

我国人工智能安全治理

本章从战略规范、法律法规、伦理治理、行政监管和标准规范等方面介绍我国人工智能安全治理体系，使读者对我国的人工智能安全治理概况有全面的了解和掌握。

3.1 我国人工智能治理情况概述

随着人工智能的广泛应用，其带来的安全风险和挑战也日益凸显，加强人工智能安全治理已成为全球共识。加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控。

我国高度重视人工智能技术的发展与安全治理，在 2015 年国务院发布的《关于积极推进“互联网+”行动的指导意见》中首次提出“培育发展人工智能新兴产业”。此后，我国对人工智能的重视程度不断提高，持续在战略层面对发展人工智能做出部署。我国在人工智能安全治理方面采取积极措施，政府高度重视人工智能技术的发展和治理，通过制定相关政策和法规来规范人工智能技术的发展和治理，强调技术应用的合法合规性，并重视保护用户数据和隐私，致力于加强人工智能的治理和监督机制。

从治理思路上看，我国人工智能治理正走向综合治理、精细治理，探索构建“柔性的伦理”与“硬性的法律”并行的人工智能治理体系。国家对人工智能产业发展与技术创新方面给予大量政策支持，也出台了一系列针对性治理与治理的法律法规，并逐步加强和明确人工智能的基本科技伦理规范。此外，人工智能治理涉及技术标准、法律法规和政策框架等多个领域的协调与合作，需要政府、社会、市场等利益相关主体共同参与、共同治理。我国政企学各界正积极推动人工智能安全的研究与合作，通过分享经验实践共同应对人工智能安全挑战。总体来看，我国人工智能安全治理呈现的特点如下。

3.1.1 治理体系不断完善

近年来，我国持续加强和完善人工智能顶层设计，密集出台各类治理政策。2017 年，国务院发布《新一代人工智能发展规划》，提出三步走的战略目标，覆盖人工智能理论和技术、产业发展、伦理规范等领域。



2019年,《新一代人工智能治理原则——发展负责任的人工智能》《新一代人工智能伦理规范》陆续出台,前者提出了人工智能治理的框架和行动指南,后者强调将伦理道德融入人工智能全生命周期。为应对数据和算法应用带来的风险,2021—2022年,我国密集出台了《数据安全法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》等多部政策文件。

2023年,为促进生成式人工智能健康发展与规范应用,中央网信办等七部门联合发布《生成式人工智能服务管理暂行办法》。至此,我国人工智能治理框架已初具雏形,《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》形成人工智能监管的“三足鼎立”之势,与《网络安全法》《数据安全法》《个人信息保护法》等共同构筑起当前人工智能治理基本框架。

3.1.2 治理方式双管齐下

从治理方式看,我国人工智能安全治理“规范立法”与“标准化建设”并举。立法监管方面,我国走在世界前列。

2020年,国家标准化委员会等五部门联合发布《国家新一代人工智能标准体系建设指南》,指导人工智能领域标准体系建设,并将安全和伦理标准作为核心组成部分。2021年印发的《国家标准化发展纲要》再次强调在人工智能领域开展标准化研究。

2022年,国家层面科技伦理治理的第一个指导性文件《关于加强科技伦理治理的意见》出台,要求在“十四五”期间重点加强人工智能、生命科学、医学等领域的科技伦理立法研究,及时推动将重要的科技伦理规范上升为国家法律法规。

2023年,人工智能法草案被列入国务院2023年立法工作计划,同年7月,我国发布了全球范围内针对生成式人工智能的首部专门立法——《生成式人工智能服务管理暂行办法》。相比立法监管,我国人工智能的标准规范更具灵活性和适应性,能够有效推进落实人工智能治理。

2023年,全国网络安全标准化技术委员会发布《人工智能安全标准化白皮书(2023版)》,指出了人工智能安全标准需求,提出了下一步开展人工智能安全标准化工作的建议,为规范引导人工智能安全标准化工作提供参考。

3.1.3 治理思路双线并行

从治理思路上看,我国既支持和鼓励人工智能创新发展,又针对潜在风险进行规制。先前发布的《互联网信息服务算法推荐管理规定》和《互联网信息服务深度合成管理规定》均以安全为导向,致力于通过制定规则和限制措施来控制技术应用的风险。最新发布的《生成式人工智能服务管理暂行办法》(下称“《办法》”)则提出坚持发展与安全并重、促进创新和依法治理相结合,强调实行包容审慎和分类分级监管,相比之前发布的征求意见



见稿，在多处放宽了监管要求，新增鼓励发展的有力措施。《办法》体现了监管生成式人工智能的中国探索与思考，坚持发展与安全并行的中国式治理思路也为全球人工智能治理开辟了新道路。

现有的政策文件体现出，我国人工智能治理一贯遵循了安全与发展并行的思路，以安全为底线护航人工智能产业发展，预计未来将呈现以下趋势。

1. 加强领域标准化顶层设计

我国在人工智能领域的顶层设计和法律规制方面正在稳步推进并日益完善的进程中。国务院在《关于印发新一代人工智能发展规划的通知》顶层设计明确采取“三步走”的战略目标，其中第二步战略（2020—2025年）要求“初步建立人工智能法律法规、伦理规范和政策体系，形成人工智能安全评估和管控能力”。各界持续推进人工智能治理领域的国家标准、行业标准、地方标准、团体标准等标准化工作。国家标准化管理委员会成立国家人工智能标准化总体组，全国信息技术标准化技术委员会设立人工智能分技术委员会，推进制定人工智能领域的国家标准。中央科技委员会组建国家科技伦理委员会，负责指导和统筹协调推进全国科技伦理治理体系建设工作。

2. 完善治理监管体系框架

基于我国现有人工智能法律法规，我国将在保持各项法规之间监管思路的延续性和逻辑自治的基础上，结合横纵两个框架的特点，从整体人工智能技术及应用领域进行通用立法，逐步完善人工智能治理监管体系框架，使其具有更高水平和更广泛的适用范围，并与我国现行科技、安全等相关法律法规相衔接。

3. 秉持包容审慎的监管态度

考虑到我国已经在人工智能关键领域适度放宽责任主体义务，对人工智能市场的发展释放出积极的信号。我国将“完善与创新相发展的科学监管方式”，并采用与国家宏观经济发展相适应的监管思路，通过监管沙盒、试点、示范应用等方式支持、促进人工智能新事物应用落地，鼓励人工智能特别在商业社会和研究应用领域的创新发展，并且激励外商对中国人工智能领域的投资。此外，我国出台一系列多元化、多层次的人工智能产业发展鼓励和支持政策，聚焦自动驾驶汽车和人工智能医疗等领域。在技术自主创新方面，鼓励和推动人工智能算法、算力、数据资源、框架、软件、芯片及配套软件平台等基础技术的全链路、全生态的技术自主创新。在应用场景创新方面，支持专精特新“小巨人”、独角兽、人工智能/互联网初创企业等积极开展场景创新。

4. 推进多元主体参与治理

我国鼓励各行业、各领域不断探索优化人工智能应用场景，支持行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等多元主体在人工智能技术创新、数据资源建设、转化应用、风险防范等方面开展协作。不同主体在人工智能治理框架下将扮演不同的角色，有助于确定社会权益的分配边界和规则，推进人工智能的良性可持续发展。



3.2 我国人工智能战略规划

我国积极倡导“以人为本”和“智能向善”，出台了一系列相关政策文件，促进人工智能产业发展和加强安全治理。总体来看，我国人工智能战略规划主要分为点状治理、回应治理和集中治理等三个阶段。

3.2.1 点状治理阶段（2013—2017 年）

从 2013 年到 2017 年，我国人工智能治理采取的是点状治理模式。这个阶段的人工智能还未能独立成为政策重点，多点状出现于其他政策文件中。2013 年，国务院发布的《关于推进物联网有序健康发展的指导意见》提出“经济社会智能化发展”，人工智能首次成为国家顶层政策文件的重要议题。2015 年《关于积极推进“互联网+”行动的指导意见》首次提出“培育发展人工智能新兴产业”。此后，国家对人工智能的重视程度不断提高，持续在战略层面对发展人工智能做出部署，促进人工智能产业发展和加强安全治理。2013—2017 年涉及人工智能的政策文件及主要内容如表 3-1 所示。

表 3-1 2013—2017 年涉及人工智能的政策文件及主要内容

时 间	文 件 名 称	涉及人工智能的要点内容
2013.2	《国务院关于推进物联网有序健康发展的指导意见》	经济社会智能化发展
2015.7	《国务院关于积极推进“互联网+”行动的指导意见》	“互联网+”人工智能；培育发展人工智能新兴产业
2015.8	《促进大数据发展行动纲要》	支持自然语言理解、机器学习、深度学习等人工智能技术创新，提升数据分析处理能力、知识发现能力和辅助决策能力
2016.3	《“十三五”规划纲要》	重点突破大数据和云计算关键技术、自主可控操作系统、高端工业和大型管理软件、新兴领域人工智能技术
2016.5	《“互联网+”人工智能三年行动实施方案》	培育壮大 AI 产业
2016.7	《“十三五”国家科技创新规划》	人工智能是新一代信息技术；重点开发移动互联、量子信息、人工智能等技术
2016.11	《“十三五”国家战略性新兴产业发展规划》	培育人工智能产业生态，促进 AI 在经济社会重点领域推广应用，打造国际领先的技术体系

3.2.2 回应治理阶段（2017—2020 年）

1. 2017 年到 2020 年我国人工智能治理采取的是回应治理模式

这个阶段的主导政策是推动人工智能技术发展，出台引导性、规范性的政策措施，形成有利于创新的环境。2017 年政府工作报告明确提出，加快培育壮大新兴产业，全面实施战略性新兴产业发展规划，加快人工智能等技术研发和转化，做大做强产业集群。这是



“人工智能”首次被写入政府工作报告。2017年7月，国务院印发《新一代人工智能发展规划》（下称《规划》），作为我国人工智能发展的纲领性文件，《规划》提出面向2030年我国新一代人工智能发展的指导思想、战略目标、重点任务和保障措施，部署构筑我国人工智能发展的先发优势。《规划》对中国人工智能发展进行了战略性部署，描绘了我国新一代人工智能发展的蓝图，提出人工智能产业2020年、2025年及2030年的“三步走”的目标，明确了我国人工智能产业发展的路线，厘清了我国人工智能发展的基本原则，对我国人工智能治理具有里程碑意义。

2. 2019年《新一代人工智能治理原则——发展负责任的人工智能》发布

提出和谐友好、公平公正、敏捷治理等在内的八条原则。在此之后，我国陆续发布了一系列产业政策文件，包括《促进新一代人工智能产业发展三年行动计划（2018—2020）》《国家新一代人工智能创新发展试验区建设工作指引》《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》（下称《规划纲要》）等，旨在引导和促进中国人工智能产业的发展。规划纲要在以下三个方面布局了人工智能发展。

（1）突破核心技术

人工智能相关技术逐步成为“事关国家安全和全局的基础核心领域”。为进一步推动解决我国人工智能核心技术中的不足和短板，《规划纲要》提出“十四五”期间将通过一批具有前瞻性、战略性的国家重大科技项目，带动产业界逐步突破前沿基础理论和算法，研发专用芯片，构建深度学习框架等开源算法平台，并在学习推理决策、图像图形、语音视频、自然语言识别处理等领域创新与迭代应用。

（2）打造数字经济新优势

发展人工智能应以产业的融合应用与产业数字化转型为核心目标，进而逐渐形成数据驱动、人机协同、跨界融合、共创分享的智能经济形态。《规划纲要》提出要以数字化转型整体驱动生产方式、生活方式和治理方式变革，充分发挥我国数据、应用场景的优势，实施“上云用数赋智”行动，促进数字技术与实体经济深度融合。通过建设重点行业人工智能数据集，发展算法推理训练场景，推进智能医疗装备、智能运载工具、智能识别系统等智能产品制造，推动通用化和行业性人工智能开发平台建设，在智能交通、智慧能源、智能制造、智慧农业及水利、智慧教育、智慧医疗、智慧文旅、智慧社区、智慧家居、智慧政务等领域形成一系列数字化、智能化应用场景。

（3）完善监管框架

针对当前学术界和产业界关心的伦理与法律风险、人工智能技术滥用、算法杀熟等人工智能健康发展的问题，《规划纲要》提出要构建与数字经济发展相适应的政策法规体系，例如在无人驾驶领域建设完善相关监管框架、法律法规和伦理审查规则。另一方面，规划纲要中一系列优化产业政策环境的措施，也将给人工智能提供肥沃的发展土壤，例如“在数字经济等领域制定实施一批国家级重点专项规划”为人工智能发展提供战略导向，“支持民营企业开展基础研究和科技创新、参与关键核心技术研发和国家重大科技项目攻关”将更加激发民营人工智能科技企业的创新活力。



3.2.3 集中治理阶段（2020 年至今）

从 2020 年开始，我国人工智能治理进入了集中治理阶段。2020 年，国家科技伦理委员会正式成立，并于 2021 年发布《新一代人工智能伦理规范》，分别阐述了人工智能的管理规范、研发规范、供应规范和使用规范。将伦理道德融入人工智能全生命周期，为从事人工智能相关活动的各类主体提供伦理指引。

2021 年，“十四五”规划和 2035 年远景目标将人工智能技术及其治理纳入国家规划，一方面要求加快人工智能核心技术突破及产业化发展，另一方面高度重视人工智能治理体系建设，集中呈现了“基于人工智能的治理”和“面向人工智能的治理”两条主线的交织融合。

2023 年，我国在第三届“一带一路”国际合作高峰论坛期间提出《全球人工智能治理倡议》（下称《倡议》），围绕人工智能发展、安全、治理三方面系统阐述了人工智能治理的中国方案，为全球人工智能治理提供了建设性解决思路。《倡议》包括推动建立风险等级测试评估体系、逐步建立健全法律和规章制度、坚持公平性和非歧视性原则、坚持伦理先行等内容，不断提升人工智能技术的安全性、可靠性、可控性、公平性。《倡议》就各方普遍关切的人工智能发展与治理问题提出了建设性解决思路，为相关国际讨论和规则制定提供了蓝本。

3.3 我国人工智能法律法规

3.3.1 整体情况

我国目前尚未出台专门针对人工智能的法律法规，法条散见于《民法典》《数据安全法》《网络安全法》《个人信息保护法》《科学技术进步法》，主要涉及科学技术发展的原则、信息安全和数据安全的保护。如《网络安全法》第二十二条规定了网络运营者应当采取技术措施和其他必要措施，防止和减少网络安全事件的发生，其中包括使用人工智能等技术手段。《个人信息保护法》包括了一些关于人工智能的规定，如禁止将个人信息用于违法活动和侵害个人权益，要求人工智能决策的透明和可解释性等。《国务院 2023 年度立法工作计划》明确将《人工智能法草案》列为“预备提请全国人大常委会审议”的法律案，可见顶层通用立法《人工智能法》已经正式进入我国的立法进程。

3.3.2 算法治理

早在 2021 年，我国就出台相关文件加强对算法的监管。在算法推荐方面，《网络信息内容生态治理规定》提出针对算法推荐技术健全人工干预和用户自主选择机制。



2022年3月1日起施行的《互联网信息服务算法推荐管理规定》是中国第一部聚焦算法治理的立法,该规定针对算法推荐服务提供者提出全方位的义务和禁止性要求,提出算法安全风险监测、算法安全评估、算法备案管理等一系列监管举措,进一步强化平台企业的算法安全责任。在算法自动化决策方面,《中华人民共和国个人信息保护法》《中华人民共和国电子商务法》《在线旅游经营管理服务暂行规定》等都有相关条款对算法歧视、大数据杀熟等不公平、不正当的算法决策行为进行规制,提出了向个人提供选择权限、开展影响评估、个人可以要求说明和享有拒绝权限等规制方式,更好地平衡个人权益保护和算法商业应用活动。在人工智能深度合成方面,相关立法积极为人工智能深度合成技术划定应用红线,促进技术正向应用,如《网络音视频信息服务管理规定》《网络信息内容生态治理规定》《互联网信息服务算法推荐管理规定》等都禁止将深度合成算法用于虚假新闻信息和其他违法违规活动。

2022年11月通过的《互联网信息服务深度合成管理规定》,总结既往监管经验,从深度合成服务提供者主体责任、深度合成信息内容标识管理等方面,对深度合成技术应用进行全面规范,通过内容标识、检测识别、内容溯源等必要的安全保障措施确保深度合成技术应用安全可靠,更好地促进技术创新发展和积极正向应用。

3.3.3 制度规范

2023年7月,国家互联网信息办公室等七部门联合发布《生成式人工智能服务管理暂行办法》,自2023年8月15日起施行。作为全球首个全面监管生成式人工智能的立法文件,该办法对生成式人工智能服务进行框架性规范。

在总体内容方面,该办法主要针对生成式人工智能的生成内容和算法设计进行约束,包括坚持社会主义核心价值观,不得生成法律、行政法规禁止的内容;提高生成内容准确性和可靠性;算法设计、训练数据选择、模型生成和优化、提供服务等过程中采取有效措施防止歧视;尊重知识产权、商业道德和他人合法权益;使用具有合法来源的数据和基础模型;提高训练数据和数据标注质量等。采用的监管方式包括分类分级、安全评估、算法备案、内容标识、监督检查、信息披露等。在模型开发阶段,主要针对数据采集、优化训练、数据标注进行了约束。

该办法对生成式人工智能服务提供者(以下简称“提供者”)提出了具体行为规范。一是明确提供者的义务。如提供者应当依法开展预训练、优化训练等训练数据处理活动,使用具有合法来源的数据和基础模型;不得侵害他人依法享有的知识产权;涉及个人信息的,应当取得个人同意或符合法定情形采取有效措施提高训练数据质量;制定标注规则、开展数据标注质量评估,并进行标注人员培训。二是明确提供者在提供服务事前、事中、事后的全流程安全责任。如在提供服务产品上线前,提供者需通过签订服务协议、开展安全评估和算法备案等方式推动责任落实。提供服务过程中,一方面要与生成式人工智能服务使用者(以下称“使用者”)签订服务协议,明确双方权利、义务;提供具有舆论属性



或者社会动员能力的生成式人工智能服务的市场主体，还需按照国家规定开展安全评估和算法备案；另一方面要通过履行个人信息保护义务、提升透明度、防沉迷、内容标识、建立安全保障措施、加强违法内容的处置和整改、投诉举报、监督检查配合等方式推动责任落实。涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形；采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性。此外，该办法还明确了数据标注的相关要求。

3.3.4 地方政策

上海、深圳等地积极推进人工智能与经济、生活、城市治理等领域深度融合。

(1)《上海市促进人工智能产业发展条例》是全国首部促进人工智能产业发展的省级地方性法规。该条例旨在促进人工智能产业的高质量发展，为上海市的人工智能产业发展提供了全面的指导和法律框架，并确保其安全、可持续性。在安全方面，该条例提出了设立人工智能伦理专家委员会，制定人工智能领域伦理规范指南，推动人工智能企业建立伦理安全治理制度等措施，以保障人工智能产业的健康、安全发展。

(2)《深圳经济特区人工智能产业促进条例》于2022年11月正式实施，为深圳市人工智能产业的发展提供了全面的法治保障和政策支持，同时也为国家层面相关立法提供了参考和经验。该条例要求设立专门的人工智能治理机构，包括人工智能伦理委员会，以加快推进人工智能伦理安全规范的制订和实施，推动构建全面、导向明确、规范有序、协调一致的人工智能伦理治理规则。

3.4 我国人工智能伦理治理

人工智能的广泛应用带来巨大便利的同时，也引发了一系列伦理挑战，如算法偏见、决策黑箱、道德伦理冲突等。科技伦理制度的调整完善成为人工智能时代的重要议题，需要从关注人的伦理转向关注技术的伦理，一方面，有助于识别和消除技术偏见，确保人工智能决策的公平性和无歧视性，另一方面，确保人工智能技术的发展和应用符合社会主义核心价值观，促进技术进步与社会伦理道德的和谐统一。随着全球科技竞争日益激化，科技伦理不仅事关科技安全风险挑战防范，也关乎国家科技竞争力。

国家对科技伦理的重视程度不断提升，明确提出科技伦理是科技活动必须遵守的价值准则，并将科技伦理纳入顶层政策设计，作为科技创新的重要支撑和保障。《中共中央关于坚持和完善中国特色社会主义制度推进国家治理体系和治理能力现代化若干重大问题的决定》和《十四五规划和2035年远景目标纲要》对科技伦理做出了顶层部署，要求健全科技伦理治理体制。目前，国家持续建立健全科技伦理制度，塑造科技向善的文化理念和保障机制，主要体现在以下三个方面。



(1) 组建国家科技伦理委员会，负责指导和统筹协调推进全国科技伦理治理体系建设工作。2019年7月，中央全面深化改革委员会第九次会议审议通过了《国家科技伦理委员会组建方案》，正式确立了中国国家级的科技伦理管理机构。2022年3月发布的《关于加强科技伦理治理的意见》进一步明确了国家科技伦理委员会的管理职责。

(2) 出台科技伦理相关的政策法规，支持科技伦理治理落地实施。2021年12月修订的《科学技术进步法》的一大亮点就是增加了科技伦理相关条款，一方面对健全科技伦理治理体制做出了整体规定，另一方面要求科学技术研究开发机构、高等学校、企业事业单位等主体履行科技伦理管理主体责任，对科学技术活动开展伦理审查，同时明确科学技术研究开发和应用活动的法律伦理底线。《关于加强科技伦理治理的意见》则对科技伦理治理提出了更为全面的要求，涉及总体要求、原则、体制、制度保障、审查和监管以及教育和宣传六大方面，为科技伦理治理的落地实施奠定了基础。后续还将制定科技伦理审查办法、科技伦理高风险科技活动清单、细分领域的科技伦理规范等配套规定。

(3) 强调人工智能领域的科技伦理治理。国家新一代人工智能治理专业委员会先后发布《新一代人工智能治理原则——发展负责任的人工智能》和《新一代人工智能伦理规范》，为发展负责任的人工智能提供伦理指南。《关于加强科技伦理治理的意见》将人工智能、生命科学、医学作为“十四五”期间的三个重点领域，要求制定专门的科技伦理规范、指南等，加强科技伦理立法，并要求相关行业主管部门推出科技伦理监管举措。随后发布的《新一代人工智能伦理规范》明确了人工智能的基本伦理规范，并提出了人工智能应用管理规范、研发规范、供应规范和使用规范等一系列规范。

企业应重点关注研发规范与供应规范，如在算法设计、实现与应用等环节，提升透明性、可解释性、可理解性、可靠性和可控性，增强人工智能系统的韧性、自适应性和抗干扰能力。在数据采集和算法开发环节，提升数据质量，考虑差异化诉求，避免可能存在的数据与算法偏见。在提供人工智能产品和服务环节，应充分尊重和帮助弱势群体与特殊群体，根据需要提供相应替代方案。同时，要保障人类拥有充分自主决策权，确保人工智能始终处于人类控制之下。

2023年4月，科技部官网发布“关于公开征求对《科技伦理审查办法（试行）》意见的公告”，旨在加强科技伦理审查和监管，促进负责任创新。该办法提到，开展可能带来伦理风险挑战的科技活动以及依据相关规定需要进行科技伦理审查的科技活动等应依照本办法进行科技伦理审查。可以预见，未来人工智能领域的科技伦理监管将得到加强。此外，地方层面的立法如《深圳人工智能产业促进条例》和《上海促进人工智能产业发展条例》都把人工智能伦理治理作为支持人工智能产业健康发展的重要保障。

从以上分析可以看出，目前的相关政策法规尝试对科技伦理管理主体责任进行细化，要求创新主体建立伦理委员会、坚持科技伦理底线、进行科技伦理审查、开展科技伦理风险监测预警和评估、开展科技伦理培训等。总体来看，我国在人工智能科技伦理治理方面的趋势是多方面的，涵盖了从顶层设计到具体实践的多个层面，旨在推动人工智能的健康发展，同时防范和治理其潜在伦理风险。



3.5 我国人工智能行政监管

在新一轮科技革命和产业变革深入发展的背景下，我国政府高度重视新一代人工智能的发展规划以及科学监管工作。为了有效规制生成式人工智能技术研发与应用，妥善协调人工智能技术与数据安全之间的关系，国家层面制定并发布了一系列核心政策法规，在《互联网信息服务算法推荐管理规定》《关于加强互联网信息服务算法综合治理的指导意见》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》等规范中逐步建立了人工智能的行政监管体系。该体系主要包括五个方面，分别是算法安全风险检测、算法安全评估、科技伦理审查、算法备案管理和涉算法违法违规行为处置。其中，算法备案管理是算法安全监管的抓手和基石；算法监督检查和算法安全制度风险监测相辅相成、互为补充；算法安全评估是监管体系的落脚点。算法备案制度与安全评估制度，是对相关产品和服务的核心监管要求。

3.5.1 算法备案制度

算法备案制度是人工智能行政监管的主要抓手。2021 年底，国家网信办等四部门联合发布《互联网信息服务算法推荐管理规定》（下称《管理规定》）对算法进行专项治理，首次提出企业“算法备案”要求。

《管理规定》第二十四条规定，具有舆论属性或者社会动员能力的算法推荐服务提供者应当在提供服务之日起十个工作日内通过互联网信息服务算法备案系统填报服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息，履行备案手续，此即算法备案制度。监管部门通过对备案算法的风险评估，可以保证风险的源头控制；通过记录算法设计和检验，也能使监管机构有效地评估、追溯和验证复杂算法。这既有助于监管机关随时系统摸排算法基本情况，为健全算法分级分类体系提供基础性支撑，也有助于为未来公众算法知情权提供制度性保障，同时还能倒逼算法推荐服务提供者积极进行风险评估和全流程记录，督促运营者在整个算法活动阶段考虑算法合规问题。

从监管实践看，算法备案制度主要针对生成合成、个性化推荐、排序精选、检查过滤及调节决策等五类算法实施无差别备案，提出了建立内部管理制度、评估验证、公示说明、用户自主管理、算法备案等要求，覆盖了人工智能产业链上所有企业的常规合规项。2023 年 6 月，网信办公开发布我国首批深度合成服务算法备案清单，共有 41 个算法进入该名单，其中不少算法出自阿里、腾讯、美团、百度和科大讯飞等知名企业。目前，国家网信办已公布三批完成备案的算法应用。

整体而言，互联网领域的算法监管强调安全可控、权益保护、公平公正、公开透明、滥用防范等多元目的，在算法应用分级分类基础上，重点规制高风险类算法应用，监管思



路与《欧盟人工智能法案》较为接近。

3.5.2 安全评估制度

早在2018年，国家网信办和公安部联合发布了《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》，其第3条规定了互联网信息服务提供者自行开展安全评估的五种情形，实践中主要由公安部进行评估。

《生成式人工智能服务管理暂行办法》第十七条再次重申对生成式人工智能的安全评估制度，要求具有舆论属性或者社会动员能力的服务提供者，按照国家有关规定开展安全评估。目前，尚未出台针对人工智能产品和服务安全评估的进一步细则和指导，但根据行业实践来看，安全评估的内容包括但不限于企业机构设置、人员配置、制度建设要求，数据安全、运营维护等多项基础安全评估内容，以及针对账号管理、各类信息审核、各类功能的专项安全评估工作。监管实践中，主要由国家网信办、工信部、公安部等部门主要负责安全评估工作。

综上所述，算法备案制度和安全评估制度的适用范围较为广泛，对于直接或间接向公众提供的产品和服务均可能适用，对于应用层的服务提供和基础层的技术支持也可能适用，在未来可能成为金融、电信、医疗等强监管领域对相关产品和服务提供商的准入要件。

3.6 我国人工智能标准规范

《国家标准化发展纲要》指出，标准化在推进国家治理体系和治理能力现代化中发挥着基础性、引领性作用。对于人工智能领域而言，标准不仅是支持、促进人工智能发展进步和广泛应用的重要手段，也是推进落实人工智能治理的有效方式，世界主要经济体均将技术标准工作列为人工智能发展战略的重要环节。

1. 我国高度重视人工智能领域的标准化建设

2018年颁布的《人工智能标准化白皮书（2018年版）》提出了六项近期亟需研制的基础和关键标准。近年来，各界持续推进人工智能治理领域的国家标准、行业标准、地方标准、团体标准等标准化工作。例如，国家标准化管理委员会成立国家人工智能标准化总体组，全国信息技术标准化技术委员会设立人工智能分技术委员会，都在推进制定人工智能领域的国家标准。2020年7月印发的《国家新一代人工智能标准体系建设指南》明确了人工智能标准领域的顶层设计，该指南将安全伦理标准作为人工智能标准体系的核心组成部分，旨在通过安全伦理标准为人工智能建立合规体系，促进人工智能健康、可持续发展。未来，人工智能治理相关标准需加快制定并落地实施，从而为人工智能领域的技术创新和产业发展提供更多保障。

《国家新一代人工智能标准体系建设指南》提出包含安全伦理在内的八个部分组成的



人工智能标准体系结构。安全伦理标准包括人工智能领域的安全与隐私保护、伦理等部分。安全与隐私保护又包括基础安全，数据、算法和模型安全，技术和系统安全，安全管理和服务，安全测试评估，产品和应用安全等六个部分。伦理标准主要规范人工智能服务冲击传统道德伦理和法律秩序而产生的要求，重点研究领域为医疗、交通、应急救援等特殊行业。

2. 国家有关标准发布实施

2022 年，国家标准《信息技术 人工智能 术语》和《信息技术 生物特征识别 人脸识别系统技术要求》发布实施，前者界定了人工智能领域中的常用术语及定义，给出了基础类、关键通用技术、关键领域技术、安全伦理四大类术语的定义，适用于人工智能领域概念的理解和信息交流，以及科研、教学和应用，为我国人工智能领域国家标准研制奠定了基础。后者规定了人脸识别系统的系统架构、业务流程、功能要求和性能要求。该标准考虑了个人信息保护、隐私安全等问题，对数据存储的形式进行了明确的规定，强调非必要情况下，系统不宜存储在验证和辨识时采集的图像数据；该标准充分考虑了不同应用场景、用户体验、可靠性等因素，给出了合理的性能评价指标。该标准适用于人脸识别系统的设计和开发，对人脸识别技术在各行业的应用以及行业应用标准的编制具有重要的指导意义。

3. 全国网络安全标准化技术委员会组织开展了一些标准化工作

全国网络安全标准化技术委员会（TC260）已在生物特征识别、汽车电子、智能制造等部分人工智能技术、产品或应用安全方面开展了一些标准化工作。在生物特征识别安全方面，TC260 已发布 GB/T 20979-2007《信息安全技术 虹膜识别系统技术要求》标准，正在研制《基于可信环境的生物特征识别身份鉴别协议》《指纹识别系统技术要求》《网络人脸识别认证系统安全技术要求》等标准；在自动驾驶安全方面，2024 年 1 月 11 日，TC260 立项《信息安全技术 汽车电子系统网络安全指南》标准项目，这是我国在汽车电子领域第一个网络安全国家标准；在智能制造安全方面，TC260 正在研制《工业控制网络监测安全技术要求及测试评价方法》《工业控制网络安全隔离与信息交换系统安全技术要求》《工业控制系统产品信息安全通用评估准则》《工业控制系统风险评估实施指南》等工控安全标准。我国人工智能标准制定现状如表 3-2 所示。

总体来看，我国人工智能安全标准主要集中在应用安全领域，缺乏人工智能自身安全或基础共性的安全标准。虽然 TC260 在生物特征识别、汽车电子安全、智能制造安全、大数据安全等领域已经开展了一些标准研究，但现有这些标准大多是规范相关领域产品或系统的自动化行为，和智能化行为安全还有差距。根据《新一代人工智能发展规划》对人工智能安全提出的监测预警、风险评估、安全问责、研发设计人员安全准则等要求，一方面，应加强人工智能基础安全标准研究，针对人工智能安全的参考架构、安全风险、伦理设计、安全评估等方面开展标准研究，并规范人工智能算法、产品和系统的安全要求和测评方法。另一方面，应进一步深化人工智能应用安全标准工作，针对已有标准完善智能安全要求，并继续开展网络安全领域的人工智能应用安全、智能机器人安全、自动驾驶安全、智能家居安全、智能交通安全、智能物流安全、智慧城市安全等领域的标准研究。

表 3-2 我国人工智能标准制定现状

模 块	A. 基础共性					B. 支撑技术与产品					C. 基础软硬件平台					D. 关键通用技术					E. 关键领域技术							
子 模 块	AA 术 语	AB 参 考 架 构	AC 测 试 评 估	AD 其 他	BA 大 数 据	BB 物 联 网	BC 云 计 算	BD 边 缘 计 算	BE 智 能 传 感 器	BF 数 据 存 储 及 传 输 设 备	BG 其 他	CA 智 能 芯 片	CB 系 统 软 件	CC 开 发 框 架	CD 其 他	DA 机 器 学 习	DB 知 识 图 谱	DC 类 脑 智 能 计 算	DD 量 子 智 能 计 算	DE 模 式 识 别	DF 其 他	EA 自 然 语 言 处 理	EB 智 能 语 言	EC 计 算 机 视 觉	ED 生 物 特 征 识 别	EE 虚 拟 现 实	EF 人 机 交 互	EG 其 他
国 家 标 准	1	0	0	1	18	78	48	0	8	9	0	0	7	0	0	1	0	0	0	0	0	0	6	1	37	3	0	0
模 块	F. 产 品 与 服 务					G. 行 业 应 用												H. 安 全 / 伦 理										
子 模 块	FA 智 能 机 器 人	FB 智 能 运 载 工 具	FC 智 能 终 端	FD 智 能 服 务	FE 其 他	GA 智 能 制 造	GB 智 能 农 业	GC 智 能 交 通	GD 智 能 医 疗	GE 智 能 教 育	GF 智 能 商 务	GG 智 能 能 源	GH 智 能 物 流	GI 智 能 金 融	GJ 智 能 家 居	GK 智 能 政 务	GL 智 能 城 市	GM 公 共 安 全	GN 智 能 环 保	GO 智 能 游 戏	GP 其 他	HA 安 全 与 隐 私 保 护	HB 伦 理	HC 其 他				
国 家 标 准	0	0	19	0	0	5	0	4	0	0	0	0	0	2	8	0	26	12	0	0	130	0	0					



展望未来,在全球人工智能安全治理的规则与实践探索进一步走深向实的趋势下,我国人工智能安全治理将继续坚持“发展”与“治理”双轮驱动的思路,强调风险预警和技术保障为主体,法律法规为辅助的治理策略,继续探索合理的人工智能治理模式。一是通过技术创新和应用提升安全水平,即加强人工智能系统的安全设计和提高系统的自主可控能力,确保技术应用的安全性和可靠性。二是完善法律框架,推进精准敏捷监管,进一步构建和优化与人工智能技术发展相匹配的法律法规体系,明确技术应用的边界与责任,针对不同的产品、服务和应用采取不同的规则。三是健全风险预警和应急响应机制,提升对人工智能安全威胁的识别和应对能力,为人工智能技术的健康发展营造一个更加稳固的安全环境,从而促进其在多个领域中的广泛而安全的应用。



习 题

1. 我国人工智能安全治理体系主要包含哪些方面?
2. 我国人工智能安全治理有哪些特点?
3. 我国人工智能安全治理未来将呈现哪些趋势?
4. 我国人工智能战略规划大致可分为哪些阶段?每个阶段的主要特点是什么?
5. “十四五”规划从哪三个方面布局我国人工智能的发展?
6. 我国规制人工智能的法律法规有哪些?
7. 我国算法治理的规范性文件主要有哪些?
8. 《生成式人工智能服务管理暂行办法》对生成式人工智能服务提供者提出哪些具体要求?
9. 我国哪些城市颁布了人工智能地方性法规?
10. 我国如何建立健全人工智能的科技伦理制度?
11. 我国人工智能行政监管的主要抓手是什么?
12. 《国家新一代人工智能标准体系建设指南》提出的人工智能标准体系由哪八个部分组成?
13. 我国人工智能的标准化工作已经在哪些领域开展?