人工智能安全发展的法治体系构建

江必新[1]　胡慧颖[2]

【内容摘要】 在全球科技创新密集活跃的智能时代背景下，将人工智能治理纳入国家治理体系现代化的大局中，实现“AI 善治”，是中国式法治现代化建设的必然要求。这意味着要通过制度创新、技术创新和管理创新，推动法治体系向智能化、精细化、高效化方向发展。构建人工智能安全发展的法治体系的基本框架，应涵盖以下几个方面：一是系统完备的法律规范体系，为人工智能的发展提供明确的法律指引和规范；二是高效协同的法治实施体系，确保法律法规得到有效执行；三是公开透明的安全监管体系，对人工智能应用进行监督和管理；四是有力的法治保障体系，为人工智能的健康、可持续发展提供坚实支撑。同时，还应抓住构建人工智能法治体系的核心智能要素，如算法可信、数据安全和 AI 伦理，以实现创新发展与安全监管的动态平衡。

【关 键 词】 人工智能法　法治体系　智能治理　法治现代化

【作　　者】 1 江必新，第十三届全国人民代表大会宪法和法律委员会副主任委员，湖南大学特聘教授。（北京 100745）

2 胡慧颖，湖南智慧法治研究院助理研究员。（长沙 410082）

【基金项目】 国家社科基金重大项目“建设中国特色社会主义法治体系研究”（232DA072）

当前全球科技创新进入密集活跃期，人工智能这一前沿技术实现了多点突破，引发了链式变革。人工智能作为引领新一轮科技革命与产业变革的战略性技术，因其无处不在的渗透性、扩散性、带动性，广泛赋能经济社会发展，同时以前所未有的速度驱动数字生态系统的深刻变革，对人类生活方式、全球产业结构、经济形态、社会进步以及国际政治经济格局等都产生了重大而深远的影响。在此背景下，党的二十届三中全会审议并通过的《中共中央关于进一步全面深化改革推进中国式现代化的决定》（以下简称《决定》），明确将人工智能列为八大重点支持的未来产业之一，并在经济高质量发展、高水平对外开放、国家安全等多个关键领域的体制机制改革中均赋予其重要地位，这充分彰显了国家对人工智能领域创新发展的高度重视。值得强调的是，《决定》

还创造性提出了“完善生成式人工智能发展和管理机制”以及“建立人工智能安全监管制度”等一系列关键性改革举措。[①]这些举措不仅是对人工智能发展与安全并重原则的深刻践行，更是对潜在风险挑战的积极回应与前瞻布局。因此，亟待构建一套科学、系统、高效的人工智能法治体系，推动人工智能技术的健康、可持续发展，引领并重塑未来社会形态的智能化转型，迈向一个更加智能、更加安全、更加美好的中国式现代化新时代。

构建人工智能法治体系的必要性

构建人工智能法治体系的必要性主要体现在以下几个方面。

第一，加快发展人工智能是促进新质生产力发展的重要引擎。人工智能技术正引领生产力由传统的“物之力”“能之力”向“智之力”的智能化进阶。[②]“当前，互联网、大数据、云计算、人工智能、区块链等新技术深刻演变，产业数字化、智能化、绿色化转型不断加速，智能产业、数字经济蓬勃发展，极大改变全球要素资源配置方式、产业发展模式和人民生活方式。”[③]“谁能把握大数据、人工智能等新经济发展机遇，谁就把准了时代脉搏。”[④]智业文明时代的到来，推动技术革命性突破、生产要素创新性配置、产业深度转型升级，带来新的发展机遇与挑战。《决定》针对“健全因地制宜发展新质生产力体制机制”，提出加强关键共性技术、前沿引领技术、现代工程技术、颠覆性技术创新，完善推动包括人工智能在内的八大战略性产业发展政策和治理体系，纵深推进人工智能这一战略性新兴产业、未来产业健康有序发展。这是发展新质生产力、赢得发展主动权的时代要求，也是全面贯彻新发展理念、扎实推动经济高质量发展的现实需要。[⑤]在以人工智能推进高质量发展的过程中，法治的护航作用不可或缺。

第二，实现人工智能创新发展与安全监管之间的动态平衡，是实现高质量发展的必然要求。高质量发展是全面建设社会主义现代化国家的首要任务，而人工智能作为具有广泛影响力的颠覆性技术，无疑是实现高质量发展的重要引擎。人工智能须被建构主体赋予价值驱动或任务驱动才能实现意识，人工智能系统往往被设计为遵循预设的规则或算法来执行任务，其“智能”表现受限于编程者的知识和预设条件。[⑥]故而，人工智能所塑造的新“人类世”赋予人类前所未有的福祉与机遇，也伴随着一系列风险，包括虚拟与现实割裂、技术失控反噬、改变就业结构、冲击法律与社会伦理、侵犯个人隐私、挑战国际关系准则等，对政府管理、经济安全、社会稳定乃至全球治理都将产生深远而复杂的影响。[⑦]当前，我国人工智能治理正进入技术创新与安全监管持续并重、动态平衡的关键阶段，其价值定位在于平衡发展创新与安全有益之间的内在张力。为此，必须把构建人工智能法治体系置于首要位置，坚持发展和安全并重、促进创新和依法治理相结合，汲取“先发展、后治理”的深刻教训，摒弃以牺牲安全为代价的粗放增长模式，实现“边发展、边治理”。同时，应准确把握技术和产业发展趋势，充分认识和评估每一项颠覆性创新可能存在的漏洞或盲点，及时采取措施加以应对。这将为人工智能的创新发展配备“安全阀”，最大化其正面效应，最小化其潜在风险，确保人工智能安全、可靠、可控、高质量发展。

第三，加强人工智能治理是智能时代全球国家治理的必然要求。随着全球科技竞争的日益激烈，竞争的焦点已跨越单一技术或设备层面，扩展至涵盖新型智能基础设施、数字文明构建以及制度生态或法治生态创新等多维度、全方位的综合较量。在此背景下，人工智能的安全监管已成为各国高度重视的议题。美国和英国均采取灵活且鼓励创新的监管策略，其中，美国通过制定地

方性法规、提倡行业自律等多种手段，构建了人工智能安全标准；英国则选择充分利用现有法律框架与监管机构进行监管。而欧盟则于2024年出台了全球首个统一人工智能监管的法案——《人工智能法案》，对高风险应用实施严格监管。[8]我国目前采取相对审慎的“小快灵”立法模式，通过制定“小切口”的伦理规范和政策法规，对人工智能的某些特定风险进行精准规制，同时为人工智能的发展创新留足空间。此外，我国还针对不同领域制定了场景化的监管规则，如互联网信息服务的深度合成技术监管规定等。[9]在全球层面，我国积极参与并推动全球人工智能治理框架与标准规范的建立，通过签署《布莱切利宣言》、提出《全球人工智能治理倡议》等举措，[10]致力于推动在全球范围内建立具有广泛共识的人工智能治理框架和标准规范，不断提升人工智能技术的安全性、可靠性、可控性和公平性。构建人工智能安全发展的法治体系，不仅是回应人工智能技术快速发展的必然要求，更是参与和引领人工智能全球治理的战略抉择。

第四，构建人工智能法治体系是中国式法治现代化的必要举措。人工智能技术的迅猛发展对现有国家治理体系构成严峻挑战。这一挑战的核心，即技术与法律之间的“科林格里奇困境”（Collingridge’s Dilemma），具体表现为人工智能技术的快速迭代与法律体系相对稳定之间的矛盾。这也直接导致天然滞后的现行法律法规难以迅速适应人工智能技术日新月异的步伐，在应对数字化变革所引发的新型风险时显得力有不逮。究其原因：其一，当前理论界对人工智能监管问题的研究探讨多集中于民法、刑法与知识产权法等传统法律领域，需进一步加强对技术监管、公平竞争等新兴法律议题的探索，以构建更加全面有效的法治体系。其二，我国现有人工智能规范体系还存在时间差、粗板块、空白区等问题，对于人工智能监管主要依赖规章、政策性文件、行业标准及技术标准等“软法”引导，其力度和效果尚显不足。[11]其三，数据安全与隐私保护、算法歧视与偏见、责任归属不清、过度依赖与法治侵蚀等经典技术法治问题依然突出。因此，必须构建适应人工智能发展创新的法治框架，整合多学科力量，从理论、立法和实践等多个层面来加强人工智能相关法律、伦理、社会问题的研究，建立健全保障人工智能健康发展的法律法规、制度体系和伦理道德。[12]这不仅是对实体规则秩序的反思与重构，更是以“未来法学”[13]问题为主旨的有益探索，对于推动中国式法治现代化具有重要意义。

构建人工智能法治体系的基本框架

在当前进一步全面深化改革、加速推进中国式现代化的关键时期，针对人工智能领域的立法工作，不能仅局限于单一的安全风险防控层面，而应秉持更加多元化的视角，致力于构建一个集促进科技创新、防范技术风险等多重目标于一体的综合性法治框架。因此，构建人工智能安全发展的法治体系，至少应包括以下内容：

第一，系统完备的人工智能法律规范体系。应坚持包容审慎与设置底线相结合、分类处理与分阶段推进相结合等立法策略，建立更加完善的人工智能法律法规、伦理规范和政策体系。其中，包容原则强调的是对人工智能技术多样性和快速迭代特性的充分认知，为创新活动预留发展空间；审慎原则体现为对法律条文严谨性和可操作性的高标准要求；设置底线是要求任何有关人工智能的设计、生产和运用都不能危害国家安全、侵犯公共利益和他人权益，都不得危害人的尊严和价值；分类处理要求同等情况同等对待，不同情况区别处理，且这种差别对待必须是正当的并说明理由；而分阶段推进则是基于技术成熟度和具体应用场景的实践考量，先逐步推进人工智能领域

的基础性、专项性立法工作，待时机成熟后再开展综合性的统一立法。

在构建纵向立法体系的过程中，还需特别注意以下几个方面的问题：一是建立健全保障人工智能创新发展的基础性法律框架，明确人工智能领域各类主体的权利、义务及责任归属等内容，并构建追溯问责机制。二是加快制定和完善与人工智能应用相关的专门法律，重点完善网络安全、数据安全、知识产权保护、个人信息处理、伦理风险规制、算法训练等专项领域的规定，加快研究制定自动驾驶、服务机器人等应用基础较好的细分领域的相关安全管理法规。三是完善科技伦理审查标准，构建包含多层次的伦理道德判断结构及人机协作的伦理框架。四是寻求多角度、常态化的人工智能监管国际磋商合作，积极参与全球人工智能治理国际规则制定，深入研究人工智能领域的重大国际共性问题，贡献中国治理方案。此外，应特别注重跨部门、跨领域的立法协调一致性，以确保各项法规之间的衔接和互补，增强法律实施的可操作性与实效性。

第二，高效协同的人工智能法治实施体系。该体系需确保在执法、司法、守法等各个环节的协调高效运作。在执法环节上，应加大对人工智能相关法律法规的执行力度，制定人工智能领域的执法程序规范，确保各项规定得到严格遵循。在司法环节上，应探索构建适应人工智能发展的司法审判机制，设立专门审判庭或审判团队，制定专门诉讼规则和程序，以增强司法对人工智能案件的审理与裁判力度，提升司法审判效能；应避免对司法机关尚难准确把握、但市场接受度高且受欢迎的技术业态产品采取过激的强制措施，以最大限度地减少司法活动对新技术发展的潜在负面影响；还需加强司法与立法的协同配合，为人工智能立法提供丰富的实践经验和案例支持。此外，还应注重执法、司法队伍的建设，通过系统培训教育提升执法人员、司法人员的专业素养与执法、司法能力，确保其能够胜任日益复杂的人工智能执法、司法工作。在守法环节上，应加大人工智能领域的法治宣传与教育力度，提高社会各界对人工智能相关法律法规的认知度和遵守度，引导社会各界自觉遵循法律法规，积极履行社会责任与道德义务。

第三，公开透明的人工智能安全监管体系。该体系可采用设计问责和应用监督并重的双层监管结构，其涵盖以下关键要素：一是监管对象的全面覆盖，既要规范设计者、开发者、制造者，也要将平台建设者、传播者及使用者纳入监管范围。二是监管流程的全链条设定，既涵盖人工智能算法设计、产品开发、成果应用等全生命周期，也要对人工智能运行过程中的数据、算法等核心要素进行智能实时监控。三是监管重点的精准把握，既对核心软件设施实施严格监管，通过制定严格的形式规则和实体正义规则，确保人工智能系统的决策过程可追溯、可解释、可审计、可问责，同时亦高度重视相关硬件设施的安全可控，确保其符合安全标准和隐私保护要求。四是监管机制和手段的创新，综合运用建立专门人工智能监管机构、推动跨部门跨领域协同监管、推广“沙盒监管”、[14]健全行业守信激励与失信惩戒机制等多种监管方式，形成对人工智能应用的监管合力。五是加强国际监管合作，推动形成统一的国际监管标准和机制。

第四，有力的人工智能法治保障体系。要统筹各领域资源，打好法治、市场、科技、政策、人才“组合拳”：一是加强党对人工智能法治保障工作的全面领导，深入贯彻《决定》精神，确保工作方向正确。二是完善促进数字产业化和产业数字化发展的政策体系，强化财政、税收、金融、价格等多方面的政策保障，鼓励社会资本参与人工智能法治保障体系建设，形成多元化的投入机制。三是运用现代信息技术为人工智能治理体系赋能，构建包括动态的人工智能研发应用评估评价机制、安全监测预警机制、跨领域的人工智能测试平台认证机制在内的人、技、物、管相配套的安全防护体系，确保人工智能安全可控。四是加快培养人工智能高端人才和高素质法治

人才专门队伍，推进人工智能法治的专业建设、学科建设以及人才储备和梯队建设，着力培养既懂法律又懂人工智能的复合型人才。五是综合运用多种策略方式，如媒体宣传、教育普及、透明度提升、案例分析、法律咨询、政策参与、伦理讨论、在线资源开发、社区参与、跨学科合作等，增强公众对人工智能应用领域的法治思维与法律认知，推动技术应用的责任感落实，同时强化对潜在风险的预防管控能力。

构建人工智能法治体系的若干智能要素

新一代人工智能依托“大模型 + 大数据 + 大算力”的路径，正以前所未有的速度推动人类逼近通用人工智能。在这一进程中，虽然技术取得了诸多突破性进展，但伴随而来的“长尾风险”持续涌现，“治理 AI 技术”与“AI 技术治理”的双向治理逻辑愈加凸显。一方面，“治理 AI 技术”强调从政策、法律、伦理等层面来规范人工智能技术的研发、应用及社会影响，确保技术发展符合公共利益、法律法规和伦理道德要求。另一方面，“AI 技术治理”侧重于利用人工智能技术本身来优化治理过程，实现技术发展与治理需求的良性互动，提升治理效能。在人工智能治理的宏观视域下，受综合治理与体系治理理念的深刻影响，构建智能要素的法治保障框架尤为重要，其为人工智能等前沿技术在社会各个领域中的广泛应用与深度融合提供坚实保障。

（一）算法可信

在人工智能技术演进过程中，传统的以算法可解释性和透明度为核心的控制主义治理范式陷入困境。原因在于，生成式人工智能（Generative AI）及其正迈向的通用人工智能（AGI，Artificial General Intelligence）都是基于“贝叶斯更新”效应[15]的算法模型迭代而涌现，直觉涌现机制和深度神经网络等复杂模型的不透明性，导致了所谓的“AI 黑箱”问题——人们无法完全理解人工智能的决策过程以及无法精准预测其输出内容。针对人工智能技术的快速迭代、高度复杂及难以预测的特性，其治理过程亟须深入洞悉技术机理，精准把握发展脉络，全面考量技术创新与治理的均衡和谐，依据相关技术原理建构具有针对性的体制机制。

有鉴于此，人工智能技术治理的策略需要从追求算法的“完全透明”，转向构建“模型可信”的框架。具体而言，应推动 AI 技术标准化，实施明确、精准、敏捷的常态化技术监管，实现人工智能大模型创新与防范算法共谋风险之间的平衡。

首先，合理设定算法模型。一是完善体系化治理，围绕算法、数据、算力三大关键要素，贯穿算法计算、机器学习与自动决策的大模型全生命周期，深入开展人工智能技术系统治理研究。二是强化算法信任，构建以透明度、可控性、可预测性为基础的算法信任体系，融入“数字人本主义”理念，利用算法备案、算法透明及解释权机制，揭开算法黑箱，消减歧视与偏见。三是优化算力布局，依法合理调配算力资源，规划算力基础设施建设布局，推动公共算力资源平台的建设与利用，提高算力资源的高效利用和开放共享。

其次，严格设定数据输入端规则。一要确保数据采集软硬件符合国家标准与行业规范；二要加强个人信息保护，明确数据采集的边界和范围，采用加密技术、匿名化处理等手段保护个人隐私；三要建立数据合规审查机制，及时处理并纠正“问题数据”，确保数据客观、全面、合法合规。

再次，硬性控制结果输出端。一是实施分级分类治理，根据应用场景、服务类型的特点以及

风险程度等因素，实施差异化监管措施，确保监管的针对性和有效性；二是针对用户制定相应的规制措施和标准，引导用户在使用人工智能技术时遵守法律、伦理和社会规范；三是优化人工智能技术与用户的互动环境，通过改善界面设计、增加用户反馈机制等方式提升用户体验。同时，关注特殊用户群体需求，确保技术能够惠及更广泛的人群。

（二）数据安全

《人工智能安全标准化白皮书（2023 版）》中提到，人工智能系统及其相关数据的网络安全属性体现为机密性、完整性、可用性以及系统应对恶意攻击的能力。特别是在迈向通用人工智能的背景下，数据作为人工智能技术研发、训练与广泛应用的关键资源，其质量与安全成为人工智能实践应用在技术原理层面的稳固保障。然而，随着大模型训练对数据需求量的激增和数据处理复杂度的提升，数据安全风险也显著攀升。如何在充分利用海量数据的同时保护公共利益，始终是人工智能治理的重要议题。这一重要议题可归纳为保障数据供给和维护数据合法性两方面。

在保障数据供给层面，首要任务在于确保数据的高质量与多样性，训练数据要满足准确性、完整性、时效性和代表性，以提升模型的鲁棒性和适应性。其次，促进数据共享与整合，应构建高效的数据流通生态，通过完善数据交易机制、推动公共数据开放政策实施、构建数据资源共享与统筹整合平台等，打破领域与机构壁垒，形成更全面的数据集，以确保高质量数据要素的可获得性、可用性和可靠性。针对数据跨境流动问题，应建立国际数据治理合作共享机制，促进全球数据资源的优化配置，同时加强跨境数据流动的法律监管，确保数据跨境流动的安全与合法。再次，数据安全与隐私保护至关重要。应实施数据分级分类保护策略，明确数据安全保障责任，强化数据泄露风险监控，构建严格的加密脱敏与访问控制体系，限制对敏感数据的访问。最后，优化内容治理机制，通过构建溯源、黑名单、用户反馈等机制，实现对恶意内容的实时监测、严格审核与有效过滤，确保机器认知框架免受不良内容侵蚀。

在维护数据合法性层面，构建多元监管机制，提升数据安全治理监管能力。具体而言，一是通过设立独立的监督机构或引入第三方审计机构，对数据处理活动实施定期与不定期的审查与评估，及时发现并纠正违规行为。同时，加强跨部门协同监管与线上线下一体化监管，形成监管合力，确保数据处理的合法性与规范性。二是强化用户数据主体权利保障，包括知情权、选择权、更正权、删除权等。对于个人信息处理行为，应贯彻最小化收集、知情同意等原则，确保信息主体的合法权益得到充分保障。在特定情境下，若数据处理仅用于纯粹算法训练且未实质性影响个人权益，可视为非个人信息处理行为，但需设定严格的限制条件，如“明确告知”与“算法训练纯技术性”等。三是实施数据全生命周期监管。采用先进技术手段如数据脱敏、敏感信息风险评估、泄漏检测、使用监控及数据库保密检查等，对数据输入、运算、储存、输出等各个环节实施全方位、全链条的合规审查与监控。四是强化数据合规意识与国际合作，引导企业和个人自觉完善数据合规工作，营造全社会尊重数据权益、遵守数据规则的良好氛围。同时，积极参与全球数字领域标准、规则的制定工作，推动构建公平、合理、高效的国际数据治理体系，通过加强国际合作与交流，共同应对数据治理中的挑战与问题。

（三）AI 伦理

技术伦理在防止技术滥用、保障公共利益方面发挥着重要作用。在“智能爆炸”的美好泡影之下，人工智能等前沿科技所蕴含的风险远不止技术滥用、隐私侵犯及算法偏见等显性伦理与法律挑战。更深层次上，新技术将对社会系统产生结构性冲击，引发一系列深层次伦理风险，包括

但不限于科学认知有限性下的未知风险、生命权利受损的不可逆后果、研发目标与人本权益冲突的价值两难、物理与数字世界界限日益模糊导致的“真相”消失、技术责任难以界定与问责的监管困境，以及国际科技霸权主义触发的国际政治伦理风险等。[16]

在我国，科技伦理治理的核心策略在于促进创新与防范风险并重，客观评估并审慎对待科技伦理风险。因此，要实现科技伦理对人工智能的更好规制，首要任务是全面提升对前沿科技伦理风险的系统认知，深化对人工智能的系统性伦理分析，力求全面洞察、前瞻预判并理性澄清其核心伦理争议，同时探讨伦理价值层面的应对策略。其二，推动法律、伦理、技术的深度融合，构建人工智能伦理治理框架，完善伦理审查与监管体系，针对重大人工智能技术项目实施严格的伦理评估与审查。其三，遵循比例原则，在有效监管与避免过度干预之间寻求平衡，确保监管措施既精准又适度，以防抑制企业创新与市场竞争的活力。其四，共同构建前沿科技伦理软着陆机制，推进科技伦理研究、传播和教育，引导企业与开发者主动进行伦理治理；加强与科技和产业部门的协同治理，如通过“人机价值对齐工程”[17]等实践探索解决技术原理层面的价值冲突；全面提升社会公众的科技伦理素养，共同营造负责任的 AI 发展环境。

鉴于技术的通用化、未来创新周期与方向的不确定性，人工智能所引发的风险呈现出系统性社会风险的显著特征，甚至有技术专家警示未来的智能模型可能具有摧毁人类文明存续的力量。[18]因此，构建面向未来的人工智能法治框架，需要处理好智能要素与法治保障的关系。一方面，人工智能法治框架的构建，需要对已有数字立法进行回顾与整理，从人工智能的角度将现行智能要素立法的有益经验加以归纳运用；另一方面，对于目前尚未达成共识的问题要敢于“留白”，等待未来智能要素立法的补充和完善，譬如数据财产权等问题。[19]

人工智能法治体系：从技术逻辑出发、回归人类需求

回顾法律历史的演进，科学技术的发展进步不仅推动了传统法律体系的理念、方式、要素的创新，而且会对深层次的基础认知逻辑、研究范式、具体规则架构乃至新的制度建构产生重要影响。[20]人工智能驱动的数字法治生态革新，其内涵远不止于数字技术、数字经济或数字社会，更将深层次地引领法治形态的重塑与法学研究范式的开创性转变。

人工智能安全发展的法治体系构建，无疑是一项复杂而系统的工程。这一进程历经早期的顶层设计，具体实施的伦理规范、政策支持、技术标准和法律规制，以及后续的监管指引、基准测试、可信认证、风险评估等多种治理范式，逐步形成了相对完善的 AI 治理系统。一些原本游离于传统法学研究边缘的议题，如科技伦理、数据安全、算法安全等，已逐步成为智能要素法治保障框架的核心内容。同时，智能时代下的“数字法学”这一新兴的交叉学科，也在经历着从研究模式相对不成熟向研究范式体系化、科学化的转变过程。[21]

未来已来，当下必须将人工智能治理纳入国家治理体系现代化的大局中。构建人工智能法治体系应始终坚定“从技术逻辑出发、回归人类需求”的理念，以技术创新与安全可信为目标，全面更新治理理念、治理内容和治理策略，通过面向未来的制度治理与过程治理，最终实现“AI 善治”，让人工智能技术更好地服务于人类福祉。此外，还应坚持法学理论创新与法治实践发展相结合，一方面强化人工智能法学理论研究与跨学科融合，以“智慧法治”和法治中国建设为主题，展开跨学科、多专业、大视野的深入研讨，重点推进计量法学、实证法学、数字法学、行政自动

化法学等新兴法学交融发展，聚焦数字化带来的物理世界相关基础制度建构、社会规则重塑、伦理认知对齐，破解人工智能创新发展中的治理难题；另一方面，持续关注并探索人工智能技术对法律制度和社会实践的实际需求，通过制度创新、技术创新和管理创新，推动法治体系向更加智能化、精细化、高效化的方向发展，以促进国家治理体系现代化，并推进人工智能的现代化进程。

注释：

①《中共中央关于进一步全面深化改革　推进中国式现代化的决定》，北京：人民出版社，2024年，第34、41页。

② 贺福初：《解放创造力》，《光明日报》2016年5月27日，第11版。

③ 习近平：《习近平向2023中国国际智能产业博览会致贺信》，《人民日报》2023年9月5日，第1版。

④ 习近平：《构建高质量伙伴关系　开启金砖合作新征程——在金砖国家领导人第十四次会晤上的讲话》，《人民日报》2022年6月24日，第2版。

⑤ 参见《〈中共中央关于进一步全面深化改革、推进中国式现代化的决定〉辅导读本》，北京：人民出版社，2024年，第31—32页。

⑥ 柳下弈：《柳下解答李德毅院士通用人工智能十疑问》，《智能系统学报》2021年第1期。

⑦《党的二十届三中全会〈决定〉学习辅导百问》，北京：学习出版社、党建读物出版社，2024年，第207页。

⑧ 参见欧盟《人工智能法案》第4b条。欧盟2024年5月通过的《人工智能法案》第4a—4c修正案中提出监管生成型人工智能的核心条款，要求任何可用于高风险应用的“通用人工智能系统”，如就业、医疗、信用评分、行政、执法等领域，都必须初步遵守《人工智能法案》对高风险系统规定的全部义务。

⑨ 例如，在司法领域，最高人民法院《关于规范和加强人工智能司法应用的意见》中提出了“辅助审判”原则，强调“无论技术发展到何种水平，人工智能都不得代替法官裁判，人工智能辅助结果仅可作为审判工作或审判监督管理的参考，确保司法裁判始终由审判人员作出，裁判职权始终由审判组织行使，司法责任最终由裁判者承担”。

⑩ 我国在第三届“一带一路”国际合作高峰论坛上，发布了《全球人工智能治理倡议》，倡导各国在人工智能治理中加强信息交流和技术合作，共同做好风险防范，推动在全球范围内建立具有广泛共识的人工智能治理框架和标准规范，不断提升人工智能技术的安全性、可靠性、可控性和公平性。

⑪ 例如《生成式人工智能服务安全基本要求》《新一代人工智能发展规划》《新一代人工智能伦理规范》《国家新一代人工智能标准体系建设指南》《网络安全标准实践指南——生成式人工智能服务内容标识方法》《关于规范和加强人工智能司法应用的意见》等。

⑫ 参见科学技术部编写组编：《深入学习习近平关于科技创新的重要论述》，北京：人民出版社，2023年，第382页。

⑬ 未来学这一概念由德国学者O. K. 福莱西泰姆于1943年首次提出。法学与未来学交叉产生了未来法学学科，它是法学研究与新兴技术的融合产物，旨在研究未来社会关系对既有法律体系和法学理论的冲击及其应对。参见张本才：《未来法学论纲》，《法学》2019年第7期。

⑭“沙盒监管”是一种创新的监管模式，指先划定一个范围，对在“盒子”里面的企业、技术或产品，采取包容审慎的监管措施，同时防止问题扩散到“盒子”外，属于在可控范围内实行容错纠错机制，并由监管部门对运行过程进行全过程监管，以保证测试的安全性并作出最终评价。

⑮ 贝叶斯更新效应，是一种统计方法，它基于贝叶斯定理，用于在已有先验知识的基础上，随着新证据的出现，逐步更新对某个事件的概率估计。这种方法在机器学习和人工智能领域尤为重要，当模型接收到新的数据时，会依据这些数据调整其内在的概率分布，也就是更新模型对各种假设或参数的信念。借此，智能体可以从每一次的观察和交互中学习，不断地改进其模型和策略，从而变得更加“聪明”。

⑯ 段伟文：《前沿科技的深层次伦理风险及其应对》，《人民论坛·学术前沿》2024年第1期。

⑰ 人工智能的价值对齐（AI Alignment），源自2017年OpenAI研究人员发表的《依托人类偏好的深度强化学习》，指的是要确保人工智能系统的行为符合人类的目标、偏好、价值观或伦理道德考量。在此基础上，可以通过人类反馈的强化学习机制，实现对算法偏见的矫正，进而促进算法结果的客观性。参见斯图尔特·罗素、彼得·诺维格：《人工智能：现代方法》下册，张博雅等译，北京：人民邮电出版社，2022年，第778页。

⑱ L. Tredinnick and C. Laybats, “The Dangers of Generative Artificial Intelligence,” *Business Information Review*, vol.4, no.21, 2023.

⑲ 杨建军等：《人工智能法：必要性与可行性》，《北京航空航天大学学报》（社会科学版）2024年第3期。

⑳ 周佑勇：《论智能时代的技术逻辑与法律变革》，《东南大学学报》（哲学社会科学版）2019年第5期。

㉑ 赵精武：《论数字法学的概念与研究定位——兼论我们需要什么样的人工智能法》，《华东政法大学学报》2024年第4期。

编辑　李　梅　高　原