 行业协议》（AI Sector Deal）为 AI 研发提供资金支持。进入 2025 年，英国政府发布了《AI 机遇行动计划》（*AI Opportunities Action Plan*）（UK Government，2025），为最大化 AI 在推动

经济增长和改善民生方面的潜力制定了清晰的路线图。该计划特别强调了对算力（包括数据中心）的投资，以及推动 AI 在公共和私营部门的广泛应用。这些战略和计划的一个核心考量是，英国若要实现其成为“全球 AI 超级大国”并营造“亲创新监管环境”的战略目标，监管方式必须与之适应。

因此，拒绝效仿欧盟制定单一的、全面的 AI 法案，英国宣称的是一种“基于原则”的治理策略。2023 年发布的《人工智能监管白皮书》（*A Pro-Innovation Approach to AI Regulation*）（Black，2012；UK Department for Science，Innovation and Technology，2023a）概述了一个“相称的、亲创新的”治理框架，明确拒绝设立新的人工智能监管机构或制定广泛的硬性法规。僵硬的、全面的立法被视为阻碍快速创新的障碍。英国选择了一种灵活的、非强制性的、基于原则的监管方法，这并非偶然，而是其国家 AI 战略的直接体现。一个致力于在 AI 这样快速发展的技术领域取得领先和持续创新的国家，自然会倾向于一个能够快速适应并且不会施加过重前期负担的监管体系。英国将责任下放给各行业的现有监管机构，要求其将五个跨领域的人工智能原则（安全、透明、公平、问责和可竞争性）整合到其规则和指南中。例如，英国金融行为监管局（FCA）负责监督金融服务领域的人工智能，英国药品和保健品管理局（MHRA）负责监督医疗设备中的人工智能，以此类推。这些监管机构被期望在其领域内发布针对具体情境的指南或行为准则，而不是执行任何单一的人工智能法律，从而构成了一种按领域划分、具体回应技术风险的治理路径。

同时，与美国以软法倡导和州立法为主的分散治理模式不同，英国在人工智能规制上采取中央主导策略的同时，也保留了具有法律约束力的硬法机制。“基于原则”的理念之下，英国选择的是以行业为单位推进专门立法的治理路径。2024 年通过的《自动驾驶汽车法案》（*Automated Vehicles Act*）（UK Parliament，2024）为全英范围内的自动驾驶技术确立了完整的法律框架，涵盖责任分配、保险机制与合规程序，是英国在特定高风险领域主动设立新法的典型实例。此外，英国议会制定的《网络安全法》（UK Parliament，2023）通过后期修正案将未经同意的深度伪造色情内容定为刑事犯罪，并加重了对私密图像滥用以及利用深度伪造进行骚扰或侮辱的处罚，使英国成为最早明确禁止滥用性深度伪造的国家之一。英国内政部（Home Office）在 2021 年更新了《监控摄像头操作规范》（Home Office，2021），指导警察和地方当局合法使用闭路电视和人工智能驱动的视频分析，明确纳入了上诉法院关于实时面部识别的裁决，强调数据保护（同意、必要原则），并要求执法部门为其在公共场所使用面部识别和其他生物识别监控的偏见、准确性和相称性负责。

与此同时，英国一方面将既有的《数据保护法》（*Data Protection Act*）和《平等法》（*Equality Act*）在垂直领域进行延展适用，另一方面也根据具体技术风险与监管需要推动制度性回应，形成一种软法硬法结合、从中央发起、面向全英统一适用的专门领域的立法模式。由政府制定专门化的而非全面的正式法规并赋予其全英适用

的强制效力，同时在制度设计上以软法为主、兼顾技术发展节奏与产业参与，这正是英国“伴侣型监管”的表现形式。

同时，对于跨行业或处于监管职责之间的人工智能应用，英国在制度与政策层面也做出了积极努力。例如，数字监管合作论坛（DRCF）（Digital Regulation Cooperation Forum，2022）与数据伦理与创新中心（CDEI）（Centre for Data Ethics and Innovation，2021）共同为 AI 治理提供协调机制与规范建议，从而构成针对性立法与制度支持的配套格局。在财政与研究层面，图灵研究所（Alan Turing Institute，2018）与 AI 安全研究所（Artificial Intelligence Safety Institute，AISI）等科研平台也被纳入国家战略资源配置体系，为特定领域立法提供知识基础与前瞻支持。此外，英国政府一直在推广“监管沙盒”（Regulatory Sandbox）和试点项目，允许人工智能开发者在监督下进行实验——承认“从实践中学习”可以为制定更好的规则提供信息（Allen，2019；Kashiwagi，2017）。“监管沙盒”最初是由英国 FCA 在金融科技领域创立的监管创新工具，允许创新企业在放松监管环境下测试新产品和服务，现已扩展到 AI 领域。由此构成的是一种政府引导、行业执行、多元支持的治理格局，体现出与技术共生、制度协同的“伴侣型监管”特征。

总之，英国模式展示了一个“伴侣型”监管者，如何在不预设统一监管框架的前提下，通过分行业、多领域的专门化立法，结合原有法律延展适用与制度协作，以软硬法相结合的方式，构建起灵活但具约束力的规制体系，在支持创新的同时有效识别并管控风险。

（四）中国：以“伴侣型监管”推进国家战略

中国式伴侣与英国式伴侣有所不同。但是中国的 AI 治理路径同样体现为一种“伴侣型”的监管范式，即国家积极与产业界合作以推动人工智能发展，同时通过将原有法律在垂直领域延展适用，并针对关键领域制定具有法律约束力的硬法，辅之以强有力的专门化规制，确保人工智能发展符合国家战略目标。

中国政府同样发布了一系列具有约束力的法规和行政措施，都是针对特定的人工智能技术和应用，而非制定一部全面的人工智能法律。例如，中国是最早明确监管算法和人工智能驱动服务的国家之一，2021 年 12 月发布的《互联网信息服务算法推荐管理规定》（国家互联网信息办公室等，2021）要求企业向监管机构注册其算法，并强制要求推荐算法的透明度和公平性。国家网信办和其他机构于 2022 年 11 月发布的《互联网信息服务深度合成管理规定》（国家互联网信息办公室等，2022），强制要求明确标记人工智能生成的媒体内容，并规定深度伪造工具的提供者和使用者需要履行获得同意、核实身份并防止滥用的严格义务——与英国一样，中国也是最早明确禁止滥用性深度伪造的国家之一。2023 年 7 月国家网信办联合七部门联合发布《生成式人工智能服务管理暂行办法》（国家互联网信息办公室等，2023），要

求生成式人工智能服务提供者确保内容合法、标记人工智能生成内容、保护个人数据并防止歧视。工业和信息化部、市场监管总局于 2025 年 2 月发布的《关于进一步加强智能网联汽车产品准入、召回及软件在线升级管理的通知》（工业和信息化部和市场监管总局，2025）要求与自动驾驶系统相关的软件在线升级（OTA）需获得监管批准，以防止汽车制造商利用软件补丁掩盖缺陷，并禁止将高级驾驶辅助系统（ADAS）功能宣传为自动驾驶，强制要求向监管机构报告 ADAS 故障或事故。国家网信办和公安部于 2025 年 3 月发布的《人脸识别技术应用安全管理办法》（国家互联网信息办公室和公安部，2025）要求提供商获得明确的知情同意，并提供非面部识别的替代方案，旨在遏制日常生活中不必要的人脸扫描，解决日益增长的隐私和安全担忧。这些规则充当了高风险人工智能的事前许可制度，确保了政府在部署前的有效监督。

此外，类似于英国，中国 AI 治理的法律工具也是一套专门化的、软法和硬法结合的规则体系。一方面，原有监管机构时常通过软法引导行业自律。例如，中国人民银行于 2021 年 3 月发布了金融服务领域算法风险管理标准，旨在促进金融科技人工智能的透明度和控制偏见。国家卫生健康委员会和国家药品监督管理局于 2021 年 7 月发布了人工智能医疗器械指南，规范人工智能驱动诊断软件分类，确保人工智能在医疗保健领域的安全。科技部于 2022 年发布了包含人工智能在内的新科技伦理指南，提出了关于“可信人工智能”发展的高层原则，强调隐私、公平和监督。另一方面，《网络安全法》（2017）、《数据安全法》（2021）和《个人信息保护法》（2021）等基础性法律预先确立了数据处理、安全评估与信息保护的底线规则，其延展适用为进一步推进人工智能在具体场景中的合规治理提供了法律基础，并为分行业、分技术类型的规制实践打开了空间。

在此基础上，中国在实践层面更多是通过专门化、行业性的行政规定和行业指南实现 AI 治理。中国政府在积极引导人工智能创新（通常与科技巨头合作）的同时，也运用一系列有针对性的行政法规将人工智能引导到既定方向。中国的监管合规相对简化——企业与政府密切合作，并迅速适应新的规则，而这些规则通常定义明确且统一适用。与美国零散的法律规定相比，中国企业可能更容易了解红线并在其范围内快速行动。这帮助中国的监管部门实现了类似英国的“敏捷治理”。

在政策支持方面，中国政府始终是人工智能发展的积极推动者，并以中央层级的战略部署和制度安排推动全国范围的技术进步与场景落地。国务院于 2017 年发布的《新一代人工智能发展规划》（国务院，2017）设定了中国到 2030 年引领世界人工智能的战略目标。科技部自 2019 年起在全国设立 AI 创新发展试验区（科技部，2019a）。一些类似政策都反映了“发展优先”的国家战略——利用产业政策、补贴和采购来加速人工智能的应用，同时，通过行业性监管确保这种发展模式不会破坏社会稳定或国家控制。国家与产业界之间的伙伴关系建立在企业合规以及双方对社

会稳定和经济增长的共同利益之上。

总之，中国的 AI 监管模式将贴近式的国家监督与强有力的国家支持相结合，体现了一种与英国相似但更为紧密的“伴侣型”监管路径：政府引导产业发展，牢牢掌握方向盘，以确保人工智能服务于经济发展和社会管理的战略目标。

三、监管结构与实施方式

任何 AI 治理模式最终都要由特定的监管部门加以实施。有的模式选择的是建立全新的、全面的监管部门，有的则仍然依赖于旧有的政府分支。

（一）欧盟：新设全面监管机构与多层次执法架构

欧盟的治理模式包含一个多层次的执法架构，将欧盟范围内的协调与成员国层面的实施结合起来（Marks and Hooghe，2001；Bache and Flinders，2004）。

为确保整个欧盟的一致性，《人工智能法案》设立了一个中央集权的欧洲人工智能委员会（European Union，2024），类似于《通用数据保护条例》的欧洲数据保护委员会（European Data Protection Board，2018；Kuner et al.，2020），以协调各成员国政府并发布指南。此外，欧盟人工智能委员会设立了一个专门的人工智能办公室，以支持该委员会并帮助制定技术标准。这个新的中央监督机构，通常被称为欧洲人工智能办公室（European Commission，2024），负责起草《人工智能法案》的解释性指南和行为准则。截至 2025 年初，该人工智能办公室一直在与数百个利益相关者进行磋商，以最终确定一份全面的行为准则（Code of Conduct），尽管因其人员配备有限以及达成泛欧共识的复杂难度而面临挑战。

同时，根据《人工智能法案》的规定，各成员国需要将特定的国家监管机构（往往是新创设的）作为 AI 治理的主要执法机构，即所谓“指定机构”（Designated Authorities）（European Commission，2021c）负责监督人工智能提供商和用户的合规情况。例如，目前在一个国家执行产品安全或数据保护的机构可能会被指定负责监督其职权范围内的人工智能系统。因此，欧盟模式具有两级执法：一个协调政策和制定标准的中央层面，以及一个进行检查、调查和制裁违规行为的成员国层面。《人工智能法案》下的执法权力是巨大的——监管机构可以对严重违规行为处以高达 3 000 万～4 000 万欧元或全球营业额 6%～7% 的罚款（这一数字与 GDPR 罚款相当或更高）（European Union，2016；European Union，2024）。欧盟致力于强力执行其预防性规则。

与此同时，传统的监管机构也将继续发挥重要作用。例如，像医疗器械管理机构或金融监管机构这样的行业机构，预计会将《人工智能法案》的义务整合到其对人工智能驱动产品的监督中，比如确保人工智能诊断工具在获准上市前已通过合格

评定（Conformity Assessment）。新旧机构之间的互动也将是一个持续的过程：现有机构带来领域专业知识，而新的人工智能委员会则确保跨行业风险（如透明度或偏见）得到一致处理。以往欧盟法规（如数据保护）的经验表明，各国执法力度存在不小的差异（European Commission，2020）。《人工智能法案》只能通过促进监管机构之间的合作和信息共享来缓解这一问题（European Union，2024）。

总之，欧盟的执法模式是一种混合模式：它设立了一个中央协调机构，并授权成员国机构，旨在实现法律统一和地方问责。这种结构反映了欧盟更广泛的治理风格——一个多层治理体系，需要布鲁塞尔与成员国首都之间持续协商，以便在人工智能监督方面“用一个声音说话”。

（二）美国：现有机构的事后监管与碎片化的执法格局

美国在联邦层面尚未设立任何专门的人工智能监管机构；相反，与人工智能相关的执法分散于现有的监管生态系统（Fragmented Regulatory Ecosystem）。美国既不存在类似于《人工智能法案》这样的统一法典和人工智能委员会这样的单一中央监督机构，又缺少适用于特定领域的专门性但系统性的硬性规则，也不存在一个补充性的承担首要责任的监管部门，由此造成了美国碎片化、相对迟缓、临时和被动的执法格局。

在 AI 时代，美国的传统监管机构继续发挥支配作用。各种联邦机构已将其执法活动扩展到人工智能领域。例如，联邦贸易委员会（FTC）根据其消费者保护授权严厉打击不公平或欺骗性的人工智能应用（Federal Trade Commission，2020）；司法部（DOJ）和平等就业机会委员会（EEOC）负责执行反歧视法，日益涵盖贷款、住房或招聘中的算法偏见（U. S. Department of Justice，2022；U. S. Equal Employment Opportunity Commission，2022）；食品药品监督管理局（FDA）在批准支持人工智能的医疗设备前专门审查安全性与有效性，并可以召回不安全的人工智能驱动产品（U. S. Food and Drug Administration，2021）；证券交易委员会（SEC）监控算法交易是否存在市场操纵（U. S. Securities and Exchange Commission，2023）。与此同时，国家公路交通安全管理局（NHTSA）负责监督自动驾驶汽车的测试，并有权调查人工智能驾驶系统的缺陷（如对特斯拉的 Autopilot 等驾驶辅助系统的调查）（National Highway Traffic Safety Administration，Office of Defects Investigation，2024）。因此，每个机构都运用其传统权力（如规则制定、许可、执法行动）来监督其管辖范围内的人工智能。这种行业执法路径（Jordana and Levi-Faur，2005）因其在相应垂直领域的丰富经验而从以往积累的专业知识中获益。因此，这些机构能够在本领域内实现灵活的、针对具体用例的及时响应：机构可以发布量身定制的指南，比如 FDA 的《人工智能/机器学习医疗软件行动计划草案》（U. S. Food and Drug Administration，2021）或 NHTSA 的自动驾驶汽车测试指南（National Highway Traffic Safety Administration，

2017），而无须新的立法，更多的是对现有法律如何适用于新技术的解释。

然而，问题也显而易见：缺乏统一的框架可能导致空白和重叠。特别是在影响许多行业的跨领域问题上，比如人工智能透明度或数据治理，各机构必须推进却往往难以有效协调。某些行业受到高度监管，而另一些行业几乎没有监管。监管机构对那些不完全属于任何单一机构权限的系统性人工智能风险必然反应较慢。华盛顿目前正在就是否设立新的人工智能监督机构或授权现有机构（如商务部或联邦贸易委员会）拥有更广泛的人工智能监管权力进行辩论，但截至 2025 年春季尚未达成共识。

美国政府也在采取一些措施来改善人工智能方面的跨机构协调。例如，2020 年的《国家人工智能倡议法案》（National AI Initiative Office，2020）设立了国家人工智能倡议办公室，负责监督人工智能研究战略，并设立了一个咨询委员会（NAIAC）（National AI Advisory Committee，2022）来推荐政策。尽管这些机构更侧重于创新政策而非执法，但目前依赖的仍是一个非正式的网络：各机构通过像联邦贸易委员会共同发起的全球隐私执法网络（Global Privacy Enforcement Network，2023）或白宫领导下的类似努力来分享最佳实践。2023 年，拜登政府成立了白宫人工智能工作组（White House AI and Tech Talent Task Force）（White House，2023），随后又成立了人工智能委员会（White House，2023），汇集了各机构负责人，旨在同步 AI 治理工作。然而，这些协调机构缺乏直接的监管权力，与欧盟人工智能委员会完全不同。

值得注意的是，美国政府在 2023 年中期促成了顶级人工智能公司（如谷歌、OpenAI、Meta 等）的自愿承诺（Voluntary Commitments），对人工智能模型进行安全测试并共享信息，这实质上是在没有立法的情况下试图强制执行规范。而在大多数情况下，执法格局可能显得临时和被动：行动往往在危机或公众压力之后发生，比如 FTC 调查导致消费者利益受损的人工智能应用程序（Federal Trade Commission，2024），或 DOJ 等机构在有关人工智能招聘工具存在偏见的报道后进行干预（U. S. Department of Justice et al.，2023）。随着新一届美国政府对拜登政府时期政治遗产的不断清除，前述协调机制作用也相应受限。但是，美国的 AI 治理模式依然没有根本性的改变：在缺乏有效协调机制的情况下，依赖于各部门的授权执法以及通过指南和自愿合规进行的软执法。

总之，美国的执法模式是去中心化的，并且在很大程度上维持现状：寄希望于由众多机构和法律组成的先行体系，如果得到适当协调，能够有效监督人工智能。因而，除非本领域内权限清楚，这些执法机构往往反应较为迟缓、临时和被动。这在客观上既给了 AI 企业更为宽阔的发展空间，又导致了由州政府和市政府主导的区域性监管的广泛兴起（后文将会进一步论述）。

（三）英国：基于主导机构和敏捷治理的事中监管

英国的执法路径与其法律策略相呼应：放弃设立单一的中央人工智能监管机构，

而是将人工智能监督纳入现有监管机构的职责范围。这与美国颇为类似。不同之处在于，英国的现行监管机构通过主导机构和敏捷治理致力于“事中”监管，而非仅仅是事后的被动回应。

2023 年的《人工智能监管白皮书》要求各个监管机构根据自身领域的特点和风险，制定具体的指导方针和工具。例如，竞争与市场管理局（CMA）（Competition and Markets Authority，2023）就 AI 基础模型发布了初步报告，而信息专员办公室（ICO）（Information Commissioner’s Office，2020）与图灵研究所合作发布了关于解释 AI 所做决策的指南。这种做法是英国在 AI 治理上的一项深思熟虑的选择，旨在避免“一刀切”的立法模式，相信特定领域的监管方法能够实现更具针对性和专业性的有效治理。AI 技术在不同行业的风险和应用差异巨大（如医疗 AI 与广告 AI），特定领域的专门规制被认为更能适应这种复杂性。

这种做法充分利用了英国现有监管机构在各自领域内积累的深厚专业知识。与其从零开始建立一个全新的 AI“超级监管机构”（这需要大量时间来培养专业能力），英国模式选择的是激活现有监管机构的潜力。行业监管机构因此受到鼓励（在某些情况下还获得了资助），以发展其在人工智能方面的治理能力。比如，竞争与市场管理局（CMA）负责关注人工智能对竞争和消费者福利的影响，而金融行为监管局（FCA）则负责审查金融科技中人工智能的使用是否存在公平性或稳定性风险。再如，如果医疗人工智能设备不安全，药品和保健品管理局（MHRA）（Medicines and Healthcare Products Regulatory Agency，2024b）可以根据医疗器械法撤销其批准；如果招聘算法存在系统性歧视，平等与人权委员会（Equality and Human Rights Commission，2022）可以根据平等法进行调查等。英国的人工智能监管体系赋予各领域监管机构相对自主的裁量空间，使其能够根据行业特点探索差异化的监管方式。这为政策实施提供了广泛的功能性实验。例如，金融科技初创企业可以在监督下测试人工智能驱动的产品（Financial Conduct Authority，2025），而医疗保健监管机构可能会采取更为谨慎的方法，并要求对人工智能临床工具进行逐案批准（Medicines and Healthcare Products Regulatory Agency，2024a）。各领域监管机构在既有职权范围内利用其现有权力“相称地”解决人工智能问题。

这原本意味着类似于美国的碎片化治理。但实际上，英国中央政府在 AI 治理方面建立了更为有效的协调机制。虽然英国没有法定的国家人工智能执法机构，白皮书也没有将新的执法权力授予任何一个机构，但在实践中信息专员办公室（ICO）作为人工智能相关数据保护问题的专责监管主体，逐步成为 AI 治理中的主导机构。作为英国 AI 治理中事实上的首要监管者，2024 年 4 月发布的《监管 AI：ICO 战略路线》（Information Commissioner’s Office，2024c）进一步确立了其以风险为本、比例适当的监管取向。此外，ICO 还发布了《AI 与数据保护指南》（Information Commissioner’s Office，2023b）（2023 年 3 月更新，详述公平、可解释、安全及自动化决策的要求）、

《AI 与数据保护风险工具包》（Information Commissioner's Office，2023a）《数据保护影响评估（DPIA）指南与模板》（Information Commissioner's Office，2024a）（将指南落地为工程师与 DPO 的风险评估工具），以及关于“生成式 AI 与数据保护”的系列咨询（Information Commissioner's Office，2024b）（聚焦合法性、透明度与 DPIA 要求），并设立“监管沙盒”“创新咨询”和“创新中心”（为企业提供先行试验环境，在合规前提下帮助加速 AI 上线，也为制定政策积累实证经验）。相比于 ICO，拟议中的新部门更多是咨询性的，比如英国于 2023 年底宣布设立一个人工智能安全研究所（UK Department for Science，Innovation and Technology，2023b），用以评估先进人工智能模型的风险，但主要是通过研究来提供政策支持，而不是充当监管机构。

不同于欧盟式（全面监管机构）和美国式（协调机构）的人工智能委员会，ICO 是一个独立的专门化的监管部门，既具有强有力的执法权，又并非无所不管，而是主要承担补充性责任。一如前述，那些职权明确的 AI 治理领域，仍然由本领域传统上的监管部门负责。ICO 首先是和其他监管部门一样的专门化的监管者，负责数据相关的 AI 治理。作为英国独立的数据保护主管机关，ICO 依据《英国 GDPR》（Regulation（EU）2016/679，2016）和《2018 年数据保护法》（UK Parliament，2018）可以发布信息/评估通知、执法通知和罚款通知，对企业处以最高 1 750 万英镑或全球营业额 4% 的罚款，并且有权命令暂停或删除相关数据处理。2023 年 10 月对 Snap chat 的“My AI”聊天机器人的预备执法通知，就是 ICO 在生成式 AI 风险评估不足时行使权力的实例（Information Commissioner's Office，2023c）。

另一方面，在超出传统职权划分的空白或交叉地带，ICO 则在英国 AI 治理中起到了主导作用。ICO 的角色是填补空白和协调重叠。这个成立于 1984 年的独立监管机构，拥有调查、整改与重罚的强大手段，使其成为英国大多数涉及个人数据处理的 AI 系统的核心监管者。为防止各自为政并确保采取一致路径，英国建立了一个 ICO 主导的协调机制。其中，最主要的是数字监管合作论坛（DRCF），这是一个由 ICO、FCA、英国通信管理局（Ofcom）（Ofcom，2010）和 CMA 组成的伞式组织（Umbrella Organization）（Mény，1993）。数字监管合作论坛（DRCF）成立于 2020 年，已将人工智能作为重点关注领域，促进就跨越监管领域的问题（如人工智能、网络安全和隐私的交叉点）开展联合研究和指导。在这一模式下，英国新旧机构之间的相互作用更多的是基于 ICO 联合主导的协调机制而非等级制度——像数字监管合作论坛（DRCF）这样的论坛充当了连接组织。数字监管合作论坛（DRCF）还推出了一个 AI 与数字中心试点项目（DRCF AI and Digital Hub），为创新者提供跨监管机构的定制化咨询服务。此外，政府还承诺投入 1 000 万英镑用于提升监管机构的 AI 专业能力。此外，通过全球隐私大会等平台，ICO 还与其他隐私监管机构共享 AI 监管经验，保持政策一致性。ICO 既通过详尽指南、工具包和创新沙盒等手段，引导产业合规创新；又借助数字监管合作论坛（DRCF）等机制，确保英国 AI 治理在

促进创新与保护权利之间取得平衡，是英国“伴侣型监管”的代表者和主导者。

英国执法结构的优势在于敏捷性和官僚机构最少化——已经了解其行业的监管机构可以相对较快地发布量身定制的人工智能指南（正如 ICO 在 2020 年发布的《人工智能审计框架》征求意见稿中所做的那样）（Information Commissioner's Office，2020），并通过沙盒项目与公司互动。英国监管机构发布的行政指南比正式立法更容易更新。同时，分阶段发布监管指南的计划以及对非立法性方法保持“持续审查”的承诺，都体现了其对敏捷性的追求。这种设计使得监管框架能够与技术发展和风险认知同步演进。它还允许在不同领域进行试验：监管机构可以在基于原则的共同精神下尝试不同方法。例如，金融行为监管局（FCA）在金融领域“采用人工智能技术”的指南可能不同于 ICO 在数据保护方面的方法。这种“准实验性”的制度环境，使英国能够在保持规则一致性的基础上实现监管多样性与快速应变。这也意味着，传统监管机构自身也在迅速变革，既需要继续发挥其传统的职能作用和部门间的紧密协调，也需要保持敏捷治理，持续更新其指南、专业知识和规则，以尽可能跟进人工智能的飞速发展。

从本质上讲，英国的执法制度体现了一种以 ICO 为主导的敏捷治理模式：中央政府设定监管预期并提供必要支持，日常的人工智能监督由各领域监管机构在自身职权范围内灵活应对，而超出传统职权划分的监管问题则由 ICO 负责并形成跨部门的协调机制，从而形成了以快速响应和制度适配为特征的“伴侣型”的治理体系。

（四）中国：现行机构密切配合的监管矩阵

与美国和英国一样，中国也没有设立专门的单一的 AI 监管的最高机构。但是，通过特定监管部门主导并联合其他部门共同组成的监管矩阵，中国政府实现了更为有力也更为高效的 AI 治理。

一个关键角色是国家网信办，其作用类似于英国的 ICO，在中国的 AI 治理中发挥着主导和补充作用，特别是在算法规则和生成式人工智能措施等人工智能法规方面。国家网信办还与工信部、公安部以及行业监管机构一起，构成了一个执行人工智能相关规则的机构矩阵（工业和信息化部，2017）。当新法规出台时（如 2022 年的深度伪造规定），中央政府通常会将监督责任分配给一个或多个机构（如国家网信办负责互联网平台，工信部负责工业人工智能等）。这些机构可以进行审计并处以罚款，并且拥有相当大的权力来调查人工智能公司——中国监管机构可以要求访问源代码或对算法进行审计（事实上，一些法规明确允许政府进行现场算法检查）。同时，中国在 2019 年成立了国家新一代人工智能治理专业委员会来制定伦理标准（科技部，2019b），但具体的执法仍然是通过既有的政府机构进行。

传统监管机构也在各自垂类领域继续发挥作用。例如，中国人民银行根据金融

科技风险指南监控金融服务领域的人工智能（如信用评分算法），国家药品监督管理局则类似于美国的食品药品监督管理局（FDA），负责监管人工智能医疗器械。这些现有监管机构在行使原本职权的同时，也会努力与国家法律和中央指令保持一致。例如，《个人信息保护法》生效后，行业监管机构和地方市场管理部门参与遏制非法使用生物特征数据的行动。

对于高风险应用，中国政府也会采取类似于欧盟的防控手段，要求强制性安全评估和许可。例如，在中国部署新的生成式人工智能服务需要通过政府对训练数据和潜在影响的安全审查，而影响公众舆论的算法必须在国家网信办注册。近年来的执法行动包括罚款和关停：2023 年，几家中国科技公司因未遵守生成式人工智能内容规定而受到警告或罚款。同时，当某些人工智能技术受到鼓励时，中央政府可能会容忍甚至鼓励一段时间的宽松执法以促进创新，然后再收紧控制。最近的一个例子是自动驾驶汽车：一系列城市被允许试验机器人出租车和无人驾驶汽车试点，随着技术相对成熟，一些城市（如北京市）正式制定了一项地方性法规（北京市人民代表大会常务委员会，2024），以安全要求来管理自动驾驶汽车的部署，有望成为更广泛的国家规则的前奏。

总之，与英国 ICO 主导下现行监管机构的敏捷治理类似，中国的执法机构在国家网信办主导下，依托强大的监管矩阵直接制定并执行关键规则，从而通过高效及时的事中监管确保 AI 发展的国家战略。正是中国和英国的这种监管模式，使其 AI 治理具有鲜明的“伴侣型”特征：不是通过预先设立全新机构全权处理所有 AI 治理问题，也不是仍由现行机构在原有职权范围内平行执法、各自为政，而是形成了一个特定监管部门主导、其他监管部门协同配合的敏捷高效的治理模式。因此，伴侣型监管既继承了垂直领域积累的原有治理经验和专业知识，又通过主导机构的补全和兜底责任，避免了监管的严重空白和过度重叠。

四、司法审查与制衡机制

不难看到，不论在何种模式下，尽管监管机制不同，行政机构都在 AI 治理中扮演了核心角色。与此同时，尽管通常是被动审查，法院不论是在“监管”这些监管机构上，还是直接作为特殊类型的 AI“监管”机构，都发挥了不可或缺的辅助作用。

（一）欧盟：积极“监管”的法院体系

在欧盟，司法审查在 AI 治理中扮演着至关重要的角色，这植根于该地区发达的法治传统和多层次的法院体系：包括成员国内的各级法院、欧洲法院（European Court of Justice，ECJ），以及处理更广泛人权问题（不限于欧盟）的欧洲人权法院（European Court of Human Rights，ECHR）。基于《人工智能法案》，欧盟的各级各类

法院实际上成为（行政）监管者的“（司法）监管者”，并且由此通过司法解释发展出新的 AI 监管规则，产生了事后救济的事前效力。

通常，欧盟的司法监督是被动的——法院在法律受到质疑后审查法律的适用情况，并可以推翻违反更高规范的决定或法律。然而，也存在一些主动/嵌入式机制。例如，在许多欧盟监管制度下（包括《人工智能法案》），都要求进行影响评估和咨询，这些在高风险系统实施前充当了内置的制衡机制。个人或倡导团体通常可以通过向独立的监督机构提起诉讼或投诉来启动司法审查（这些机构的决定可以向法院上诉）。荷兰 SyRI 案（District Court of The Hague，2020）是 2024 年之前司法干预算法治理的一个突出例子。SyRI 是荷兰政府用于标记福利欺诈的算法系统，一个由非政府组织和公民组成的联盟在法庭上对其提出了质疑。海牙地方法院裁定该系统违反了《欧洲人权公约》（*European Convention on Human Rights*，ECHR）所载的隐私和人权原则，有效地阻止了其使用。这是一项被动审查——在 SyRI 部署之后，但它展示了司法机构针对不透明人工智能系统执行基本权利的意愿。这表明，欧洲法院在《人工智能法案》颁布之前就已经成为受人工智能影响的个人的后盾，甚至审查监管者是否遵循了适当的程序（透明度、公平性评估等）以及给予个人权利应有的尊重。

如今，根据《人工智能法案》，司法审查将不可避免地扩展到执法行动：被监管机构罚款或处罚的公司可以向成员国法院并最终向欧盟法院上诉，以质疑这些决定甚至条款的有效性。进而，还存在宪法审查的前景：如果一个成员国或公民认为《人工智能法案》或实施该法案的国家措施侵犯了基本权利或欧盟条约，他们可能会向欧洲法院（ECJ）提起诉讼。虽然《人工智能法案》的制定旨在与《欧盟基本权利宪章》（European Union，2000）保持一致，但其适用（如“实时生物特征识别”禁令的确切范围）未来可能会面临法院的法律解释。此外，欧洲人权法院（位于斯特拉斯堡，独立于欧盟但对许多欧盟国家具有约束力）可以审理关于国家使用人工智能（如监控系统）是否侵犯人权的案件。

总之，欧洲的司法审查尽管是事后的和原则驱动的，然而强有力的司法救济的存在本身就具有事前效力：欧洲政府和公司知道它们的人工智能部署可能会在法庭上受到质疑，从而激励它们采取预防措施。在一起发生在意大利的案件中，一家法院裁定 Deliveroo 的骑手调度算法具有歧视性，并下令进行修改（Soluzioni Lavoro. it，2020；Business & Human Rights Resource Centre，2021）。当时，《人工智能法案》尚未颁布，法官将欧盟数据保护和劳动法原则应用于算法决策，已经在欧盟的 AI 治理实践中发挥作用。不难想见，随着《人工智能法案》的颁布实施，法院将会通过解释法律的方式为人工智能应用设定更多界限，从而实行一种类似美国式“判例法”的嵌入式监督。2024 年已经有成员国法院向欧盟法院提交了解释欧盟《人工智能法案》的初步裁决请求。

因此，欧盟在国家和联盟层面的司法系统是 AI 治理的关键组成部分，它确保了贯穿欧盟法律的高层价值观（人的尊严、自由、民主等）在人工智能背景下得到执行。随着时间的推移，随着 AI“判例法”的发展，我们可能会看到司法机构阐明针对人工智能的特定原则（类似于法院在互联网时代为数据隐私制定的检验标准）。这种不断发展的法官制定法将会持续补充成文法规。

从本质上讲，在欧洲，任何受人工智能驱动决策影响的人都有寻求补救的法律途径，而且法院在干预人工智能纠纷方面也颇为积极——这使得司法审查成为欧盟模式下可持续的、尊重权利的 AI 治理的有力保障。法院必将成为《人工智能法案》下的一种特殊类型的“监管”部门，不仅监管使用 AI 技术的企业，也监管那些监管 AI 应用的监管机构。

（二）美国：消极司法审查和原有判例法体系的纳入

在美国，迄今为止对人工智能问题的司法审查在很大程度上不仅是被动的，而且是消极的。由于没有一般性的人工智能法律供法院解释，因此涉及人工智能的案件通常援引宪法原则、反歧视法、行政法或侵权责任等法规，通过现有法律框架下的诉讼进行。由此导致的结果是，与人工智能有关的判例分散在不同领域，随时可以纳入联邦和各州的现行判例体系。并且，迄今为止尚没有一例针对 AI 相关诉讼的联邦最高法院判决。与行政监管碎片化一致的是司法审查的碎片化。

美国法院习惯于裁决新技术问题，但人工智能的复杂性在责任归属和证据透明度方面带来了新的挑战。美国法院在人工智能系统使用前不会对其进行审查或许可。相反，司法干预发生在原告提起诉讼、声称人工智能驱动的行为侵犯了其权利或违反了法律之时，因此自然而然被嵌入了原本所在的法律领域。例如，在消费者保护法领域，美国公民自由联盟（ACLU）根据伊利诺伊州的《生物特征信息隐私法》（Illinois General Assembly，2008）起诉了 Clearview AI（一家面部识别公司），最终在 2022 年达成和解，限制了 Clearview AI 的销售并设定了合规要求（Circuit Court of Cook County，Illinois，2022）；在就业领域，求职者已就算法招聘工具提起诉讼，声称存在歧视，比如根据《民权法案》（*Civil Rights Act*）对一家算法招聘公司提起的集体诉讼（U. S. District Court，Northern District of California，2023）。总体而言，美国司法机构保持的依然是其基本立场：将现有法律原则应用于新的事实。这意味着，如果一个人工智能系统导致歧视，法院会像对待任何其他导致歧视的做法一样予以审视。

相比于美国法院在 AI 相关的民事案件中有很多判决，司法监督在行政案件中则鲜有作为。与欧盟类似，如果联邦机构就人工智能发布规则或采取执法行动，企业和个人可以在法庭上质疑这些行为超出权限或过于恣意。例如，虽然目前尚未专门针对人工智能，但是可以设想，一家公司质疑联邦贸易委员会（FTC）对某项人工

智能实践的制裁，这将导致法院审查 FTC 是否正确地将消费者保护法应用于人工智能（Federal Trade Commission，2021）。但是，与欧盟不同的是，美国法院在“监督”监督者方面没有那么积极。美国司法审查的性质是消极的和分散的——它缺乏欧盟法院的那种全面的监督功能，更没有一部《人工智能法案》可以依据。在欧盟，监管决定会受到法院的普遍审查，以确保其与广泛的权利框架保持一致。相反，美国原告必须在现有法律中找到依据。更为重要的是，美国法院在技术问题上通常都会给予行政机构相当大的尊重，这意味着如果监管机构颁布人工智能法令，只要这些法令是合理的，法院就可能予以维持（尽管雪佛龙尊重原则已被 Loper Bright 案改变）（Supreme Court of the United States，2024）。相反，如果一个机构未能监管明显的危害（这也是目前的常见情况），通常无法通过法院强制其行动（除非通过广泛的宪法主张，而这显然相当困难）。

这当然不意味着美国法院在 AI 治理实践中的态度是完全消极的。司法参与的一个特定领域是透明度——美国法院一直在努力解决法律程序中“黑箱”算法的问题。在一些情况下，被告试图获取法医人工智能工具（如 DNA 软件或酒精测试仪算法）的专有源代码以质疑其准确性。当公司拒绝时，法院有时会下令披露或宣布证据无效，理由是被告有权审查证据，这表明正当程序可以通过法院强制实现算法透明度（Superior Court of New Jersey，Appellate Division，2021）。一些法官在注意到局限性的同时接受了人工智能的输出作为证据。在刑事判决中使用算法风险评估已受到宪法方面的质疑。在 State v. Loomis 案（Wisconsin Supreme Court，2016）中，威斯康星州最高法院审查了被告对在判决中使用 COMPAS 风险评分（Angwin et al.，2016）提出的质疑，尽管最终允许使用，但警告说它不能成为决定性因素，并强调了正当程序方面的担忧。在另一个领域，一些因面部识别匹配错误而被错误逮捕的个人已起诉警察部门。例如，在密歇根州的案件中，无辜的黑人原告被警察人工智能系统错误识别，导致了和解和政策改变（U. S. District Court，Eastern District of Michigan，2024）。但是，这些司法“监管”仍然局限于个人权利保护，而不是像欧盟那样扩展为对 AI 监管机构的司法审查。

总之，美国模式下的司法审查目前充当着补救措施的角色，虽然在个案中可能很强大（如为被错误逮捕的人获得赔偿，或通过援引现有法律下令修改有偏见的算法），但并非人工智能领域的系统性治理。随着时间的推移，随着更多人工智能纠纷诉诸法庭，美国“人工智能法”的判例体系将更为支离破碎。因此，美国的司法体制难以发挥欧盟式的家长制作用，而是在守夜人的国家角色中为特定领域的企业行为划定底线——一条模糊的虚线。美国法院和行政机构共同构成了其保守主义 AI 治理模式的监管结构，都是由现行机构在原有职权范围内依据原有法律对新的 AI 问题做出回应。

（三）英国：司法审查作为对政府人工智能应用的关键制衡

在英国，司法审查是质疑公共机构行为合法性的既定机制，这也延伸到政府机构部署人工智能的情况。在“监督”监督者方面，英国法院与欧盟法院更为相似。

英国没有像美国那样的成文宪法，但它拥有强大的行政法原则，并且受到《欧洲人权公约》及其自身《人权法案》（*Bill of Rights*）的约束。因此，当个人和组织认为公共机构使用的人工智能系统侵犯了隐私、平等或行政法标准（如公平行事或考虑相关因素的义务）时，可以提起诉讼。Bridges 案（Court of Appeal（Civil Division），2020）是人工智能背景下被动司法审查的一个里程碑式案例。此案涉及警方在公共场所使用实时面部识别技术，公民埃德·布里奇斯（Ed Bridges）以侵犯隐私（《欧洲人权公约》第 8 条）和平等为由对此提出质疑。上诉法院裁定该部署不合法——法院认为，缺乏明确的法律指导方针以及缺乏检查算法偏见的措施，意味着警方侵犯了隐私权和公共部门平等义务。这一判决不仅迫使南威尔士警方，而且实际上迫使所有英国警察部队，暂停或重新考虑面部识别计划，等待适当的监管。它体现了英国法院如何通过为合法使用人工智能设定条件来主动影响政策。Bridges 案是被动的，因为它是在技术使用后发生的，但其影响是前瞻性的：它确立了在没有充分保障和明确性的情况下此类人工智能的禁止使用规则。

总的来说，英国的司法监督虽然是事后的，但并不消极。法官们表现出对技术的了解和要求透明度的意愿——在一个值得注意的 2019 年的案件中（High Court of Justice Queen’s Bench Division Administrative Court，2020），高等法院斥责内政部使用了一个可能存在偏见的签证算法，导致内政部在法律和公众的联合压力下放弃了该算法。另一个领域是 2020 年 A-level 考试算法丑闻（Adams，2020）的余波。当一个算法评分系统不成比例地降低了来自弱势学校学生的分数时，引发公愤并被废除。虽然政府在全面司法审查之前改变了做法，但至少有一宗诉讼被提起，法律压力（认为其违反了平等法且不合理）促成了政策的逆转。这一事件显示出英国司法审查充当护栏的可能作用——政府知道，如果算法任意影响个人，法院可能会介入。

英国在某些领域也设有专门的法庭，对于监管机构的决定可以向法庭起诉，确保对执法进行独立的司法审查。由于信息专员办公室（ICO）在某种程度上通过评估政府高风险数据项目（如 ICO 对政府算法的审计）提供嵌入式监督，一旦 ICO 决定本身受到质疑，也可以由法院审查。例如，由于 ICO 在 2022 年对 Clearview AI 开出 755 万英镑的罚款，并要求其删除英国公民的面部识别数据，Clearview AI 于 2023 年 10 月在 First-tier Tribunal 上诉并胜诉：法院判定其行为不在 UK GDPR 的管辖范围内，并撤销罚款（First-tier Tribunal（GRC），2023）。

英国司法机构因而充当着关键的制衡力量，尤其是在政府使用人工智能方面，确保法治和权利得到维护。它倾向是被动的——在问题出现时介入，但这些干预往

往具有广泛的政策影响。这种司法后盾在整个 AI 治理环境中非常重要：法院随时准备制止滥用行为，即使没有详细的立法，也能提供一定的保障。这意味着英国政府和公司必须准备好根据现有法律（在没有专门化立法的领域）在法庭上为其人工智能实践辩护，如果它们越界，法官可能会加以约束。这和美国法院礼让监管部门的传统非常不同。

展望未来，由于英国倾向于基于原则而非规则的监管方法，可能需要法院来充实这些原则在实践中的具体要求。缺乏成文的《人工智能法案》意味着普通法的发展——一种非常英国式的路径——可能会发挥作用。例如，当算法对某人产生重大影响时获得解释的权利，正如欧盟《通用数据保护条例》（GDPR）可能要求的那样。此外，由于英国仍然是《欧洲人权公约》的缔约国，如果国内补救措施失败，个人可以将申诉提交给欧洲人权法院，这可能导致欧洲人权法院做出影响英国人工智能政策的判决（如果类似 Bridges 的案件提交到斯特拉斯堡，它可能会为人工智能监控设定欧洲范围的人权标准）。这些法院设立的硬规则，不同于监管部门发布的那些建议性和引导性的软法，可以和《自动驾驶汽车法案》一样在 AI 治理中发挥关键作用。

正是在这个意义上，英国法院和英国行政部门一道表现出了一种“伴侣型监管”的特征：政府通过专门化的软法为企业提供指引，而法院则在一些关键领域通过硬性规则划定底线，包括对政府行为本身的监管。这既区别于欧盟式的全面监管的家长式防控，又不同于美国法院过于消极的保守角色。英国“伴侣型监管”中的硬法是立法和法院共同设定的。

（四）中国：行政监管的司法补充

相比之下，由于行政部门已经为相关行业设立了专门化的硬性规则，中国法院通常是在涉及人工智能的民事纠纷中，特别是在隐私或消费者权益方面，发挥一定的作用。

在一个具有里程碑意义的案件中，杭州一家法院引用《个人信息保护法》，裁定一家野生动物园对游客使用面部识别侵权，确认未经同意收集面部数据违反了合法和必要的数据使用原则（浙江省杭州市富阳区人民法院，2020）。最高人民法院不久后在全国范围内发布了关于合法使用面部识别的司法解释（最高人民法院，2021）。

同时，中国法院也一直在处理技术问题。在杭州、北京、广州等城市设立了专门的互联网法院（最高人民法院，2019），以有效处理数字纠纷——这些法院处理了从电子商务算法到虚假内容（深度伪造）问题的人工智能相关案件（北京互联网法院，2021；杭州互联网法院，2023）。法院的处理方式通常是通过惩罚违法者来执行新法律，从而有助于实现政府部门的监管目标。

一方面，与美国法院类似，中国法院也会对行政机构的决定保持礼让，特别是

在 AI 领域。最高人民法院发布的关于人工智能相关问题的指导性案例和解释（如面部识别、电子证据和数据隐私）有效地制定了与行政部门监管政策一致的司法政策。另一方面，中国监管机构和最高人民法院密切协调，以合作的方式推进嵌入式监督——当监管机构制定专门化的人工智能法规时，它们通常会邀请法律专家和法官参与，以确保其在法庭上得到适用，并为司法机构实施该法规做好准备。从治理的完整性来看，法院在解释与人工智能相关的新法规以及在私人行为者之间执行合规性方面仍然扮演着重要角色。

因此，与英国类似，中国的“伴侣型监管”也是一种行政与司法互补的格局。但与英国不同的是，中国是由政府监管部门制定专门化的硬性规则，并作为主要的规则设定者，而法院则是在与政府目标保持一致的情况下发挥补充性的作用。

五、地方实验与权力分配

上文讨论围绕的是中央层面在 AI 治理中的立法、执法和司法实践。实际上，地方政府同样扮演着至关重要的角色。任何一种 AI 治理模式都涉及广泛的地方实验以及中央与地方之间的权力分配。在这个层面上，我们能够看到不同模式之间的差异：有些实行的是高度中央集权化的治理策略，留给地方政府的余地较小；有些主要依赖地方政府提供因地制宜的解决方案；还有些则在中央和地方之间形成了积极互动——中央授权地方试点，再将成熟经验提升为普遍范式。

（一）欧盟：跨国中央集权体系下的成员国实施空间

欧盟模式寻求高度的中央统一，《人工智能法案》是一项欧盟法规，直接适用于所有成员国（最大限度地减少国家差异）。然而，依照辅助性原则（Subsidiarity Principle）这一欧盟治理的核心原则，即决策应在最接近公民的层级进行，只有在下级无法有效处理时才上移到更高层级，欧盟将具体执法和某些措施留给了成员国，这必然会产生一些地方差异（Bermann，1994；Føllesdal，1998）。

总的来说，欧盟路径比通常的联邦制更具中央集权特征：它强调单一市场（Single Market），即欧盟内部商品、服务、人员和资本自由流动的统一市场体系（Pelkmans，2006），旨在防止出现 27 种不同的人工智能法规的拼凑局面。尽管如此，成员国在如何履行某些义务方面仍拥有回旋余地。例如，它们可以指定自己的人工智能监督主管部门，并可以制定自己的国家指南或监管沙盒，前提是不与《人工智能法案》冲突。一些欧盟国家已经在欧盟框架内开展了地方监管创新。例如，西班牙在 2022 年成立了一个新的国家人工智能监督机构——西班牙人工智能局（European Commission，2022），使其成为欧洲第一个专门的人工智能监管机构。该机构可以被视为在欧盟要求之前主动塑造 AI 治理的地方（成员国）实验。其他国家，

如法国，在其数据保护机构的监督下，利用现有欧盟法律（通用数据保护条例）的灵活性，启动了实验性人工智能沙盒（如在医疗保健领域），允许在真实数据下对人工智能进行受控试验（Commission Nationale de l'Informatique et des Libertés，2021）。欧盟在《人工智能法案》中鼓励此类试点项目和监管沙盒，以此作为协调创新与监管的一种方式。

此外，由于成员国仍然控制许多政策领域（如执法、教育），它们可以在国内对某些人工智能应用施加更严格的规则。一个例子是，在欧盟就面部识别进行辩论的同时，一些成员国或城市并没有等待——比利时警方在获得更明确的法律依据之前被要求暂停了人工智能面部识别的使用（Peeters，2020），而法国则根据一项临时法律在 2024 年奥运会期间试验了人工智能辅助视频监控（Government of France，2023），测试了欧盟可能允许的界限。这些地方行动将反馈到欧盟的政策辩论中。

从结构上看，欧盟 AI 治理中的权力分配可以描述为：中央（欧盟）制定规则，地方（成员国）执行规则，并可以在执行方式上进行一些调整。例如，风险等级定义是中央设定的，但国家监管机构可能会为其国内特定行业发布更详细的指南（只要与法案一致）。欧盟还容忍成员国制定补充欧盟规则的特定人工智能战略。例如，德国可能会更多地投资于工业人工智能，并通过其研究机构制定技术标准（Die Bundesregierung，2020；Deutsches Institut für Normung and Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE，2023）；而爱沙尼亚则侧重于政府服务中的人工智能，并拥有自己的伦理框架（Ministry of Economic Affairs and Communications of Estonia，2019）。这些是政策重点而非法律要求的实验形式。

重要的是，如果一个成员国希望在保护方面走得更远（超出欧盟法律），它们会受到一定程度的限制——欧盟法规通常在该领域优先于国家法律（Schütze，2018）。例如，成员国是否可以禁止欧盟法案仅将其标记为高风险的人工智能实践，比如基于伦理理由或国家安全例外（National Security Exception）原则（即成员国基于国家安全考虑偏离欧盟法律的权利）（Cremona and Scott，2019）。实际上，非正式的差异始终存在。例如，匈牙利在 2023 年通过了一项扩大生物特征监控的法律（European Center for Not-for-Profit Law，2023），一些机构认为该法律实际上与欧盟法案即将实施的禁止无差别面部识别的规定相冲突。因此，欧盟的多层次治理在很大程度上优先考虑中央集权以确保统一标准，地方实验主要在于实施技术或自愿倡议，而非核心监管标准。挑战在于协调执法——如前所述，行政能力的差异意味着一些监管机构会比其他监管机构更严格或更迅捷。欧盟试图通过在国家政府之间建立合作网络和施加同侪压力来缓解这种情况（“布鲁塞尔效应”也影响成员国自身的政策，以与欧盟规范保持一致）。

总体而言，与美国或中国相比，欧盟体系是一个最明确的跨国中央集权体系（Supranational Centralized System）（Haas，1958；Moravcsik，1998）：允许一定程度的

地方调整和创新测试，前提是不得违反欧盟的统一规则。从以往欧盟在网络法、数据法的实施经验来看，欧盟统一规则往往起到的作用是限制乃至扼杀地方实验。人工智能领域大概率也会如此，特别是作为其监管基础的风险分类本身具有很大的干预空间。

（二）美国：联邦制下州和地方政府的主导作用

美国在宪法设计上允许广泛的地方治理实验，人工智能政策也不例外。由于没有总体性的联邦人工智能法律，各州甚至市政府填补了这一空白，充当了人工智能监管的“实验室”（States as Laboratories）（Supreme Court of the United States，1932；Friedman，1997）。

这导致了权力的广泛分配：联邦政府设定了某些基线（特别是在民权和安全方面），但人工智能规则制定的大部分工作都在州及以下各级政府进行。伊利诺伊州是早期行动者（Illinois General Assembly，2008），并于 2020 年通过了一项法律（Illinois General Assembly，2020），规范在视频招聘面试中使用的人工智能（要求披露和同意）；加利福尼亚州已将算法透明度纳入其隐私法（CPRA）（California State Legislature，2018；Solove and Schwartz，2021），并正在考虑制定针对人工智能的法案；而其他州（如弗吉尼亚州和犹他州）则颁布了法律来管理政府对面部识别的使用（通常要求搜查令或审计）（Virginia Legislature，2021；Utah State Legislature，2025）。截至 2025 年，美国至少有 15 个州制定了某种法律来处理警察或政府使用面部识别技术的问题，而一些州正在辩论立法，以设定在就业或保险等领域人工智能驱动决策的界限。

与此同时，市政府在某些领域甚至更为积极。值得注意的是，近年来，包括旧金山、奥克兰、波士顿、波特兰等城市都对公共部门的面部识别实施了禁令或严格限制。纽约市实施了全美首个法令（New York City Council，2021；NYC Department of Consumer and Worker Protection，2023），要求对该市雇主使用的自动化招聘工具进行偏见审计。这些地方举措反映了社区标准和担忧（Buolamwini and Gebru，2018；Benjamin，2019），公众对种族偏见和监控的强烈不满推动了一系列城市禁止警察使用面部识别的浪潮。地方政府在运营中采用人工智能时也存在不同的监督程度。例如，一个县的儿童保护服务机构可能会使用算法来标记高风险儿童，而另一个城市则出于伦理考虑拒绝使用此类工具（Allegheny County Analytics，2019；The Associated Press，2022）。这些选择创造了一个关于 AI 治理的自然实验，可以为其他地方提供借鉴。

一如前述，司法裁决也是主要发生在地方层面或各巡回区。到目前为止，还没有案件直接就 AI 治理问题上诉至美国联邦最高法院：尽管未来几年，美国联邦最高法院很可能会被要求解决诸如使用算法是否会侵犯宪法权利等问题，比如监控人工智能可能涉及第四修正案的搜查和扣押问题，而政府算法做出决定可能涉及第十四

修正案的正当程序问题。在此之前，许多问题将仍然在下级法院和州法院得到解决。这也必然导致不同的司法解释和规则适用（如一个州的最高法院可能禁止某种类型的人工智能证据，而另一个州则允许）。

监管权力的分配在特定行业背景下也很明显。例如，交通运输主要由州监管，比如加利福尼亚州机动车辆管理局制定自动驾驶汽车测试规则（California Office of Administrative Law，2014），并有权暂停许可（California Office of Administrative Law，2020），正如其在 2023 年发生安全事故后对一家机器人出租车公司所做的那样）（California Department of Motor Vehicles，2023）；而其他行业（如航空）则由联邦优先管辖（Federal Preemption）（Gardbaum，1994）。这是因为，传统上，无人机和飞机上的人工智能属于联邦航空局的管辖范围，地方发言权较少；而交通法规则是由州法规定。

因此，美国模式的去中心化不仅促成了多个监管机构的共管，而且促成了多种“规则集”的共存。其结果是全美范围内的法规碎片化：那些在全美范围内运营的公司必须应对不同的规则，而公民的权利保护也因地而异。这符合美国的联邦制，但也引发了不平等和低效率的问题：某些州的公民是否会比其他州的公民更好地免受人工智能的危害？公司是否会涌向规则较宽松的司法管辖区，从而造成事实上的监管套利（Regulatory Arbitrage）（Romano，1985）。

与此同时，虽然这种碎片化可能会对合规性造成问题，但也产生了积极的竞争效果：一个州的有效措施可以激励其他州。事实上，已经发生了政策扩散（Policy Diffusion）（Gray，1973；Berry and Berry，2018）。例如，伊利诺伊州在 2020 年通过其招聘人工智能法后，至少有其他三个州提出了类似的法案（New Jersey Legislature，2022—2023；New York State Legislature，2023—2024；Maryland General Assembly，2024）。这表明某些地方性规则可能具有超乎寻常的影响力，产生所谓“加州效应”（California Effect）（如加利福尼亚州的汽车排放标准塑造了全美汽车市场）（Vogel，1997）。

未来的挑战在于，联邦决策者最终是将这些实验整合为一个更统一的路径，还是本着竞争性治理的精神继续让多样性盛行。2025 年 5 月 14 日，美国众议院能源与商业委员会通过了一项预算协调法案，其中包括名为“第 43201 节：人工智能与信息技术现代化倡议”的条款（U. S. House Committee on Energy and Commerce，2025），提议在法案生效后，实施为期 10 年的联邦禁令，禁止各州及其下属政治实体执行任何监管人工智能模型、系统或自动决策系统的法律或法规。提出这一禁令的理由是建立统一的联邦 AI 治理框架，以避免各州之间监管政策的持续碎片化。然而，该提案遭到了广泛的两党反对。包括 40 位州检察长在内的批评者认为，这项禁令将削弱各州保护消费者免受人工智能技术潜在危害的能力。他们担心，在缺乏全面联邦监管的情况下，长达十年的州级监管禁令可能会使消费者面临风险（Godoy，2025）。

此外，该提案在程序上也面临挑战。由于人工智能禁令被纳入预算协调法案，根据参议院的伯德规则，该法案必须符合与预算相关的规定，因而可能对禁令最终纳入立法方案构成了不小的障碍。

总之，在美国的 AI 治理模式中权力分配最为分散。联邦政府提供指导并执行基本法律，但许多前沿治理工作都在州和市一级进行。地方实验不仅被允许，而且是固有的：它可以产生创新的解决方案（或揭示陷阱），并最终反馈到更广泛的政策中，尽管代价是日益严重的碎片化局面。

（三）英国：中央集权战略下的部分权力下放

英国作为一个单一制国家，其人工智能政策比美国更为中央集权，但它仍然允许有针对性的地方实验，并且能够有效管理下放的权限。

英国政府的人工智能监管战略适用于英格兰、苏格兰、威尔士和北爱尔兰，但在某些领域（如医疗保健和教育）它把权力下放给了地区政府，这可能会影响这些领域的 AI 治理。2023 年的《人工智能监管白皮书》（UK Department for Science, Innovation and Technology, 2023a）及其原则涵盖整个英国；然而，实施方式可能有所不同。例如，英格兰的国民医疗服务体系（NHS）可能会推出人工智能诊断工具试点（National Institute for Health and Care Research, 2021），而苏格兰的医疗服务体系可能会根据自己的指南独立试验类似的人工智能。苏格兰还发布了一项侧重于伦理原则的人工智能战略，并拥有自己的咨询小组（The Scottish Government, 2021）。

某些市议会或警察部队已尝试将人工智能用于公共服务，但须遵守国家法律（如《数据保护法》）（UK Parliament, 2018）的监督。一个著名的例子是前文提到的南威尔士警方的实时面部识别试验（Court of Appeal（Civil Division）, 2020）。这是地方警察部队的一项举措，触及了法律界限，因而招致了司法介入。另一个例子是地方政府在福利金评估中使用人工智能。一些议会部署了算法来检测欺诈或分配资源，而另一些议会在公众咨询后选择不这样做，这导致了英国境内实践的迥异局面。

随着此类地方应用引发担忧，国家机构也开始制定应对措施（如 ICO 发布关于公共部门算法的指南）。因此，英国的权力分配仍然相当集中：英国政府制定基调和基本规则，而地方实体（包括权力下放的行政机构）则在该框架内实施或进行实验。英国以行业为主导的路径含蓄地赋予了权力下放实体在其权限范围内的自主权。事实上，由于每个地方政府都可以制定一些领域的政策，它们可以为这些领域制定自己的人工智能伦理指南。英国国内也存在一些紧张关系——一个常被提及的担忧是，威斯敏斯特为吸引科技投资而推行的轻触式监管，可能与苏格兰等地区更为谨慎、优先考虑伦理的路径相冲突。如果这种紧张关系加剧，可能会导致实践上的分歧。例如，一种人工智能应用可能在英格兰的医疗系统获准使用，但在苏格兰则不然，如果后者的政府对其安全性并不信服。到目前为止，协调一直占主导地位，像数字

监管合作论坛（DRCF）这样的全英机构包括其管辖范围遍及整个英国的监管部门，致力于消除区域性和行业性差异。

英国地方实验的一个典型方式是政府明确建议设立监管沙盒，以支持在自动驾驶汽车、无人机和机器人等高增长、具有监管挑战的 AI 领域的创新。新成立的 AI 安全研究所（AISI），虽然首要任务是保障先进 AI 的安全性，但也将“构建 AI 治理产品”并“改进评估科学”，从而可能为沙盒活动和实验性治理提供信息和支持，其研究成果将为政策制定提供参考。数字监管合作论坛（DRCF）的 AI 与数字中心试点项目则为企业提供了一个测试新 AI 产品和服务的平台，并就相关监管问题提供咨询。此外，药品和保健品管理局（MHRA）设立的“AI Airlock”是针对医疗器械领域 AI（AIaMD）的特定行业沙盒。这些沙盒不仅是创新的试验场，更是“基于证据的监管学习”的关键工具。中央政府积极推动并资助这些实验性举措，确保从实验中获得的经验教训能够反馈到国家政策的制定中。这种结构化的实验为政策学习提供了宝贵机会，由此形成了一个反馈循环：地方性或行业性的实验产生数据和洞见，这些数据和洞见又反过来为国家层面敏捷调整监管方法提供依据。

我们可以将英国模式描述为以中央集权为主的权力分配：中央通过原则和协调确保凝聚力，但允许不同地区调整执行方式。与欧盟不同，英国的地方实验具有更大的自主性，不仅仅是在法律执行层面。也与美国不同，英国的地方实验主要以国家监管机构策划或批准的试点项目和沙盒的形式进行，而非独立的地方立法。英国的权力分配是中央主导的，致力于在中央对人工智能原则和目标的政策控制与行业和地区在实施中的自主权之间取得平衡，允许一定程度的实验，又不会导致整体监管格局的碎片化。

（四）中国：中央引导下的地方试点

和英国一样，中国也是单一制国家。中国的政治结构虽然是中央集权的，但有地方试点项目的长期传统，即由特定城市或省份在中央指导下测试新的政策（“试点先行”）（Zhu and Zhao，2021）。

中央与地方机构在执法中的互动是中国治理的一个显著特征。虽然中央监管机构制定规则和采取高级别执法行动，但许多日常执法工作是由国家监管机构的地方分支机构或地方政府部门在中央指导下执行的。省市级政府经常在中央指导下实施试点项目和执法行动。反过来，行之有效的地方执法或司法方案能够被中央政府推广为国家标准——与美国自下而上的自发路径不同，地方实验的最初动力往往来自中央推动。中国的执法可以被描述为“中央决策，分散执行”。中央政府确保统一的目标（如不允许威胁社会稳定的人工智能风险隐患），而地方监管机构和公安部门则执行检查，通常为定期形式（如定期清查利用人工智能技术的欺诈行为或检查应用程序是否符合算法备案规则）。

在 AI 治理方面，这种模式表现为中央政府严格控制政策目标和监管标准，同时指定某些地方进行人工智能发展和监管的实验。例如，北京市和上海市已被授权为自动驾驶汽车试验和智慧城市人工智能部署的先行城市。北京市政府与各部委密切合作，发布了中国首批关于自动驾驶的综合性地方法规《北京市自动驾驶汽车条例》（2025 年生效）（北京市人民代表大会常务委员会，2024），规范了从道路测试到无人驾驶出租车部署的方方面面。这项法规既含有地方的意愿，也是中央目标（发展与安全）在地方的具体实施，反映了中央政府如何允许一个主要城市填补监管细节，并随后可能在全国范围内复制经验。同样，以特区灵活性著称的深圳，在 2022 年通过了地方规定，为人工智能和数据用于技术创新提供便利（深圳市人民代表大会常务委员会，2021；深圳市人民代表大会常务委员会，2022），实质上充当了法律改革的沙盒（如在某些条件下确立更宽松的数据规则以训练人工智能）。中国地方政府也通过提供激励措施和建立产业园来竞争吸引人工智能公司——这些政策支持是地方性的，但与国家战略保持一致，并且通常由中央拨款资助。

监管权力的分配在执行上是分散的，但在决策上并非如此。中央政府设定明确的界限（如所有生成式人工智能必须遵守的内容审查规则），但如何在一省推广（如，制造业中的人工智能），则可能留给该省去探索。如果成功，中央政府可能会在全国范围内将其做法制度化。这在其他领域也是如此（例如，金融科技监管通常始于一个城市）。在人工智能领域，我们可以看到许多地方行政实验。例如，不同城市在管理公共场所面部识别方面采取了不同的方法：重庆市等一些城市尝试要求安装公共面部识别摄像头需要备案（重庆市人民政府，2016），而其他城市则只是遵循中央指示整合监控。这种动态机制表明，虽然中央政府占主导地位，但地方是实施和创新的代理人。地方不能违反中央规则，但它们可以增强或以不同方式执行政策。中央政府在需要时采用全国性的“整顿”“治理”行动，可以指示所有城市同时打击某种人工智能滥用行为，从而在这些时刻减少地方自主权。

因此，与其他模式相比，中国在形式上拥有高度集中的权力分配，但能够有效地利用受控的地方实验作为一种治理工具。一个明显的例子是网约车的人工智能算法：中央政府最初让公司自我监管，但在出现问题后，国家网信办联合其他部门于 2021 年直接出台了相应规则，取代了任何地方安排（交通运输部办公厅等，2022）。即便如此，地方政府仍然积极与中央优先事项保持一致，主动充当 AI 治理方法的试验场，为全国推广提供经验信息。例如，广州市人工智能创新区的成功或杭州市社会信用算法的实验，可能导致这些模式在其他地方被采用（杭州市人民代表大会常务委员会，2022；张露和陈钧圣，2025）。

总之，中国在 AI 治理方面的中央—地方动态合作是一个自上而下的体系，辅之以经过校准的自下而上的实验。不同于欧盟高度受限的地方创新空间，也不同于美国式的独立的地区监管自主权，而是与英国类似，中国的地方是中央统一领导下的

政策实验室。这种方法可以快速产生结果，并首先根据当地情况调整规则，但终归与国家战略保持一致——地方实验是完善国家政策的一种手段，而不是地方权力本身的终点。

六、全球视野下的 AI 治理

上述三种 AI 治理模式以欧盟、美国以及中国和英国为代表，但不限于此。许多国家和地区的 AI 治理实践都可以纳入这一分析框架。

（一）家长式防控

1. 转向欧盟模式：韩国、巴林

近年来，韩国在 AI 治理路径上有一个显著的防控式转向。在此之前，韩国表现为一定程度的伴侣型监管：一方面制定了《人工智能国家战略》（*National Strategy for Artificial Intelligence*）（Government of the Republic of Korea，2019），在侧重促进人工智能行业发展的同时，建立了一个“质量已验证的人工智能产品”的认证体系；另一方面在自动驾驶汽车领域制定了硬法和软法规则，积极引导产业方向。但到 2024 年末，韩国成为继欧盟之后第二个通过全面人工智能法律的国家。其《人工智能框架法》（*Artificial Intelligence Framework Act*）（通常称为《人工智能基本法》）（National Assembly of the Republic of Korea，2024）建立了一个广泛的治理结构，旨在预防人工智能风险并建立公众信任。该法定义了“高影响力人工智能”（影响安全或权利的 11 个敏感领域）和“生成式人工智能”的类别，并计划对这些类别施加定制的义务。该法也建立了一个中心化的监督系统：科学和信息通信技术部（Ministry of Science and ICT）是主要的监管机构，总统级别的国家人工智能委员会（National Artificial Intelligence Committee）负责政策制定，而人工智能安全研究所（AI Safety Institute）将支持监测和风险评估。

值得注意的是，同样类似于欧盟，韩国的人工智能法律尽管是预防性的（保护公民权利和安全是核心目的），但也包含了以受控方式进行的促进人工智能创新（如资助研发、数据基础设施和人工智能示范集群）的条款。这表明其允许地方实验（例如，自动驾驶汽车的专用试验台）但始终要在政府监督和伦理准则下进行。总的来说，韩国的法律工具将严格的事前保障措施（如对高风险人工智能进行分类和监管）与国家强有力的政策指导相结合，反映了家长式防控模式中“胡萝卜加大棒”的双重方法。

另一个欧盟的典型追随者是巴林。巴林议会（Shura Council）（协商会议）于 2024 年 4 月 28 日批准的《人工智能监管法》（The Daily Tribune — News of Bahrain，2024）是该地区首部全面的人工智能法律，旨在规范人工智能的开发和使用。该法

规定了防止人工智能滥用的义务和禁令，并对违法行为施以处罚（包括监禁和罚款）。与此同时，巴林还在 2024 年全球人工智能峰会上宣布其《人工智能伦理指南》（2024 年）（Information & eGovernment Authority，Kingdom of Bahrain，2024），概述其人工智能的伦理原则。

2. 尚在途中：巴西、加拿大和土耳其

还有一些国家处于朝向家长式防控模式迈进的途中。巴西是一个典型的代表。巴西国会于 2021 年开始审议的《人工智能法案》（*Marco Legal da Inteligência Artificial*）（Chamber of Deputies of Brazil，2021），建立了一个类似欧盟模式的基于权利、风险分层的框架。该草案将彻底禁止某些高风险的人工智能使用。例如，禁止使用潜意识技术（Subliminal Techniques）或剥削弱势群体的人工智能系统，并禁止公共机构使用社会评分（Social Scoring）或对公众进行广泛的生物识别监控（Biometric Surveillance）。该草案还强制要求对人工智能系统进行事前风险评估，并定义了“高风险”人工智能的类别（如在关键基础设施、教育、就业、信贷、健康领域），要求给予严格监督。一个新的国家人工智能监管机构将被指定来执行这些规则，并有权处以巨额罚款［最高 5 000 万雷亚尔（约为 1 000 万美元）或营业额的 2%，与欧盟《通用数据保护条例》（GDPR）的处罚标准相当］。在司法方面，该草案中包含了透明度、可质疑性和人工监督等原则，确保个人可以对人工智能驱动的决策提出质疑并寻求救济。这种强有力的监管立场（侧重于预防危害和保障权利）是家长式防控模式的典范。

加拿大是另一个例证。其早期治理模式类似于美国，中央层面发布指南类的软法，而安大略、阿尔伯塔等省级地方政府则制定或提议硬法规则。但是，2022 年以来，其 AI 治理在路径选择上日益接近欧盟，体现出较为明显的家长式防控特征。其中央政府提出了《人工智能和数据法案》（*Artificial Intelligence and Data Act*，AIDA）（House of Commons of Canada，2022），这是一项将为人工智能系统的设计、开发和使用（特别是那些被认为是“高影响力”的系统）建立通用要求的法律。与欧盟一样，《人工智能和数据法案》也采取了基于风险的监督立场：它不会一概禁止整个人工智能类别，而是将要求人工智能开发者和部署者进行影响评估，确保透明度，并为高影响力系统实施风险缓解。《人工智能和数据法案》下的执法结构将监督权交给联邦部长（并可能是一个新的人工智能/数据专员），有权审计系统并对不遵守规定者处以罚款。尽管截至 2025 年《人工智能和数据法案》仍在议会审议阶段“死亡”，但加拿大的努力仍在继续。此外，加拿大政府还发布了一份针对行业的《人工智能自愿行为准则》（*Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*）（Innovation，Science and Economic Development Canada，2023），与全球最佳实践保持一致，以弥合正式法规出台前的差距。这种多管齐下的战略（拟议立法、新设监管机构和临时自愿标准）展示了加

拿大的家长式防控理念：对 AI 危害的事前预防和风险防范。

土耳其正在成为防控模式的又一例证。原本其采取的也是美国式的看护策略，依赖多项由政府部门发布的人工智能的软法指南（如关于人工智能在金融和公共服务中的使用）和地方性人工智能实验［如土耳其的各种人工智能初创公司和在伊斯坦布尔、安卡拉等地开展的多个智慧城市试点项目（Yigitcanlar et al.，2018）］。但是，土耳其政府正在通过立法程序推进人工智能监管法案（Grand National Assembly of Türkiye，2024），可能会正式限制此前允许的人工智能实践，以防止滥用。这项立法草案旨在为土耳其的人工智能制定具有约束力的规则，可能会借鉴欧盟式的风险类别和基本权利原则（因为土耳其的科技政策通常与欧洲标准保持一致）。尽管细节仍在不断完善，但人工智能法案的存在表明了其对人工智能施加事前控制的预防性立场。土耳其路径预计还将包括建立一个新的治理机构或扩大现有监管机构的授权以执行人工智能要求，这与防控式的执法结构也颇为相符。

3. 全球 AI 治理的“布鲁塞尔效应”

伴随着欧盟在 AI 治理领域的激进实践，其家长式防控理念也产生了全球性影响。它不仅在一些发展中国家形成了示范效应，更为重要的是，深刻影响了国际共识的塑造和国际软法治理的发展。

除了前述国家，非洲的尼日利亚（Federal Ministry of Communications，Innovation & Digital Economy，2023）和肯尼亚（Republic of Kenya，2025），以及拉丁美洲的阿根廷（Chamber of Deputies of Argentina，2024）和智利（Ministry of Science，Technology，Knowledge and Innovation of Chile，2021）等国也已开始起草风险预防性的人工智能框架。尽管这些国家尚未颁布全面的人工智能法律，但是其近年来的官方战略和白皮书经常强调“负责任的人工智能”原则（Floridi et al.，2018），并呼吁在必要时探索立法以维护公共利益。这一趋势表明家长式防控模式在全球范围内的影响力，超越了通常的西方背景。

更大的影响来自于国际共识的塑造。例如，联合国呼吁就“安全、可靠和值得信赖的人工智能”达成全球共识（United Nations General Assembly，2024），鼓励各国实施监管和治理措施以预防危害。这些努力虽然尚未完全实现，但反映了更广泛范围内的家长式防控精神：它们优先考虑保护性规则和强有力的监督机制，以预防人工智能带来的威胁。这也反映在近年来的各种国际论坛和联盟中。一个主要例子是经济合作与发展组织（Organization for Economic Co-operation and Development，OECD）的人工智能原则（AI Principles）（全球 40 多个国家认可），敦促成员国确保人工智能透明、安全、负责任并受到以人为本的监督。经济合作与发展组织的原则虽然不具有法律约束力，但已促使各国制定人工智能政策，授权监管机构监督人工智能［经济合作与发展组织还启动了一个人工智能政策观察站（OECD，2020），以帮助跟踪和指导这些努力］。

另一个例子是全球人工智能伙伴关系（Global Partnership on AI）（Global Partnership on AI Founding Members，2020），由加拿大、法国、印度、日本等国于2020 年成立，作为一个多利益攸关方的家长委员会：它汇集了政府、专家和行业代表，研究人工智能影响并制定监管和监督的最佳实践。在亚太地区，七国集团的“广岛人工智能进程”（G7 Hiroshima AI Process）（G7 Leaders，2023）（在日本 2023 年担任七国集团轮值主席国期间启动）体现了多边层面的预防性路径——七国集团就“值得信赖的人工智能和人工智能治理”的指导原则达成一致，并承诺协调监管方法以管理人工智能风险（如在各国之间统一人工智能系统的风险评估框架）。

同时，联合国教科文组织的《人工智能伦理建议书》（*Recommendation on the Ethics of Artificial Intelligence*）（UNESCO，2021）获得 193 个国家的通过，鼓励各国建立监督机构（如人工智能伦理委员会）并对人工智能系统进行影响评估。许多西方以外的国家正在采取行动：联合国教科文组织报告称，截至 2023 年，已有 50 多个国家（包括发展中国家）正在根据该建议制定国家 AI 治理框架。这些国际措施不施加直接限制，但致力于促使各国加强国内监督机制，比如算法审计要求、透明度标准和公共部门监督途径（Raji et al.，2020）。除了并非硬法，所有这些都是家长式防控模式的标志。

东南亚国家联盟于 2024 年发布了《东盟人工智能治理和伦理指南》（*ASEAN Guide on AI Governance and Ethics*）（Association of Southeast Asian Nations，2024），这是一份全面但自愿的区域组织指南，鼓励东盟企业和政府在公平、透明、安全和人机协同决策等原则上保持一致，甚至建议在组织内部设立人工智能伦理委员会或董事会。非洲联盟于 2022 年发布《非洲大陆人工智能战略》（*The Continental Artificial Intelligence Strategy*）（African Union，2022），同样提出了“非洲大陆人工智能治理统一方法”的愿景。此外，电气电子工程师学会（IEEE）和国际标准化组织（ISO）都制定了侧重风险管理、透明度的 AI 治理标准（IEEE，2021；ISO/IEC，2023）。这些国际性软法也都体现出明显的预防性特征（Shaffer and Pollack，2010）。

但是，随着“普鲁塞尔效应”的共识塑造日益成功，国际软法有望进一步发展为具有约束力的国际公约。例如，欧洲委员会（Council of Europe）（由包括非欧盟国家在内的 46 个国家组成）正在起草一项《人工智能、人权、民主和法治公约》（*Convention on AI*，*Human Rights*，*Democracy and the Rule of Law*）（Council of Europe，2024）。这项公约将要求缔约国通过立法，确保人工智能系统不侵犯基本权利。这实际上是一种预防性的国际法律工具，是欧盟模式的国际法延伸。这并不奇怪，国际法的底色一直是欧洲意识形态。

（二）守夜人看护

像美国这样采取守夜人式的放任策略的国家不多，毕竟不是任何一个国家都具

有美国监管机构这样的专业能力，而人工智能技术和应用的泛滥首先会危害到本国的生产生活。但是，像印度、沙特、阿联酋和以色列这样在人工智能领域抱有雄心的国家，则可能甘冒风险，争取本国企业在全球 AI 格局中占有一席之地。

1. 印度

印度迄今为止主要采取了一种“守夜人看护”立场，倾向于依靠软法的最低限度的监管策略。印度政府在 2023 年明确表示，近期“不打算监管人工智能的增长”（The Economic Times，2023），称人工智能是一种不应被扼杀的战略技术。相反，负责电子治理、网络安全和数字政策的中央政府部门印度电子信息技术部（Ministry of Electronics and Information Technology）一直专注于通过政策框架解决伦理风险，并依赖于现有法律。印度正在依靠其 2018 年发布的《国家 AI 战略》（*National Strategy for Artificial Intelligence*）［“人工智能造福所有人”（AIfor All）］和随后的政策文件来指导生态系统（NITI Aayog，2018）。这些文件倡导“印度作为人工智能车库”（India as AI Garage）（NITI Aayog，2018）的理念，以期将印度建设成为一个全球人工智能发展中心。印度国家转型委员会（NITI Aayog）通过发布负责任的人工智能指南，确定人工智能解决方案的优先领域（如农业和医疗保健），并启动监管障碍最小的概念验证项目（Proof of Concept Projects）（NITI Aayog，2021）（如人工智能驱动的作物咨询系统或用于公共医院结核病诊断的人工智能工具）。尽管印度也已发出负责任人工智能的期望信号，例如政府支持由行业团体开发的偏见和歧视测试工具包，并鼓励科技公司采用国际标准［如 ISO/IEC 人工智能管理标准（ISO/IEC，2023）］，但是政府的主要角色是 AI 创新的推动者——提供资金、数据（通过开放数据倡议）和问题陈述——同时就伦理准则提供建议。

缺乏人工智能领域的全面性或专门化的硬性法律意味着，根据《2023 年数字个人数据保护法》（*Digital Personal Data Protection Act*）成立的印度数据保护局（Data Protection Board of India）（Parliament of India，2023）将在其一般授权下充当守夜人，处理与人工智能相关的投诉（如数据滥用或算法歧视）。在金融领域，印度储备银行（Reserve Bank of India）（Reserve Bank of India，2017）和证券监管机构已发布算法交易和金融科技人工智能指南（Securities and Exchange Board of India，2024）；在医疗保健领域，印度医学研究理事会（Indian Council of Medical Research）发布了详细的《医疗保健人工智能伦理指南》（*Ethical Guidelines for Application of Artificial Intelligence in Healthcare and Research*）（Indian Council of Medical Research，2023），以确保医疗人工智能应用中的患者安全、算法公平性和透明度。这种鼓励创新的政策也延伸到邦一级——一些印度邦已成立人工智能工作组或创新园区，与初创公司密切合作。例如，泰米尔纳德邦 2020 年通过的《安全和伦理人工智能政策》（*Safe and Ethical Artificial Intelligence Policy*）（Government of Tamil Nadu，2020）设定了广泛原则，但主要用于指导政府使用人工智能，而不是强制私人行为。印度对地方实验的态度总

体上是积极的：政府已在农业、教育和智慧城市领域启动或支持人工智能试点项目，将其作为适当治理的学习场所。本质上，印度通过现有机构视角监控人工智能并制定非约束性指南，遵循“守夜人看护”模式，仅在情况需要时介入正式监管。

2. 沙特、阿联酋和以色列

几个重要的中东国家不约而同地采取了美国式的“守夜人看护”模式：中央进行软法引导，地方通过硬性规则进行多样性实验。

沙特的 AI 治理实践提供了一个典型例证。沙特于 2020 年启动了一项全面的《国家数据和人工智能战略》（*National Strategy for Data and AI*）（Saudi Data & AI Authority，2020），旨在到 2030 年使沙特成为人工智能领域的全球领导者。为了实现这一目标，沙特政府避开了早期监管，而是专注于制定伦理准则和投资人工智能研发能力。该战略促成了沙特数据与人工智能管理局（SDAIA）（Government of Saudi Arabia，2019）的成立，该机构还制定了“沙特人工智能原则”（Saudi Data & AI Authority，2023）——一套高层次原则，如公平、透明和安全，旨在指导政府和行业的人工智能发展。

与此同时，沙特监管机构正在更新相关法律（如加强数据保护和网络安全法律），并为某些人工智能用途（如深度伪造）发布指导文件。其重点是标准和认证，例如，鼓励公司获得 SDAIA 认可的人工智能伦理标准合规认证。同时，沙特大力促进地方实验：它开设了人工智能研究中心，举办了国际人工智能峰会（Global AI Summit）（Saudipedia，2020）以促进公共和私人对话，甚至在其未来城市项目（如 Neom）（Fattah and Martin，2024）中纳入了人工智能驱动的服务作为试验台。政府经常承担人工智能的试点部署（如教育中的人工智能或智能政府服务），以展示潜力，与科技公司成为合作伙伴。这种密切合作，以广泛的伦理承诺为指导，但没有详细的人工智能法规，是“守夜人模式”的典型特征。其在咨询机构和广泛规范的软监督下，给予企业创新的空间，只有当结果可能违反核心价值观或国家安全时才进行干预。

阿联酋的 AI 治理路径同样遵循的是美国模式。阿联酋政府没有制定和施加全面监管的严格法律，而是利用行政指令和战略举措将人工智能融入社会。阿联酋是首批任命人工智能部长（Minister of Artificial Intelligence）（2017 年）（Arabian Business，2017）并发布全面国家人工智能战略的国家之一。阿联酋人工智能部于 2022 年 12 月颁布了阿联酋人工智能伦理原则与指南（UAE AI Office，2022），旨在促进公共和私营部门合乎伦理地使用人工智能。阿联酋内阁发布的 2031 年国家人工智能战略（2018 年发布，至 2023 年持续更新）（UAE Government，2018b）包括了确保对人工智能进行“强有力治理和有效监管”的目标。阿联酋于 2024 年 6 月，通过了一项包含 12 项原则的人工智能宪章（UAE AI Office，2024），以指导负责任的人工智能发展；于 2024 年 9 月发布了一项关于人工智能的国际政策（Ministry of Foreign Affairs

and Minister of State for Artificial Intelligence and Digital Economy and Remote Work Applications Office，2024），阐明其在防止人工智能滥用方面的全球立场，并作为高层政策框架。类似于美国，阿联酋也是在现行法律体系内进行 AI 监管。例如，利用其 2018 年第 25/2018 号联邦法令（UAE Government，2018a），允许对现有法律未涵盖的人工智能项目进行试点许可，以便在临时基础上启用新兴的人工智能应用。

阿联酋的 AI 地方实验是在其两个主要金融自由区进行：阿布扎比全球市场（Abu Dhabi Global Market）（Abu Dhabi Global Market，2024）和迪拜国际金融中心（Dubai International Financial Centre）（Dubai International Financial Centre，2024））。两者都更新了其数据保护法规，以涵盖人工智能处理，要求自动化决策的透明度和人工审查，但同样是在现有数据法律的背景下，而不是制定新的人工智能法案。迪拜地方政府甚至发布了一份《人工智能伦理工具包》（AI Ethics Toolkit）（Smart Dubai，2019），指导城市机构和公司负责任地使用人工智能，该工具包已在智能交通系统等城市项目中进行试点。迪拜政府的 2023 年第 9 号迪拜法律为迪拜酋长国的自动驾驶汽车建立了监管框架（Government of Dubai，2023），规范了自动驾驶汽车的运营，要求运营商从道路与运输管理局获得许可，并遵守安全和网络安全标准。随后，阿布扎比政府的 2024 年第 3 号法律成立了一个人工智能与先进技术委员会（Abu Dhabi Media Office，2024），负责监管和监督阿布扎比的人工智能项目和投资，并为在阿布扎比酋长国部署的人工智能技术制定具有约束力的规则和标准。同时，监管沙盒也得到了积极推广。例如，迪拜道路管理局（Roads and Transport Authority Dubai）运行了一个自动驾驶汽车测试沙盒；中央银行等监管机构正在运行金融科技沙盒，将人工智能驱动的金融产品置于监管指导下进行测试，但没有惩罚性规则。

简而言之，这一监管策略旨在迅速将阿联酋定位为人工智能友好型中心，相信有指导的自律和政府以身作则的领导足以应对风险，而无须广泛的法律控制。

以色列的人工智能以其在国防和科技领域的创新而闻名，奉行的也是守夜人式的宽松监管政策。以色列政府已发布政策计划［如 2024 年《以色列国家人工智能计划》（*ISRAEL AI*：*National AI Program*）（Government of Israel，2024）］并成立了咨询委员会，但截至 2025 年，没有人工智能领域的特定法律生效。相反，以色列通过部门指南和部际协调促进“负责任的人工智能创新”。例如，以色列隐私保护局（Israel Privacy Protection Authority）发布了关于人工智能和大数据使用的指南（Israel Privacy Protection Authority，2025），建议各类组织在现有《隐私保护法》（*Privacy Protection Law*）（Knesset，1981）框架内避免偏见和尊重隐私。同样，以色列创新局（Israel Innovation Authority）制订了鼓励医疗保健领域人工智能初创公司的计划（Israel Innovation Authority，2022），与监管机构密切合作，确保遵守现行法规，而无须引入新的人工智能法律。以色列的法院和法律尚未出现主要的人工智能判例，监管理念是利用和促进以色列强大的科技产业，而避免过度干预。

（三）伴侣型监管

随着 AI 技术和应用的负面影响日益显著，如果不具备美国监管机构类似的专业能力，“守夜人模式”将会付出越来越大的代价。而欧盟“防患于未然”的传统策略，又很容易扼杀本国的 AI 发展。相比之下，一些希望在 AI 产业有所作为的国家开始选择或秉承与中国和英国类似的伴侣型监管模式，即在鼓励 AI 创新的同时给予贴近和敏捷的有效监管。

1. 新加坡和日本：通过软法工具实现“伴侣型监管”

秉承其 AI 国家战略，新加坡选择的是更类似于英国的监管策略，即通过敏捷治理在促进人工智能发展的同时进行有效监管。但是，新加坡的 AI 治理路径更加倾向于软法工具。

与中国和英国一样，新加坡也明确提出了自己的 AI 国家战略。新加坡于 2019 年推出了其《国家人工智能战略》（*National AI Strategy*，NAIS）（Smart Nation and Digital Government Office，2019），并于 2023 年 12 月更新至 NAIS 2.0 版本（Smart Nation Singapore，2023），旨在将新加坡打造成为 AI 领域的全球领导者，重点关注人才培养、产业发展和前沿研究。同时，其 Veritas 项目旨在推动金融领域负责任的 AI 应用（Monetary Authority of Singapore，2019）。同时，AI 新加坡（AISG）作为一个全国性计划（Halimweb，2017），致力于构建新加坡在 AI 领域的深厚国家能力。此外，新加坡政府还通过多项举措支持 AI 研发、推广“AI 赋能产业”（AI for Industry）项目，并大力发展 AI 相关技能培训（AI Singapore，2018）。

2019 年，新加坡个人数据保护委员会（Personal Data Protection Commission Singapore）（Personal Data Protection Commission Singapore，2024a）发布了《人工智能治理模型框架》（*Model AI Governance Framework*）（Personal Data Protection Commission Singapore，2019），这是世界上首批全面的 AI 治理指南之一。该框架（2020 年更新）提供了关于负责任人工智能的详细指导，涵盖了透明度、公平性和可解释性等原则，并建议了部署人工智能的组织内部治理实践。随之而来的是《组织实施和自我评估指南》（*Implementation and Self-Assessment Guide for Organizations*）（Personal Data Protection Commission Singapore，2020a），这是一个自愿性工具包，旨在帮助公司根据这些原则评估其人工智能系统。新加坡没有通过法律强制执行这些规定，而是通过强调可信赖人工智能的竞争优势和认可遵守模型框架的公司来鼓励采用。新加坡政府还建立了人工智能伦理和数据使用委员会（Advisory Council on the Ethical Use of AI and Data）（一个讨论人工智能问题的多利益攸关方机构）等结构，并启动了人工智能验证倡议（*AI Verify Initiative*）（Infocomm Media Development Authority，2022a），这是一个测试沙盒和衡量工具，公司可以使用它来审计其人工智能系统是否符合新加坡的治理原则。针对生成式 AI 的快速发展，新加坡于 2024 年 1 月提出了新的《生成式

AI 示范性治理框架》草案（5 月正式发布）（Infocomm Media Development Authority and AI Verify Foundation，2024b）。《SingPass 面部验证》（*SingPass Face Verification*）由新加坡政府科技局于 2020 年推出（Government Technology Agency of Singapore，2020），2024 年扩大应用，确立了在政府数字身份系统中面部识别技术的使用规范。

新加坡的现行监管机构将现有法规有选择地适用于人工智能领域，但没有一个总体的人工智能法案。例如，银行和金融监管机构（Monetary Authority of Singapore）（Monetary Authority of Singapore，2024）发布了关于人工智能在信贷评分中使用的指南，确保人工问责制。而且，部门监督（如金融、医疗保健、交通）与自愿标准协同工作。同时，新加坡立场明确地支持创新实验。它举办人工智能创新挑战赛，允许自动驾驶汽车试点，甚至指定无人机和机器人测试区域（Civil Aviation Authority of Singapore，2021）——所有这些都在政府的观察下进行，但具有灵活的许可。结果是形成一个共同监管（Co-regulation）的环境——一种政府监管和行业自律相结合的合作关系。这种方法因其敏捷性而受到赞扬——新加坡可以随着技术的发展迅速更新其人工智能指南，避免了立法的缓慢步伐。它体现了伴侣型规制如何依赖信任和伙伴关系，正如一位新加坡官员所指出的，目标是“帮助公司做正确的事情”，而不是惩罚它们，同时确保人工智能技术与公共价值观广泛一致（Ministry of Digital Development and Information，2024）。

新加坡的 AI 治理主要由其现有的、具备相关领域专业知识的监管机构负责，并且形成了主要负责的主导部门。信息通信媒体发展局（IMDA）和个人数据保护委员会（PDPC）是新加坡推动 AI 治理的核心机构，它们负责发布示范性治理框架，发挥着类似于英国信息专员办公室（ICO）和中国国家网信办的作用。个人数据保护委员会于 2022 年发布《生物识别数据安全应用指南》（*Guide on Responsible Use of Biometric Data in Security Applications*）（Personal Data Protection Commission Singapore，2022），规范安全摄像头和生物识别数据的使用，包括面部识别系统的安全应用；于 2024 年发布《AI 推荐和决策系统中个人数据使用咨询指南》（*Advisory Guidelines on the Use of Personal Data in AI Recommendation and Decision Systems*）（Personal Data Protection Commission Singapore，2024b），明确《个人数据保护法》在 AI 系统开发、测试和部署中的应用。信息通信媒体发展局和 AI Verify 基金会于 2024 年 5 月发布《生成式 AI 治理框架草案》（*Draft Model AI Governance Framework for Generative AI*）（Infocomm Media Development Authority and AI Verify Foundation，2024a），涵盖生成式 AI 的九个治理维度，包括问责制、数据质量、透明度和安全性等。金融管理局（MAS）则通过 FEAT 原则和 Veritas 项目来监管金融领域的 AI 应用（Monetary Authority of Singapore，2018）。这些机构之间也存在密切合作。并且新加坡的 AI 治理模式高度重视公私协作。AI Verify 基金会便是一种信息通信媒体发展局与多家科技巨头（如谷歌、微软、IBM

等）及其他行业成员共同参与的公私合作伙伴关系（Personal Data Protection Commission Singapore，2023）。Veritas 联盟同样汇集了众多金融机构和科技公司的参与（Monetary Authority of Singapore，2019）。这都显示出新加坡有意吸纳行业专业知识，并确保治理工具具有实用性和广泛采纳性。这种合作模式有助于加速负责任 AI 实践的开发和推广。

新加坡采取了一种“敏捷”的监管方法，即在必要时针对 AI 的特定领域或用途进行引导。其治理工具主要是软法，表现为非约束性的框架和指南，使得治理措施能够随着技术的演进保持灵活性并及时更新。这种敏捷性也体现在其治理框架的迭代发展上。《示范性 AI 治理框架》从最初版本到第二版（Personal Data Protection Commission Singapore，2020b），再到针对生成式 AI 提出新的框架草案，清晰地展示了其对 AI 技术最新进展的快速响应。同时，新加坡政府也通过设立多种形式的监管沙盒和实验平台，在中央机构的引导下，对 AI 治理方法进行实践检验。例如，金融科技监管沙盒（Monetary Authority of Singapore，2016），由金融管理局设立，允许金融科技公司在受控环境中测试创新产品和服务，其中也包括 AI 应用；隐私增强技术（PET）沙盒（Infocomm Media Development Authority，2022b），由信息通信媒体发展局推出，旨在鼓励企业探索和应用 PET 解决方案，以在利用数据的同时保护隐私；生成式 AI 评估沙盒（Infocomm Media Development Authority，2023），由信息通信媒体发展局和 AI Verify 基金会共同发起，旨在推动生成式 AI 模型评估基准的开发；AI Verify 基金会与工具包（Infocomm Media Development Authority，2022a），本身即是一个由信息通信媒体发展局推动的、联合产业界共同开发 AI 治理测试框架和工具的重大实验项目。这些沙盒和实验项目均由金融管理局、信息通信媒体发展局等中央政府机构设立或大力支持，表明了中央层面的引导和从实验中学习的战略意图，也提供了一个结构化的环境，用于在实践中测试 AI 治理原则。新加坡利用这些平台，不仅旨在促进产品创新、进行专门化测试和完善 AI 治理方法，而且致力于构建国际认可的标准，反哺更广泛的政策制定和议题设定。

总之，新加坡并未将 AI 治理视为一种限制性力量，而是将其看作是促进信任和创新的赋能手段。尽管与在特定领域制定和适用硬法的中国和英国有所不同，但新加坡“敏捷”的框架制定方式以及 AI Verify、Veritas 等实用工具包的开发，同样能够为企业提供清晰指引和支持，从而使治理能够跟上技术快速变化的步伐。因而，新加坡也构成了中国和英国之外“伴侣型监管”的又一个代表国家，并且其自身也能够成为一个亚型。

与新加坡类似，日本也是通过监管机构及时发布指南类的细致软法来进行积极监管，同时又与中国和英国类似，通过发布 AI 产业促进法来明确国家战略，通过修改原有法律来建立专门化的硬法规则。

日本政府于2025 年 5 月通过了《人工智能相关技术促进法》（*Act on the Promotion of Research, Development and Utilization of AI-Related Technologies*）（Diet of Japan，2025），旨在推进 AI 技术研发和应用，并设立 AI 战略本部（AI Strategic Headquarters）。这是少有的人工智能促进法，相比于作为软法的国家战略和规划，更加显示出日本在 AI 产业发展中的雄心。此外，日本个人信息保护委员会将会提出《个人信息保护法 AI 修正案》（*Personal Information Protection Act AI-related Amendments*）（Cooper et al，2025），针对 AI 训练数据使用个人信息设置特别规则，放宽本人同意要求。

日本国会修订后的《道路交通法》（Road Traffic Act）（2023 年 4 月生效）（Diet of Japan，2022），在有限条件下，将 L4 级自动驾驶在公共道路上的运营合法化。此次修订允许全无人驾驶车辆（如无人驾驶自动穿梭巴士）在农村公交线路等指定区域运营，并将逐步扩展到高速公路上的私家车和卡车。这是日本继 2020 年颁布修订的《道路运输车辆法》（*Road Transport Vehicle Act*）、允许 L3 级自动驾驶汽车在有驾驶员待命的情况下行驶之后（Diet of Japan，2019），进一步推动了更高程度的自动化驾驶。同时，日本立法者已开始考虑针对某些人工智能危害进行有针对性的“硬法”干预，特别是随着生成式人工智能的兴起。此外，日本法院一直是通过解释和应用现有法律概念（隐私、产品责任等）在既定法律范围内监督人工智能。

同时，日本的 AI 治理主要以“软法”和现有机构监管为特征。日本没有颁布一项全面的或专门化的人工智能法规，而是发布了伦理准则和部门指南来引导人工智能发展。自 2019 年以来，日本政府一直在其《以人为本的人工智能社会原则》（*Social Principles of Human-Centric AI*）（Cabinet Office of Japan，2019）和人工智能战略更新等文件中推广“以人为本的人工智能”原则，虽不具有法律约束力，但可作为行业和政府机构的指南。在实践中，日本利用其现有监管机构做出敏捷治理。例如，日本政府与行业合作，于 2021 年启动“通往 L4 之路”项目，发布技术标准和安全指南，指导 L4 级车辆的安全推广（Toshio Yokoyama，2021）。个人信息保护委员会（Personal Information Protection Commission）监督人工智能中的隐私问题，确保人工智能训练和使用中遵守数据保护原则（Personal Information Protection Commission of Japan，2024）；消费者事务厅（Consumer Affairs Agency）和竞争监管机构可以利用现有法律处理人工智能驱动的欺诈或串通（Japan Fair Trade Commission，2021）；经济产业省与总务省（Ministry of Economy，Trade and Industry）发布的企业人工智能治理指南（Ministry of Economy，Trade and Industry and Ministry of Internal Affairs and Communications，2022），涵盖了人工智能部署中的风险管理、透明度和人工监督。

新加坡、日本的 AI 治理实践尽管并非典型，但是具有“伴侣型监管”的色彩。相对中国和英国而言，新加坡和日本两国的监管策略旨在通过原有机构不断发展的详尽指南，在控制潜在风险的情况下，确保措施有力，促进人工智能产业发展。就“伴侣型监管”所要达到的贴近式、伴随式的治理目标而言，更新及时、规则细致的

软法同样可以达到和硬法类似的效果。

2. 澳大利亚和新西兰：转向“伴侣型监管”

类似于中国和英国，澳大利亚出台了专门性法律以规制人工智能应用。澳大利亚政府于2024年12月发布《隐私和其他立法修正法案》（*Privacy and Other Legislation Amendment Act 2024*）（Parliament of Australia，2024），创设严重侵犯隐私的法定侵权行为。同时，拟议中的《自动驾驶车辆安全法》（*Automated Vehicle Safety Law*）（Department of Infrastructure，Transport，Regional Development，Communications and the Arts，2024）已经由基础设施、交通、区域发展、通信和艺术部于2024年4月公开咨询，预计2026年生效，旨在建立自动驾驶系统实体责任制，规范自动驾驶车辆的安全操作。

在此之前，澳大利亚实行的是一种类似美国的“守夜人看护”模式：没有专门的人工智能法律，一直依赖一般法律和自愿伦理原则实现AI治理；现有监管机构［如负责在线安全监管的电子安全专员办公室（eSafety Commissioner，2015）］通过发布人工智能指南等软法方式进行治理。然而，福利服务中自动化“Robodebt”丑闻等备受瞩目的算法决策失败案例［澳大利亚2016—2019年间使用自动化债务追讨系统计算福利债务，导致大量错误追债和公众痛苦（Royal Commission into the Robodebt Scheme，2023）］，进一步激发了澳大利亚加强对算法系统进行监督以防止危害的决心。

在就“安全和负责任的人工智能”进行广泛的公众咨询后，澳大利亚政府得出结论，现有系统“不适合应对人工智能带来的独特风险”。作为回应，澳大利亚政府于2024年制定了《AI监管路线图2024—2030》（*AI Regulation Roadmap 2024—2030*）（Department of Industry，Science and Resources，2024），针对高风险AI应用提出强制性测试、透明度和问责制要求。澳大利亚于2023—2024年发布了一份提案文件（Department of Industry，Science and Resources，2023），概述了高风险环境中人工智能的强制性“护栏”。这些规则将对就业、金融和安全等关键领域中使用的人工智能施加具有约束力的要求（例如透明度、公平性评估和可能的外部符合性评估）。此外，澳大利亚政府引入了一项自愿性人工智能安全标准（*AS/NZS ISO/IEC 23053：2022*）（Standards New Zealand，2022），其中包含一套十项“人工智能伦理原则”或“护栏”，鼓励公司立即遵循。这些自愿性“护栏”——涵盖问责制、透明度、隐私、稳健性、公平性、利益相关者参与等——与正在审议的强制性“护栏”保持一致，从而确保安全标准的早期采用者能够在立法生效时做好充分准备。

新西兰也是最先对深度伪造做出强制性法律规制的国家之一。现行深度伪造相关法规包括《有害数字通信法》（*Harmful Digital Communications Act 2015*）第22A条（Parliament of New Zealand，2015）和《公平交易法》（*Fair Trading Act 1986*）（Parliament of New Zealand，1986）第9条和第13条，由新西兰国会、商务委员会制

定，规制非同意的深度伪造内容和误导性深度伪造广告。此外，隐私专员办公室于 2024 年 12 月发布《生物识别处理隐私守则》（*Biometrics Processing Privacy Code*）第二版草案（Office of the Privacy Commissioner of New Zealand，2024），预计 2025 年中期生效，以规范生物识别信息的收集、使用和存储，包括面部识别技术的比例性评估。这些硬法规则的扩展适用与软法指引的结合，提供了足以确保相应领域 AI 治理的有效方案。

七、一个初步的比较分析框架

本文总结的这三种模式——“家长式防控”“守夜人看护”与“伴侣型监管”——构成了研究 AI 治理的一个比较分析框架。

三种治理模式体现了在创新和安全之间不同的平衡策略。“家长式防控”模式对 AI 治理采取预防性、自上而下的路径，政府扮演家长式的管控角色，旨在危害发生前进行预防。该模式下的法律工具往往全面覆盖且具有规避风险的倾向，通常会禁止或严格限制某些被认为具有危险的人工智能应用。执法通过强大的机构集中进行，司法框架强调基本权利和事前保护。地方实验受到限制或严格控制（如在特殊区域或试点项目中），以确保创新不会超越安全措施。欧盟是其中的代表，而韩国、巴西等许多国家紧随其后。

“守夜人看护”模式强调 AI 治理中的事后监督和问责制度，为企业的制度创新提供了宽松的法律环境。该模式下政府扮演守夜人或管家的角色，通过现有监管机构在各自职权范围内来监督人工智能的开发和使用。立法和司法都有意避免高度限制性的禁令。行政监管机构被授权监督人工智能的实施，通常以高级原则或行为准则为指导。在司法方面，该模式依赖问责机制（如审计、影响评估、责任框架）来处理问题。它通常鼓励创新，但坚持通过监控实现“值得信赖的人工智能”，并可能利用上市后监督（Post-Market Surveillance），而不是严格的上市前控制（Pre-Market Control）（European Medicines Agency，2017）。美国以其在 AI 领域公认的领先地位和巨大影响力确立了这一 AI 治理模式，但是鲜有国家能够复制或持续实践这一路径。加拿大、澳大利亚这类英美法系国家纷纷在 AI 治理实践中转向欧盟模式或中国和英国路径，并非偶然。

在伴侣型监管模式下，政府将自己定位为人工智能发展的长期合伙人和引导者。立法者尽可能避免预先施加无所不包的严格规则，而是通过指南、框架、准则、战略和协作平台阐明治理原则和国家导向，同时在特定领域划定规则底线，以确保人工智能负责任的开发或符合国家发展战略。这一路径既会在一些领域（特别是政策鼓励的领域）默认企业创新和行业自律，适用软法治理，保持法律工具轻触或间接特征，也会在特定领域设定行业性的硬规则，划定行为红线。治理主体以原有的行

政性的行业监管机构为核心，既各司其职，也有一两个专门机构承担填补空白和协调重叠的首要监管职责。司法机构与行政部门形成某种互补关系，共同发挥对 AI 应用的监管作用。强有力的中央政府更有自信也更乐于进行地方实验：政府建立沙盒、创新中心和试点项目，为中央层面的规则制定和政策实施积累经验教训。不仅是中国和英国，新加坡、日本乃至澳大利亚、新西兰这些制度迥异的国家，都致力于建立一个能够有效“陪伴”人工智能技术和产业发展的监管体制。

当然，各国 AI 治理在实践上有很多共通之处。划分的这三种模式属于社会科学所谓的“理想类型”。我们可以借此看到不同的国家和地区在 AI 治理模式方面的显著特征。这些“理想类型”是对现实制度实践中若干关键变量进行提炼后的抽象结构。本文选取了立法、执法、司法以及央地关系中的一些代表性特征。各国 AI 治理进程中的举措纷繁复杂，很多可能只是偶然或昙花一现。我们在研究中努力选取那些相对稳定的结构性的代表性特征。例如，各国 AI 治理模式背后更深层次的制度传统。

欧盟与美国都并非单一制国家，在超国家治理与联邦制的制度架构下，中央政府在人工智能领域的监管权限原本相对有限，需要依赖成员国或地方政府的立法与执行。两个政治体看似相反的 AI 治理模式，实际上是对类似处境的不同反应。欧盟在成员国主权限制下，只能通过制定全面、严格、细致的预防性立法和设立全面监管的全新机构来打破原有的治理结构——其中心化方案正源于去中心化处境。而在美国，中央政府由于党派政治而难以通过全美性立法，监管部门各司其职，道路交通等应用领域又多由各州政府主导，法院长期避免对行政分支主动干预，因此人工智能的规制呈现出高度分散的碎片化特征。这些都不是短时间或者单一力量能够改变的。

相比之下，中英两国在政治体制上均具备较强的单一制色彩，中央政府无须制定全面立法就能够掌控局面，行政部门传统上具备较强的监管能力，能够在行政体系内直接统筹 AI 政策的规划与执行，同时与司法部门形成有效互动，从而有能力选择一种紧密高效的“伴侣型监管”路径。这种 AI 治理模式更具协调力和系统性，但也对国家的制度能力与组织水平提出了更高要求。因此，能够采取“伴侣型监管”模式的国家不可能很多，都是既致力于 AI 产业发展又同时具备协同治理能力的国家。而多数国家跟随欧盟范式选择防御式的监管策略也是无奈之举：本国的科技企业不具备引领 AI 发展的产业实力，而本国的政府机构也缺乏伴随性、贴近性的监管能力。或许只有美国这样的国家可以做到：本国的 AI 企业在全世界居于领先地位，而传统的监管部门和许多地方政府都具有较高的 AI 治理水平，不依赖中央政府的统一协调，仍能确保 AI 技术发展和产业应用不至于失控。从这一意义上看，这三种治理模式分别根植于更为基础性的特定的制度演化结构，实际上具有内在的逻辑性和必然性。

不过，“理想类型”并不意味着治理实践被僵化地“绑定”在某一模式中。相比于本文详述的四个代表国家或地区，其他各国的 AI 治理实践往往混合了三种模式中的多种元素，以适应其特定的政治、经济、社会和文化背景。特别是，随着人工智能的持续发展，一些国家可能会在不同阶段发生模式转化。例如，加拿大早期较多采纳美国式的“守夜人看护”逻辑，依赖既有机构进行松散的事后监督，但近年来在隐私保护与算法歧视领域日益趋向欧盟式的预防性规制，甚至提出了类似于欧盟的全面人工智能立法，反映出其政策重心与监管哲学的动态调整；而在全面 AI 立法失败后，则可能转向“伴侣型监管”。同样，一个国家可能从“守夜人看护”开始（原有机构的事后督察），并随着人工智能的迅猛发展而转向“伴侣型监管”（给予资金资助和政策支持的同时通过有针对性的约束规则）。“理想类型”的价值正在于为我们理解这一制度演变过程提供比较清晰的分析框架，而不是静态地定义国家治理的终局状态。

因此本文提出的关于 AI 治理模式的比较分析框架，有助于阐明为什么 AI 治理在世界各地看起来如此不同，又在许多方面展现出惊人的相似之处。一方面，尽管每个国家都宣称追求“可信、可靠、可控”的 AI 实践，都在尝试最大限度地发挥人工智能对社会的功用，同时最大限度地降低其风险，但是最终呈现的却是判然有别的治理路径。另一方面，“家长式防控”“守夜人看护”与“伴侣型监管”三种 AI 治理模式的区分，又在很大程度上可以概括并揭示出各国在技术创新与风险控制的平衡方面做出的典型选择。同时，这也并不妨碍我们在每种类型中再做区分。例如，“伴侣型监管”中对于硬法和软法工具的选择，以及行政和司法的关系，各国实践往往有所不同。这些亚型的区分同样有助于我们对于 AI 治理模式的深入理解。

每种模式都有其利弊得失。随着 AI 技术和应用的负面影响日益显著，美国式的“守夜人看护”模式虽有利于初期创新活力的释放，但也面临对系统性风险响应不足的困境，监管滞后带来的制度真空正逐渐累积为成本高昂的治理风险。而欧盟式的“家长式防控”模式尽管影响广泛，却因监管门槛过高、预设限制过多，容易扼杀本土企业的成长空间和发展机遇。相比之下，中国和英国共同实践的“伴侣型监管”模式，在鼓励 AI 创新的同时给予贴近和敏捷的有效监管，在全球 AI 治理中提供了欧美之外的另一个可供选择的实践方案，但是要实现这样一种贴近的、介入式的协同发展目标，政府和企业都需要做出格外的努力。

长远来看，一个国家如果致力于在 AI 技术、产业和应用领域拥有自主地位，必然要有与之匹配的政府能力。尤其是，要实现对 AI 这类飞速发展、高度复杂、自主性日增、嵌入性极强的新型技术的有效治理，追求一种更为紧密的“伴侣型监管”，政府不能仅仅作为严厉管控的父权制君主或者冷漠放任的自由主义管家，而是需要像已婚人士在长期的伴侣型婚姻中对待自己的配偶那样，给予热情支持并建立互信，保持沟通并营造共识，同时对其言行保持敏锐感知、高度关注和积极回应，在产生

分歧时就事论事，并在对方出格时给予及时警告乃至严厉惩罚，从而实现一段持久关系的协同演化。在 AI 治理上，立法和司法依旧可以有所作为，但总体而言难以胜任，能够发挥关键作用的主要是行政机关及其专业化的监管能力——一种足以匹敌和驾驭人工智能的人类智能。

参考文献

[1] 北京互联网法院，2021.（2021）京 0491 民初 30475 号民事判决书[Z]. 北京：北京互联网法院.

[2] 北京市人民代表大会常务委员会，2024. 北京市自动驾驶汽车条例[S]. 北京：北京市人民代表大会常务委员会.

[3] 重庆市人民政府，2016. 重庆市公共安全视频图像信息系统管理办法[R/OL].（2016-06-23）[2025-05-28]. http://www.cq.gov.cn/zwgk/zfxxgkml/szfwj/zfgz/zfgz/201606/t20160623_8836437.html.

[4] 工业和信息化部，2017. 促进新一代人工智能产业发展三年行动计划（2018-2020 年）[R/OL].（2017-12-13）[2025-05-28]. https://www.cac.gov.cn/2017-12/15/c_1122114520.htm.

[5] 工业和信息化部，市场监管总局，2025. 关于进一步加强智能网联汽车产品准入、召回及软件在线升级管理的通知[R/OL].（2025-02-28）[2025-05-28]. https://www.samr.gov.cn/zw/zfxxgk/fdzdgknr/zlfzs/art/2025/art_8126cfad252445428e7e5ccdec9f8df4.html.

[6] 国家互联网信息办公室，工业和信息化部，公安部，2022. 互联网信息服务深度合成管理规定 [R/OL].（2022-11-25）[2025-05-28]. https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm.

[7] 国家互联网信息办公室，工业和信息化部，公安部，国家市场监督管理总局，2021. 互联网信息服务算法推荐管理规定[R/OL].（2021-12-31）[2025-05-28]. https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.

[8] 国家互联网信息办公室，公安部，2025. 人脸识别技术应用安全管理办法[R/OL].（2025-03-13）[2025-05-28]. https://www.gov.cn/zhengce/zhengceku/202503/content_7016075.htm.

[9] 国家互联网信息办公室，国家发展和改革委员会，教育部，科学技术部，工业和信息化部，公安部，国家广播电视总局，2023. 生成式人工智能服务管理暂行办法[R/OL].（2023-07-10）[2025-05-28]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

[10] 国务院，2017. 新一代人工智能发展规划[R/OL].（2017-07-08）[2025-05-28]. https://app.www.gov.cn/govdata/gov/201707/20/408540/article.html.

[11] 杭州互联网法院，2023.（2023）浙 0192 民初 4563 号民事判决书[Z]. 杭州：杭州互联网法院.

[12] 杭州市人民代表大会常务委员会，2022. 杭州市社会信用条例[S]. 杭州：杭州市人民代表大会常务委员会.

[13] 交通运输部办公厅，工业和信息化部办公厅，公安部办公厅，人力资源社会保障部办公厅，人民银行办公厅，税务总局办公厅，市场监管总局办公厅，网信办秘书局，2022. 关于加强网络预约出租汽车行业事前事中事后全链条联合监管有关工作的通知[R/OL].（2022-02-07）[2025-05-28]. https://www.gov.cn/zhengce/zhengceku/2022-02/15/content_5673773.htm.

[14] 科技部，2019a. 国家新一代人工智能创新发展试验区建设工作指引[R/OL].（2019-08-29）[2025-05-28]. https://www.gov.cn/zhengce/zhengceku/2019-12/03/content_5457884.htm.

[15] 科技部，2019b. 新一代人工智能发展规划推进办公室召开 2019 年工作会议[R/OL].（2019-02-20）[2025-05-28]. https://www.most.gov.cn/kjbgz/201902/t20190220_145130.html.

[16] 深圳市人民代表大会常务委员会，2021. 深圳经济特区数据条例[S]. 深圳：深圳市人民代表大会常务委员会.

[17] 深圳市人民代表大会常务委员会，2022. 深圳经济特区人工智能产业促进条例[S]. 深圳：深圳市人民代表大会常务委员会.

[18] 张露，陈钧圣，2025. 在广州，“人工智能 +”落地开花[N]. 广州日报，01-03(04).
[19] 浙江省杭州市富阳区人民法院，2020.（2019）浙 0111 民初 6971 号民事判决书[Z]. 杭州：浙江省杭州市富阳区人民法院.
[20] 最高人民法院，2019. 中国法院的互联网司法[M]. 北京：人民法院出版社.
[21] 最高人民法院，2021. 关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的规定[S]. 北京：最高人民法院.
[22] ABBOTT K W, SNIDAL D, 2000. Hard and soft law in international governance[J]. International Organization, 54(3): 421-456.
[23] ABU DHABI GLOBAL MARKET, 2024. The ADGM legal framework[R]. Abu Dhabi: Abu Dhabi Global Market.
[24] ABU DHABI MEDIA OFFICE, 2024. In his capacity as ruler of Abu Dhabi, UAE president issues law establishing artificial intelligence and advanced technology council[R]. Abu Dhabi: Abu Dhabi Media Office.
[25] ADAMS R, 2020. A-levels and GCSE: how did the exam algorithm work? [N/OL]. BBC News, 2020-08-20 [2025-06-01]. https://www.bbc.com/news/explainers-53807730.
[26] AFRICAN UNION, 2022. The African Union artificial intelligence continental strategy for Africa[R]. Addis Ababa: African Union Commission.
[27] AI SINGAPORE, 2018. AI for industry (AI4I) programme[R]. Singapore: AI Singapore.
[28] ALAN TURING INSTITUTE, 2018. 2017/18 at the turing: a year in review[R]. London: The Alan Turing Institute.
[29] ALLEGHENY COUNTY ANALYTICS, 2019. Developing predictive risk models to support child maltreatment hotline screening decisions[R]. Pittsburgh: Allegheny County Department of Human Services.
[30] ALLEN H, 2019. Regulatory sandboxes[J]. George Washington Law Review, 87(3): 579-645.
[31] ANGWIN J, LARSON J, MATTU S, KIRCHNER L, 2016. Machine bias[EB/OL].（2016-05-23）[2025-06-01]. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[32] ARABIAN BUSINESS, 2017. UAE appoints first minister for artificial intelligence[EB/OL].（2017-10-19）[2025-05-28]. https://www.arabianbusiness.com/politics-economics/381648-uae-appoints-first-minister-for-artificial-intelligence.
[33] ASSOCIATION OF SOUTHEAST ASIAN NATIONS, 2024. ASEAN guide on AI governance and ethics[R]. Jakarta: Association of Southeast Asian Nations.
[34] BACHE I, FLINDERS M, 2004. Multi-level governance[M]. Oxford: Oxford University Press.
[35] BALDWIN R, CAVE M, LODGE M, 2011. Understanding regulation: theory, strategy, and practice[M]. Oxford: Oxford University Press.
[36] BENJAMIN R, 2019. Race after technology[M]. Cambridge: Polity Press.
[37] BERMANN G, 1994. Taking subsidiarity seriously[J]. Columbia Law Review, 94(2): 331-456.
[38] BERRY F, BERRY W, 2018. Innovation and diffusion models in policy research[M]. Boulder: Westview Press.
[39] BLACK J, 2001. Decentring regulation: understanding the role of regulation and self-regulation in a ‘post-regulatory’ world[J]. Current Legal Problems, 54(1): 103-146.
[40] BLACK J, 2010. Risk-based regulation: choices, practices and lessons[M]//OECD. Risk and regulatory policy: improving the governance of risk. Paris: OECD Publishing, 185-224.
[41] BLACK J, 2012. Paradoxes and failures: ‘new governance’ techniques and the financial crisis[J]. Modern Law Review, 75(6): 1037-1063.
[42] BRADFORD A, 2020. The brussels effect: how the European Union rules the world[M]. Oxford: Oxford University Press.
[43] BUOLAMWINI J, GEBRU T, 2018. Gender shades[J]. Proceedings of Machine Learning Research, 81: 1-15.
[44] BUSINESS & HUMAN RIGHTS RESOURCE CENTRE, 2021. Italy: court rules against Deliveroo’s rider

algorithm, citing discrimination [EB/OL]. (2021 - 01 - 05) [2025 - 06 - 01]. https://www.business - humanrights.org/en/latest - news/italy - court - rules - against - deliveroos - rider - algorithm - citing - discrimination/.

[45] CABINET OFFICE OF JAPAN, 2019. Social principles of human - centric AI[R]. Tokyo: Cabinet Office of Japan.

[46] CALIFORNIA DEPARTMENT OF MOTOR VEHICLES, 2023. DMV statement on Cruise LLC suspension for immediate release[EB/OL]. (2023 - 10 - 24)[2025 - 05 - 28]. https://www.dmv.ca.gov/portal/news - and - media/dmv - statement - on - cruise - llc - suspension/.

[47] CALIFORNIA OFFICE OF ADMINISTRATIVE LAW, 2014. Cal. Code Regs. Tit. 13, § § 227.00 - 227.54 [S]. California: California Office of Administrative Law.

[48] CALIFORNIA OFFICE OF ADMINISTRATIVE LAW, 2020. Cal. Code Regs. Tit. 13, § § 250.00 - 253.02 [S]. California: California Office of Administrative Law.

[49] CALIFORNIA STATE LEGISLATURE, 2018. California consumer privacy act, Cal. Civ. Code § 1798.100 et seq[S]. California: California State Legislature.

[50] CENTRE FOR DATA ETHICS AND INNOVATION, 2021. AI barometer[R]. London: Centre for Data Ethics and Innovation.

[51] CHAMBER OF DEPUTIES OF ARGENTINA, 2024. Bill 3003 - D - 2024 [S]. Buenos Aires: Chamber of Deputies of Argentina.

[52] CHAMBER OF DEPUTIES OF BRAZIL, 2021. Bill No. 21/2020[S]. Brasília: Chamber of Deputies of Brazil.

[53] CIRCUIT COURT OF COOK COUNTY, ILLINOIS, 2022. American civil liberties union v. Clearview AI, Inc., No. 2020 CH 04353 (Ill. Cir. Ct. Cook Cty. settlement approved May 19, 2022) [Z]. Chicago: Circuit Court of Cook County, Illinois.

[54] CIVIL AVIATION AUTHORITY OF SINGAPORE, 2021. Singapore designates One - North as first drone estate [EB/OL]. (2018 - 02 - 05) [2025 - 06 - 01]. https://www.caas.gov.sg/who - we - are/newsroom/Detail/singapore - designates - one - north - as - first - drone - estate.

[55] COMMISSION NATIONALE DE L'INFORMATIQUE ET DES LIBERTÉS, 2021. Intelligence artificielle et services publics: la CNIL publie le bilan de son《 bac à sable 》[EB/OL]. (2025 - 04 - 11) [2025 - 06 - 01]. https://www.cnil.fr/fr/bilan - bac - a - sable - IA - services - publics.

[56] COMPETITION AND MARKETS AUTHORITY, 2023. AI foundation models: initial report[R]. London: Competition and Markets Authority.

[57] CONGRESS, 2022. CHIPS and Science Act[S]. Washington D.C.: Congress.

[58] COOPER D, RYOKO M, DE MENESES A S O, 2025. Japan plans to adopt AI - friendly legislation[EB/OL]. (2025 - 03 - 24) [2025 - 06 - 01]. https://www.insideprivacy.com/international/japans - plans - to - adopt - ai - friendly - legislation/.

[59] COUNCIL OF EUROPE, 2024. Framework convention on artificial intelligence and human rights, democracy and the rule of law (CETS No. 225) [S]. Strasbourg: Council of Europe.

[60] COURT OF APPEAL (CIVIL DIVISION), 2020. Bridges v. South Wales police, Ewca civ 1058[Z]. London: Court of Appeal (Civil Division).

[61] CREMONA M, SCOTT J, 2019. EU law beyond EU borders[M]. Oxford: Oxford University Press.

[62] DE SADELEER N, 2002. Environmental principles: from political slogans to legal rules[M]. Oxford: Oxford University Press.

[63] DEPARTMENT OF INDUSTRY, SCIENCE AND RESOURCES, 2023. Safe and responsible AI in Australia: discussion paper[R]. Canberra: Department of Industry, Science and Resources.

[64] DEPARTMENT OF INDUSTRY, SCIENCE AND RESOURCES, 2024. Australian government's interim response to safe and responsible AI consultation[R]. Canberra: Department of Industry, Science and Resources.

[65] DEPARTMENT OF INFRASTRUCTURE, TRANSPORT, REGIONAL DEVELOPMENT, COMMUNICATIONS AND THE ARTS, 2024. Call to help shape the future of automated vehicles in Australia[R]. Canberra: Department of Infrastructure, Transport, Regional Development, Communications and the Arts.

[66] DEUTSCHES INSTITUT FÜR NORMUNG, DEUTSCHE KOMMISSION ELEKTROTECHNIK ELEKTRONIK INFORMATIONSTECHNIK IN DIN UND VDE, 2022. Deutsche normungsroadmap künstliche intelligenz – version 2[R]. Berlin: Deutsches Institut für Normung (DIN), Deutsche Kommission Elektrotechnik Elektronik Informationstechnik (DKE) in DIN und VDE.

[67] DIE BUNDESREGIERUNG, 2020. Strategie künstliche intelligenz der bundesregierung[S]. Berlin: Die Bundesregierung.

[68] DIET OF JAPAN, 2019. Partial amendment of the road transport vehicle act (act No. 14 of 2019)[S]. Tokyo: National Diet of Japan.

[69] DIET OF JAPAN, 2022. Partial amendment of the road traffic act (act No. 32 of 2022)[S]. Tokyo: National Diet of Japan.

[70] DIET OF JAPAN, 2025. Act on the promotion of research and development, and utilization of AI-related technologies (act No. 53 of 2025)[S]. Tokyo: National Diet of Japan.

[71] DIGITAL REGULATION COOPERATION FORUM, 2022. Digital regulation cooperation forum: plan of work for 2022 to 2023[R]. UK: Digital Regulation Cooperation Forum.

[72] DISTRICT COURT OF THE HAGUE, 2020. Case C-09-550982-HA ZA 18-388 (ECLI: NL: RBDHA: 2020: 865)[S]. The Hague: District Court of The Hague.

[73] DUBAI INTERNATIONAL FINANCIAL CENTRE, 2024. Comprehensive laws and regulations in Dubai[R]. Dubai: Dubai International Financial Centre.

[74] EQUALITY AND HUMAN RIGHTS COMMISSION, 2022. Equality watchdog takes action to address discrimination in use of artificial intelligence[R]. England: Equality and Human Rights Commission.

[75] ESAFETY COMMISSIONER, 2015. About eSafety[EB/OL]. (2024-12-20)[2025-06-01]. https://www.esafety.gov.au/about-us/what-we-do.

[76] EUROPEAN CENTER FOR NOT-FOR-PROFIT LAW, 2023. Hungary's new biometric surveillance laws violate the AI act[EB/OL]. (2025-04-28)[2025-06-01]. https://ecnl.org/news/hungarys-new-biometric-surveillance-laws-violate-ai-act.

[77] EUROPEAN COMMISSION, 2020. Communication-two years of application of the GDPR, COM (2020) 264 final[R]. Brussels: European Commission.

[78] EUROPEAN COMMISSION, 2021a. Coordinated plan on artificial intelligence 2021 review[R]. Brussels: European Commission.

[79] EUROPEAN COMMISSION, 2021b. Horizon Europe: The EU research and innovation programme (2021-2027)[R]. Brussels: European Commission.

[80] EUROPEAN COMMISSION, 2021c. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union Legislative Acts, COM (2021) 206 final, article 59[R]. Brussels: European Commission.

[81] EUROPEAN COMMISSION, 2022. Headquarters of the Spanish agency for the supervision of artificial intelligence[R]. Brussels: European Commission.

[82] EUROPEAN COMMISSION, 2024. Commission decision establishing the European AI office[R]. Brussels: European Commission.

[83] EUROPEAN COMMISSION, 2025. The digital Europe programme[R]. Brussels: European Commission.

[84] EUROPEAN DATA PROTECTION BOARD, 2018. European data protection board rules of procedure[R]. Brussels: European Data Protection Board.

[85] EUROPEAN MEDICINES AGENCY, 2017. Good pharmacovigilance practices[R]. Amsterdam: European Med-

icines Agency.

[86] EUROPEAN UNION, 2000. Charter of fundamental rights of the European Union[S]. Strasbourg: European Union.

[87] EUROPEAN UNION, 2016. Regulation (EU) 2016/679 (general data protection regulation), article 83[S]. Brussels: European Union.

[88] EUROPEAN UNION, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)[S]. Brussels: European Union.

[89] EXECUTIVE OFFICE OF THE PRESIDENT, 2023. Safe, secure, and trustworthy development and use of artificial intelligence (executive order 14110)[S]. Washington, D. C: Executive Office of the President.

[90] FATTAH Z, MARTIN M, 2024. Saudis scale back ambition for $1.5 trillion desert project Neom[N]. Bloomberg News, 05-23.

[91] FEDERAL MINISTRY OF COMMUNICATIONS, INNOVATION & DIGITAL ECONOMY, 2023. National artificial intelligence strategy[EB/OL]. (2023-08-28)[2025-06-01]. https://fmcide.gov.ng/initiative/nais/.

[92] FEDERAL TRADE COMMISSION, 2020. Using artificial intelligence and algorithms[R]. Washington, D. C: Federal Trade Commission.

[93] FEDERAL TRADE COMMISSION, 2021. California company settles FTC allegations it deceived consumers about use of facial recognition in photo storage app[EB/OL]. (2021-01-11)[2025-06-01]. https://www.ftc.gov/news-events/news/press-releases/2021/01/california-company-settles-ftc-allegations-it-deceived-consumers-about-use-facial-recognition-photo.

[94] FEDERAL TRADE COMMISSION, 2024. FTC announces crackdown on deceptive AI claims and schemes[R]. Washington, D. C: Federal Trade Commission.

[95] FINANCIAL CONDUCT AUTHORITY, 2025. FCA set to launch live AI testing service[EB/OL]. (2025-04-29)[2025-06-01]. https://www.fca.org.uk/news/press-releases/fca-set-launch-live-ai-testing-service.

[96] FIRST-TIER TRIBUNAL (GRC), 2023. Clearview AI Inc. v. information commissioner, [2023] UKFTT 819 (GRC)[Z]. London: First-tier Tribunal (GRC).

[97] FLORIDI L, COWLS J, BELTRAMETTI M, CHATILA R, CHAZERAND P, DIGNUM V, LUETGE C, MADELIN R, PAGALLO U, ROSSI F, SCHAFER B, VALCKE P, VAYENA E, 2018. AI4People — an ethical framework for a good AI society[J]. Minds and Machines, 28(4): 689-707.

[98] FØLLESDAL A, 1998. Survey article: subsidiarity[J]. Journal of Political Philosophy, 6(2): 190-218.

[99] FRIEDMAN B, 1997. Valuing federalism[J]. Minnesota Law Review, 82: 317-412.

[100] G7 LEADERS, 2023. G7 Leaders' statement on the Hiroshima AI process[R]. Japan: G7/G20 Documents Database.

[101] GARDBAUM S, 1994. The nature of preemption[J]. Cornell Law Review, 79(4): 767-815.

[102] GLOBAL PARTNERSHIP ON AI FOUNDING MEMBERS, 2020. Joint statement from founding members of the global partnership on artificial intelligence[EB/OL]. (2020-06-15)[2025-06-01]. https://www.gov.uk/government/publications/joint-statement-from-founding-members-of-the-global-partnership-on-artificial-intelligence/joint-statement-from-founding-members-of-the-global-partnership-on-artificial-intelligence.

[103] GLOBAL PRIVACY ENFORCEMENT NETWORK, 2023. Action plan for the global privacy enforcement network (GPEN)[R]. Washington, D. C: Global Privacy Enforcement Network.

[104] GODOY J, 2025. AI regulation ban meets opposition from state attorneys general over risks to US consumers[N/OL]. Reuters, 2025-05-16[2025-06-01]. https://www.reuters.com/sustainability/boards-policy-regulation/ai-regulation-ban-meets-opposition-state-attorneys-general-over-risks-us-2025-05-16/.

[105] GOVERNMENT OF DUBAI, 2023. Law No. (9) of 2023 regulating the operations of autonomous vehicles in the Emirate of Dubai[S]. Dubai: Government of Dubai.

[106] GOVERNMENT OF FRANCE, 2023. Loi n° 2023-380 du 19 mai 2023 relative aux jeux olympiques et paralympiques de 2024[S]. Paris: Government of France.

[107] GOVERNMENT OF ISRAEL, 2024. ISRAEL AI: national AI program[R]. Jerusalem: Government of Israel.

[108] GOVERNMENT OF SAUDI ARABIA, 2019. Saudi Arabia royal order No. 471/A/1440 on the establishment of an authority called of (Saudi authority for data and artificial intelligence) (SDAIA)[S]. Riyadh: Government of Saudi Arabia.

[109] GOVERNMENT OF TAMIL NADU, 2020. Tamil Nadu safe and ethical artificial intelligence policy 2020[R]. Chennai: Government of Tamil Nadu.

[110] GOVERNMENT OF THE REPUBLIC OF KOREA, 2019. National strategy for artificial intelligence[S]. Seoul: Government of the Republic of Korea.

[111] GOVERNMENT TECHNOLOGY AGENCY OF SINGAPORE, 2020. New two factor authentication features in Singpass to provide more convenience and accessibility[EB/OL]. (2020-12-16)[2025-06-01]. https://www.tech.gov.sg/media/2020-12-16-singpass-2fa.

[112] GRAND NATIONAL ASSEMBLY OF TÜRKYE, 2024. Artificial intelligence law (draft bill) (2/2234)[R]. Ankara: Grand National Assembly of Türkiye.

[113] GRAY V, 1973. Innovation in the states[J]. American Political Science Review, 67(4): 1174-1185.

[114] HAAS E, 1958. The uniting of Europe: political, social, and economic forces[M]. Stanford: Stanford University Press.

[115] HALIMWEB, 2017. AISG: new national programme to catalyse, synergise and boost Singapores artificial intelligence capabilities[N/OL]. AI Singapore, 2017-05-02[2025-06-01]. https://aisingapore.org/aisg-new-national-programme-to-catalyse-synergise-and-boost-singapores-artificial-intelligence-capabilities/.

[116] HIGH COURT OF JUSTICE QUEEN'S BENCH DIVISION ADMINISTRATIVE COURT, 2020. Joint council for the welfare of immigrants v. the president of the upper tribunal (immigration and asylum chamber), ewhc 3103 (admin)[Z]. London: High Court of Justice Queen's Bench Division Administrative Court.

[117] HOME OFFICE, 2021. Surveillance camera code of practice[R]. London: Home Office.

[118] HOUSE OF COMMONS OF CANADA, 2022. Bill C-27, An act to enact the consumer privacy protection act, the personal information and data protection tribunal act and the artificial intelligence and data act and to make consequential and related amendments to other acts[S]. Ottawa: House of Commons of Canada.

[119] HUTTER B, 2005. The attraction of risk-based regulation: accounting for the emergence of risk ideas in regulation[EB/OL]. (2005-03)[2025-06-01]. https://www.lse.ac.uk/accounting/assets/CARR/documents/D-P/Disspaper33.pdf.

[120] IEEE, 2021. IEEE 7001-2021-IEEE standard for transparency of autonomous systems[S]. New York: IEEE.

[121] ILLINOIS GENERAL ASSEMBLY, 2008. Biometric information privacy act[S]. Springfield: Illinois General Assembly.

[122] ILLINOIS GENERAL ASSEMBLY, 2020. Artificial intelligence video interview act[S]. Springfield: Illinois General Assembly.

[123] INDIAN COUNCIL OF MEDICAL RESEARCH, 2023. Ethical guidelines for application of AI in biomedical research and healthcare[R]. New Delhi: Indian Council of Medical Research.

[124] INFOCOMM MEDIA DEVELOPMENT AUTHORITY, 2022a. AI verify: an AI governance testing framework and toolkit[EB/OL]. (2022-05-25)[2025-06-01]. https://file.go.gov.sg/aiverify-primer.pdf.

[125] INFOCOMM MEDIA DEVELOPMENT AUTHORITY, 2022b. IMDA and PDPC launch Singapore's first privacy enhancing technologies sandbox as they mark decade-long effort of strengthening public trust[EB/OL]. (2022-07-20)[2025-06-01]. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2022/imda-and-pdpc-launch-sg-first-privacy-enhancing-technologies-sandbox.

[126] INFOCOMM MEDIA DEVELOPMENT AUTHORITY, 2023. First of its kind Generative AI evaluation sandbox for trusted AI by AI verify foundation and IMDA[EB/OL]. (2023-10-31)[2025-06-01]. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox.

[127] INFOCOMM MEDIA DEVELOPMENT AUTHORITY, AI VERIFY FOUNDATION, 2024a. Model Al governance framework for Generative Al[EB/OL]. (2024-05-30)[2025-05-28]. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/gen-ai-and-digital-foss-ai-governance-playbook.

[128] INFOCOMM MEDIA DEVELOPMENT AUTHORITY, AI VERIFY FOUNDATION, 2024b. Proposed model AI governance framework for Generative AI[R]. Singapore: Infocomm Media Development Authority, AI Verify Foundation.

[129] INFORMATION COMMISSIONER'S OFFICE, 2020. Guidance on the AI auditing framework[R]. Wilmslow: Information Commissioner's Office.

[130] INFORMATION COMMISSIONER'S OFFICE, 2023a. AI and data protection risk toolkit[R]. Wilmslow: Information Commissioner's Office.

[131] INFORMATION COMMISSIONER'S OFFICE, 2023b. Guidance on AI and data protection[R]. Wilmslow: Information Commissioner's Office.

[132] INFORMATION COMMISSIONER'S OFFICE, 2023c. UK information commissioner issues preliminary enforcement notice against Snap[R]. Wilmslow: Information Commissioner's Office.

[133] INFORMATION COMMISSIONER'S OFFICE, 2024a. Data protection impact assessments[R]. Wilmslow: Information Commissioner's Office.

[134] INFORMATION COMMISSIONER'S OFFICE, 2024b. ICO consultation series on generative AI and data protection: the transparency duty[R]. Wilmslow: Information Commissioner's Office.

[135] INFORMATION COMMISSIONER'S OFFICE, 2024c. Regulating AI: the ICO's strategic approach[R]. Wilmslow: Information Commissioner's Office.

[136] INFORMATION & EGOVERNMENT AUTHORITY, KINGDOM OF BAHRAIN, 2024. Al Qaed: "Bahrain's adoption of AI in government sector in line with global privacy, security, and ethical standards"[EB/OL]. (2024-09-10)[2025-06-01]. https://www.iga.gov.bh/en/article/al-qaed-bahrains-adoption-of-ai-in-government-sector-in-line-with-global-privacy-security-and-ethical-standards.

[137] INNOVATION, SCIENCE AND ECONOMIC DEVELOPMENT CANADA, 2023. Voluntary code of conduct on the responsible development and management of advanced generative AI systems[R]. Ottawa: ISED Canada.

[138] ISO/IEC, 2023. ISO/IEC 23053:2022 framework for artificial intelligence (AI) systems using machine learning (ML), 23894:2023 information technology-artificial intelligence-guidance on risk management[S]. Geneva: ISO/IEC.

[139] ISRAEL INNOVATION AUTHORITY, 2022. Israel innovation authority iaunches new incubators program[EB/OL]. (2022-02-22)[2025-06-01]. https://innovationisrael.org.il/en/press_release/israel-innovation-authority-launches-new-incubators-program/.

[140] ISRAEL PRIVACY PROTECTION AUTHORITY, 2025. Draft guideline of the privacy protection authority: applicability of the provisions of the privacy protection law to artificial intelligence systems[R]. Israel: Israel Privacy Protection Authority.

[141] JAPAN FAIR TRADE COMMISSION, 2021. Report on algorithms/AI and competition policy[R]. Tokyo:

Japan Fair Trade Commission.

[142] JORDANA J, LEVI-FAUR D, 2005. The politics of regulation[M]. Cheltenham: Edward Elgar.

[143] KASHIWAGI R, 2017. The rise of the regulatory sandbox[N]. Financial IT, 04-12.

[144] KNESSET, 1981. Privacy protection law, 5741-1981[S]. Jerusalem: Knesset.

[145] KUNER C, BYGRAVE L A, DOCKSEY C, 2020. The EU general data protection regulation (GDPR)[M]. Oxford: Oxford University Press.

[146] MARCH J G, OLSEN J P, 1989. Rediscovering institutions: the organizational basis of politics[M]. New York: The Free Press.

[147] MARKS G, HOOGHE L, 2001. Multi-level governance and European integration[M]. Lanham: Rowman & Littlefield.

[148] MARYLAND GENERAL ASSEMBLY, 2024. MD. Code Ann., Lab. & Empl, § 3-717 [S]. Maryland: Maryland General Assembly.

[149] MEDICINES AND HEALTHCARE PRODUCTS REGULATORY AGENCY, 2024a. AI airlock: the regulatory sandbox for AIaMD[R]. London: Medicines and Healthcare Products Regulatory Agency.

[150] MEDICINES AND HEALTHCARE PRODUCTS REGULATORY AGENCY, 2024b. Medical devices: compliance and enforcement of the regulations[R]. London: Medicines and Healthcare Products Regulatory Agency.

[151] MÉNY Y, 1993. Government and politics in Western Europe[M]. Oxford: Oxford University Press.

[152] MINISTRY OF DIGITAL DEVELOPMENT AND INFORMATION, 2024. Opening keynote by Minister Josephine Teo at asia tech x artificial intelligence (ATxAI) conference[R]. Singapore: Ministry of Digital Development and Information.

[153] MINISTRY OF ECONOMIC AFFAIRS AND COMMUNICATIONS OF ESTONIA, 2019. Estonia's national artificial intelligence strategy 2019—2021[R]. Tallinn: Ministry of Economic Affairs and Communications of Estonia.

[154] MINISTRY OF ECONOMY, TRADE AND INDUSTRY, MINISTRY OF INTERNAL AFFAIRS AND COMMUNICATIONS, 2022. Governance guidelines for implementation of AI principles (version 1.1)[R]. Tokyo: Ministry of Economy, Trade and Industry, Ministry of Internal Affairs and Communications.

[155] MINISTRY OF FOREIGN AFFAIRS, MINISTER OF STATE FOR ARTIFICIAL INTELLIGENCE AND DIGITAL ECONOMY AND REMOTE WORK APPLICATIONS OFFICE, 2024. UAE position on AI policy [R]. Abu Dhabi: Ministry of Foreign Affairs, Minister of State for Artificial Intelligence and Digital Economy and Remote Work Applications Office.

[156] MINISTRY OF SCIENCE, TECHNOLOGY, KNOWLEDGE AND INNOVATION OF CHILE, 2021. National artificial intelligence policy[R]. Santiago: Ministry of Science, Technology, Knowledge and Innovation of Chile.

[157] MONETARY AUTHORITY OF SINGAPORE, 2016. Fintech regulatory sandbox guidelines[R]. Singapore: Monetary Authority of Singapore.

[158] MONETARY AUTHORITY OF SINGAPORE, 2018. Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore's financial sector[R]. Singapore: Monetary Authority of Singapore.

[159] MONETARY AUTHORITY OF SINGAPORE, 2019. MAS partners financial industry to create framework for responsible use of AI[R]. Singapore: Monetary Authority of Singapore.

[160] MONETARY AUTHORITY OF SINGAPORE, 2024. Who we are[R]. Singapore: Monetary Authority of Singapore.

[161] MORAVCSIK A, 1998. The choice for Europe[M]. Ithaca: Cornell University Press.

[162] NATIONAL AI ADVISORY COMMITTEE, 2022. Working group reports[EB/OL]. (2022-05-04)[2025-

06-01]. https://data.aclum.org/storage/2025/01/NAIAC_ai_gov_naiac.pdf.

[163] NATIONAL AI INITIATIVE OFFICE, 2020. H. R. 6216 - national artificial intelligence initiative act of 2020 [S]. Washington D. C.: United States Congress.

[164] NATIONAL ASSEMBLY OF THE REPUBLIC OF KOREA, 2024. Framework act on artificial intelligence development and establishment of a foundation for trustworthiness [S]. Seoul: National Assembly of the Republic of Korea.

[165] NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION, 2017. Automated driving systems 2.0: a vision for safety[R]. Washington D. C.: U. S. Department of Transportation.

[166] NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION, OFFICE OF DEFECTS INVESTIGATION, 2024. ODI resume for EA22 - 002 (Tesla autopilot) [R]. Washington D. C.: U. S. Department of Transportation.

[167] NATIONAL INSTITUTE FOR HEALTH AND CARE RESEARCH, 2021. AI in health and care award-funded projects 2021[R]. London: National Institute for Health and Care Research.

[168] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, 2023. AI risk management framework (AI RMF 1.0), NIST AI 100-1[R]. Gaithersburg: National Institute of Standards and Technology.

[169] NEW JERSEY LEGISLATURE, 2022-2023. N. J. assembly bill A4909. regulates use of automated tools in hiring decisions to minimize discrimination in employment[S]. Trenton: New Jersey Legislature.

[170] NEW YORK CITY COUNCIL, 2021. NYC local law 144 of 2021[S]. New York: New York City Council.

[171] NEW YORK STATE LEGISLATURE, 2023-2024. N. Y. assembly bill A567. establishes criteria for the use of automated employment decision tools[S]. Albany: New York State Legislature.

[172] NITI AAYOG, 2018. National strategy for artificial intelligence: #AIforAll[R]. New Delhi: NITI Aayog.

[173] NITI AAYOG, 2021. Responsible AI # AIFORALL: approach document for India part 1—principles for responsible[R]. New Delhi: NITI Aayog.

[174] NYC DEPARTMENT OF CONSUMER AND WORKER PROTECTION, 2023. Automated employment decision tools[R]. New York: NYC Department of Consumer and Worker Protection.

[175] OECD, 2020. OECD AI policy observatory (OECD. AI)[EB/OL]. (2025-07-10)[2025-07-15]. https://oecd.ai/en/dashboards/policy-initiatives/oecd-ai-policy-observatory-oecdai-9635.

[176] OFCOM, 2010. What is Ofcom[R]. London: Office of Communications.

[177] OFFICE OF THE PRIVACY COMMISSIONER OF NEW ZEALAND, 2024. Biometric processing privacy code -draft guide[R]. Wellington: Office of the Privacy Commissioner of New Zealand.

[178] PARLIAMENT OF AUSTRALIA, 2024. Privacy and other legislation amendment act 2024[S]. Canberra: Parliament of Australia.

[179] PARLIAMENT OF INDIA, 2023. The digital personal data protection act, 2023 (No. 22 of 2023) [S]. New Delhi: Parliament of India.

[180] PARLIAMENT OF NEW ZEALAND, 1986. Fair trading act 1986 (1986 No. 121) [S]. Wellington: Parliament of New Zealand.

[181] PARLIAMENT OF NEW ZEALAND, 2015. Harmful digital communications act 2015 (2015 No. 63)[S]. Wellington: Parliament of New Zealand.

[182] PEETERS B, 2020. Facial recognition at Brussels Airport: face down in the mud[EB/OL]. (2020-03-17)[2025-06-01]. https://www.law.kuleuven.be/citip/blog/facial-recognition-at-brussels-airport-face-down-in-the-mud/.

[183] PELKMANS J, 2006. European integration: methods and economic analysis[M]. Toronto: Pearson Education Canada.

[184] PERSONAL DATA PROTECTION COMMISSION SINGAPORE, 2019. Model artificial intelligence governance framework (first edition) [R]. Singapore: Personal Data Protection Commission Singapore.

[185] PERSONAL DATA PROTECTION COMMISSION SINGAPORE, 2020a. Implementation and self-assessment guide for organisations[R]. Singapore: Personal Data Protection Commission Singapore.

[186] PERSONAL DATA PROTECTION COMMISSION SINGAPORE, 2020b. Model artificial intelligence governance framework (second edition)[R]. Singapore: Personal Data Protection Commission Singapore.

[187] PERSONAL DATA PROTECTION COMMISSION SINGAPORE, 2022. Guide on the responsible use of biometric data in security applications[R]. Singapore: Personal Data Protection Commission Singapore.

[188] PERSONAL DATA PROTECTION COMMISSION SINGAPORE, 2023. Launch of the AI verify foundation to shape the future of AI standards through global collaboration[EB/OL]. (2023-06-07)[2025-06-01]. https://www.pdpc.gov.sg/news-and-events/announcements/2023/06/launch-of-ai-verify-foundation-to-shape-the-future-of-ai-standards-through-collaboration.

[189] PERSONAL DATA PROTECTION COMMISSION SINGAPORE, 2024a. About us[EB/OL]. (2023-11-8)[2025-07-15]. https://www.pdpc.gov.sg/who-we-are/about-us.

[190] PERSONAL DATA PROTECTION COMMISSION SINGAPORE, 2024b. Advisory guidelines on the use of personal data in AI recommendation and decision systems[R]. Singapore: Personal Data Protection Commission Singapore.

[191] PERSONAL INFORMATION PROTECTION COMMISSION OF JAPAN, 2024. Interim report on considerations for the triennial review of the act on protection of personal information[R]. Tokyo: Personal Information Protection Commission of Japan.

[192] PIERSON P, 2004. Politics in time: history, institutions, and social analysis[M]. Princeton: Princeton University Press.

[193] RAJI I D, SMART A, WHITE R N, MITCHELL M, GEBRU T, HUTCHINSON B, SMITH-LOUD J, THERON D, BARNES P, 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing[C]. New York: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, 33-44.

[194] REPUBLIC OF KENYA, 2025. Kenya artificial intelligence strategy 2025-2030[R]. Nairobi: Republic of Kenya.

[195] RESERVE BANK OF INDIA, 2017. Master direction—information technology framework for the NBFC sector[R]. Mumbai: Reserve Bank of India.

[196] ROMANO R, 1985. Law as a product: some pieces of the incorporation puzzle[J]. Journal of Law, Economics, and Organization, 1(2): 225-283.

[197] ROYAL COMMISSION INTO THE ROBODEBT SCHEME, 2023. Report of the Royal Commission into the Robodebt Scheme[R]. Canberra: Royal Commission into the Robodebt Scheme.

[198] SAUDI DATA & AI AUTHORITY, 2020. National strategy for data & AI[R]. Riyadh: Saudi Data & AI Authority.

[199] SAUDI DATA & AI AUTHORITY, 2023. AI ethics principles[R]. Riyadh: Saudi Data & AI Authority.

[200] SAUDIPEDIA, 2020. Global AI summit[R]. Riyadh: Saudipedia.

[201] SCHÜTZE R, 2018. European Union law[M]. Cambridge: Cambridge University Press.

[202] SECURITIES AND EXCHANGE BOARD OF INDIA, 2024. Master circular for stock brokers[R]. Mumbai: Securities and Exchange Board of India.

[203] SHAFFER G, POLLACK M A, 2010. Hard vs. soft law: alternatives, complements, and antagonists in international governance[J]. Minnesota Law Review, 94(3): 706-799.

[204] SMART DUBAI, 2019. AI ethics principles and guidelines[R]. Dubai: Smart Dubai.

[205] SMART NATION AND DIGITAL GOVERNMENT OFFICE, SINGAPORE, 2019. National artificial intelligence strategy[R]. Singapore: Smart Nation and Digital Government Office.

[206] SMART NATION SINGAPORE, 2023. National artificial intelligence strategy 2.0[R]. Singapore: Smart

Nation Singapore.

[207] SNYDER F, 1994. Soft law and institutional practice in the European community[M]//MARTIN S. The construction of Europe. Dordrecht: Springer Dordrecht, 197-225.

[208] SOLOVE D, SCHWARTZ P, 2021. Information privacy law[M]. Alphen aan den Rijn: Wolters Kluwer.

[209] SOLUZIONI LAVORO. IT, 2020. Giurisprudenza - tribunale di Bologna - ordinanza 31 dicembre 2020[N]. Soluzionilavoro. it, 12-31.

[210] STANDARDS NEW ZEALAND, 2022. Framework for artificial intelligence (AI) systems using machine learning (ML)[S]. Wellington: Standards New Zealand.

[211] SUNSTEIN C, 2005. Laws of fear: beyond the precautionary principle[M]. Cambridge: Cambridge University Press.

[212] SUPERIOR COURT OF NEW JERSEY, APPELLATE DIVISION, 2021. State v. Pickett, 466 n. j. super. 270 (app. div. 2021)[Z]. Newark: Superior Court of New Jersey, Appellate Division.

[213] SUPREME COURT OF THE UNITED STATES, 1932. New state ice Co. v. liebmann, 285 U. S. 262[Z]. Washington D. C.: Supreme Court of the United States.

[214] SUPREME COURT OF THE UNITED STATES, 2024. Loper Bright Enterprises v. Raimondo, 603 U. S. 369 (2024)[Z]. Washington D. C.: Supreme Court of the United States.

[215] THE ASSOCIATED PRESS, 2022. Oregon is dropping an artificial intelligence tool used in child welfare system [EB/OL]. (2022-06-02)[2025-06-01]. https://www. npr. org/2022/06/02/1102661376/oregon-drops-artificial-intelligence-child-abuse-cases.

[216] THE DAILY TRIBUNE—NEWS OF BAHRAIN, 2024. Bahrain shura council approves law to curb AI abuse [N]. News of Bahrain, 04-29.

[217] THE ECONOMIC TIMES, 2023. Government not considering regulating AI growth, says IT Minister Vaishnaw [N]. The Economic Times, 04-05.

[218] THE SCOTTISH GOVERNMENT, 2021. Scotland's artificial intelligence strategy: trustworthy, ethical, and inclusive[R]. Edinburgh: The Scottish Government.

[219] TOSHIO YOKOYAMA, 2021. "ROAD to the L4" symposium: launch of the "research and development and social implementation project for advanced mobility services such as level 4 automated driving"—overview of the project[R]. Tokyo: National Institute of Advanced Industrial Science and Technology.

[220] U. S. DEPARTMENT OF JUSTICE, 2022. Justice Department secures groundbreaking settlement agreement with Meta platforms, formerly known as Facebook, to resolve allegations of discriminatory advertising [R]. Washington, D. C.: U. S. Department of Justice.

[221] U. S. DEPARTMENT OF JUSTICE, CFPB, FTC, EEOC, 2023. Joint statement on enforcement efforts against discrimination and bias in automated systems[N]. Federal Trade Commission, 04-25.

[222] U. S. DISTRICT COURT, EASTERN DISTRICT OF MICHIGAN, 2024. Williams v. City of Detroit, No. 21-10827 (E. D. Mich. settled)[Z]. Detroit: U. S. District Court, Eastern District of Michigan.

[223] U. S. DISTRICT COURT, NORTHERN DISTRICT OF CALIFORNIA, 2023. Mobley v. Workday, Inc., No. 3: 23-cv-00770-hsg (n. d. cal. filed Feb. 21, 2023)[Z]. San Francisco: U. S. District Court, Northern District of California.

[224] U. S. EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, 2022. The Americans with disabilities act and the use of software, algorithms, and artificial intelligence to assess job applicants and employees: technical assistance document[R]. Washington, D. C.: U. S. Equal Employment Opportunity Commission.

[225] U. S. FOOD AND DRUG ADMINISTRATION, 2021. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan[R]. Silver Spring: U. S. Food and Drug Administration.

[226] U. S. HOUSE COMMITTEE ON ENERGY AND COMMERCE, 2025. Artificial intelligence and information technology modernization initiative. § 43201, 119th Cong[R]. Washington, D. C.: U. S. House Committee

on Energy and Commerce.

[227] U. S. SECURITIES AND EXCHANGE COMMISSION, 2023. SEC proposes new requirements to address risks to investors from conflicts of interest associated with the use of predictive data analytics[R]. Washington, D. C.: U. S. Securities and Exchange Commission.

[228] UAE AI OFFICE, 2022. AI ethics: principles & guidelines[R]. Abu Dhabi: UAE AI Office.

[229] UAE AI OFFICE, 2024. The UAE Charter for the development & use of artificial intelligence[R]. Abu Dhabi: UAE AI Office.

[230] UAE GOVERNMENT, 2018a. Federal Decree-Law No. (25) of 2018 on the projects of future nature [S]. Abu Dhabi: United Arab Emirates Government.

[231] UAE GOVERNMENT, 2018b. National strategy for artificial intelligence 2031[R]. Abu Dhabi: United Arab Emirates Government.

[232] UK DEPARTMENT FOR SCIENCE, INNOVATION AND TECHNOLOGY, 2023a. A pro-innovation approach to AI regulation[R]. London: UK Department for Science, Innovation and Technology.

[233] UK DEPARTMENT FOR SCIENCE, INNOVATION AND TECHNOLOGY, 2023b. Introducing the AI safety institute[R]. London: UK Department for Science, Innovation and Technology.

[234] UK GOVERNMENT, 2021. National AI strategy[R]. London: UK Government.

[235] UK GOVERNMENT, 2025. AI opportunities action plan[R]. London: UK Government.

[236] UK PARLIAMENT, 2018. Data protection act 2018, chapter 12[S]. London: UK Parliament.

[237] UK PARLIAMENT, 2023. Online safety act 2023[S]. London: UK Parliament.

[238] UK PARLIAMENT, 2024. Automated vehicles act 2024[R]. London: UK Parliament.

[239] UNESCO, 2021. Recommendation on the ethics of artificial intelligence[R]. Paris: UNESCO.

[240] UNITED NATIONS GENERAL ASSEMBLY, 2024. Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development[R]. New York: United Nations General Assembly.

[241] UTAH STATE LEGISLATURE, 2025. Utah Code Ann., § 77-23F-101 ET SEQ[S]. Salt Lake City: Utah State Legislature.

[242] VIRGINIA LEGISLATURE, 2021. VA. Code Ann., facial recognition technology; approval; penalty. § 15.2-1723.2[S]. Richmond: Virginia Legislature.

[243] VOGEL D, 1997. Trading up: consumer and environmental regulation in a global economy[M]. Cambridge: Harvard University Press.

[244] WHITE HOUSE, 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence[S]. Washington, D. C.: The White House.

[245] WHITE HOUSE OFFICE OF SCIENCE AND TECHNOLOGY POLICY, 2022. Blueprint for an AI bill of rights [R]. Washington, D. C.: White House Office of Science and Technology Policy.

[246] WISCONSIN SUPREME COURT, 2016. State v. Loomis, 881 n. w. 2d 749 (wis. 2016) [Z]. Madison: Wisconsin Supreme Court.

[247] YIGITCANLAR T, KAMRUZZAMAN M, BUYS L, IOPPOLO G, SABATINI-MARQUES J, DA COSTA E M, YUN J J, 2018. Understanding 'smart cities': intertwining development drivers with desired outcomes in a multidimensional framework[J]. Cities, 81: 145-160.

[248] ZHU X, ZHAO H, 2021. Experimentalist governance with interactive central—local relations: making new pension policies in China[J]. Policy Studies Journal, 49(1): 13-36.

Paternal Prevention, Guardian Supervision, and Companion Regulation
—A Comparative Framework of Global AI Governance

Bin Ling* Runyi Ma
(Law School, Peking University)

Summary: The study offers a systematic and comparative framework of three ideal-typical models of global artificial intelligence (AI) governance: Paternal Prevention (exemplified by the European Union), Guardian Supervision (characteristic of the United States), and Companion Regulation (manifested in differing forms in China and the United Kingdom). Through in-depth analysis across four key governance dimensions—legal tools and policy orientation, administrative enforcement and structure, judicial review and mechanisms of checks and balances, and local experimentation and power allocation—the study reveals the normative foundations, institutional logics, and strategic approaches each model employs to balance AI innovation with risk mitigation. Particular emphasis is placed on the rising prominence of Companion Regulation as a potentially adaptive and globally influential governance path.

The Paternal Prevention model pursued by the European Union embodies a risk-averse, precautionary logic, centered on preemptively constraining AI applications that may infringe upon fundamental rights. Anchored in the binding AI Act, the EU constructs a unified, risk-tiered regulatory framework that applies directly across member states. This framework mandates ex-ante compliance measures for high-risk systems, such as human oversight and conformity assessments, while prohibiting certain high-risk uses altogether. The EU's approach is supported by a multilevel enforcement architecture, including the European AI Board and designated national supervisory authorities, with substantial sanctioning powers. The model is undergirded by a strong judiciary capable of reviewing both administrative actions and legislative compliance, further reinforcing fundamental rights. However, the supranational and centralized nature of this model limits member states' autonomy and scope for localized experimentation, potentially constraining innovation flexibility.

By contrast, the Guardian Supervision model exemplified by the United States emphasizes post hoc oversight within a market-oriented, innovation-friendly environment. Lacking a comprehensive

* **Corresponding Author:** Bin Ling, Law School, Peking University, E-mail: lingbin@pku.edu.cn.

federal AI law, the U. S. relies on a decentralized patchwork of sector-specific regulations, supplemented by soft law instruments such as the NIST AI Risk Management Framework and executive guidance like the "Blueprint for an AI Bill of Rights". Enforcement is fragmented across existing agencies (e. g. , FTC, FDA, EEOC), with no centralized authority for AI regulation. The judiciary intervenes only after harm has occurred, adjudicating AI-related disputes through the application of general legal principles rather than AI-specific norms. Local jurisdictions, particularly states and municipalities, serve as regulatory innovators, adopting diverse measures that reflect localized priorities but also contribute to regulatory fragmentation. This model privileges technological dynamism but raises concerns about delayed responses to systemic harms and governance incoherence.

The Companion Regulation model, observed in both China and the United Kingdom, seeks to align public governance with industry innovation through flexible, collaborative, and context-sensitive regulatory mechanisms. In the UK, this model is instantiated through a "pro-innovation" approach that emphasizes principles-based, sector-led guidance over comprehensive legislative codification. Regulators such as the Information Commissioner's Office and Financial Conduct Authority lead AI oversight within their sectors, supported by coordination platforms like the Digital Regulation Cooperation Forum. Judicial interventions, as seen in key cases on facial recognition and algorithmic bias, reinforce rights-based accountability. While the UK system is more centralized than that of the U. S. , it still permits targeted experimentation through regulatory sandboxes and devolved competencies.

China's version of Companion Regulation is more interventionist and regulatory. It combines robust top-down mandates with strategic state—industry coordination. Regulatory instruments include binding measures for specific technologies (e. g. , generative AI, recommendation algorithms), supported by broader legal frameworks such as the Cybersecurity Law and the Personal Information Protection Law. Enforcement is led by the Cyberspace Administration of China and implemented through a vertically integrated regulatory matrix spanning multiple ministries. While judicial review plays a supplementary role, local pilot zones (e. g. , in Shanghai and Beijing) enable experimentation with regulatory approaches under central guidance. Successful local practices are often scaled nationally, reflecting a model of iterative governance rooted in strong administrative capacity.

This tripartite framework also applies to understanding other countries in global AI governance. For example, South Korea, Bahrain, Brazil, Canada, and Turkey, as well as many international forums, tend toward EU-style Paternal Prevention, while India, Saudi Arabia, the UAE, and Israel are closer to U. S. -style Guardian Supervision, and Singapore, Japan, Australia, and New Zealand exhibit Companion Regulation characteristics similar to those of the UK and China.

The study concludes that these three models represent divergent responses to the governance

challenges posed by AI's rapid development and social entrenchment. The EU model prioritizes legal certainty and rights protection through preemptive regulation; the U. S. approach champions innovation and institutional pluralism but often lags in anticipatory oversight; the China's and UK pathway, through different institutional arrangements, attempt to harmonize regulatory responsiveness with developmental goals. Among these, Companion Regulation emerges as a particularly salient alternative, offering a dynamic balance between flexibility and control. Its success, however, depends on the state's capacity to deploy technical expertise, coordinate across sectors, and adapt regulatory strategies in real time. As AI technologies continue to evolve, this model—grounded in adaptive governance and collaborative oversight—may offer a more effective pathway toward a responsible and sustainable AI future.

Keywords: Artificial Intelligence Governance; Regulatory Mechanisms; Innovation-Risk Balance; Paternal Prevention; Guardian Supervision; Companion Regulation

JEL Classification: K23; K20; O33