人工智能安全标准化白皮书

（2019版）

全国信息安全标准化技术委员会
大数据安全标准特别工作组

2019年10月

人工智能安全标准化白皮书

（2019版）

全国信息安全标准化技术委员会
大数据安全标准特别工作组

2019年10月

前 言

习近平总书记在十九届中央政治局第九次集体学习时明确指出，要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。人工智能（Artificial Intelligence，简称AI）经过60多年的演进，已发展成为研究和开发用于模拟、延伸和扩展人类智能的学科。近年来，在算法、算力和数据三大因素的共同驱动下，人工智能进入加速发展的新阶段，成为经济发展的领头雁和社会发展的加速器。目前世界主要国家均把人工智能作为国家发展战略。2017年我国发布《新一代人工智能发展规划》，将发展新一代人工智能上升至国家战略高度。随着人工智能在相关行业和人民社会生活中的深度融合应用，由此带来的国家安全、社会伦理、网络安全、人身安全和隐私保护多个层面的风险和挑战，也引起了社会的广泛关注。

人工智能安全标准化是人工智能产业发展的重要组成部分，在激发健康良性的人工智能应用、推动人工智能产业有序健康发展方面发挥着基础性、规范性、引领性作用。《新一代人工智能发展规划》中明确提出了“要加强人工智能标准框架体系研究，逐步建立并完善人工智能基础共性、互联互通、行业应用、网络安全、隐私保护等技术标准”，切实加强人工智能安全标准化工作，是保障人工智能安全的必由之路。

为推动人工智能技术健康、快速、安全、有序的发展和推广应用，全国信息安全标准化技术委员会（以下简称“全国信安标委”）下设的大数据安全标准特别工作组启动了《人工智能安全标准化白皮书》编制工作，本白皮书主要围绕人工智能本身的安全，详细分析人工智能发展现状，面临的主要安全威胁和风险挑战，梳理总结国内外人工智能安全法规政策和标准化组织标准化工作进展。在此基础上，对人工智能安全标准化需求进行深入辨析，提出人工智能安全标准框架和标准化工作建议。

人工智能安全标准化白皮书（2019版）

编写单位

中国电子技术标准化研究院
清华大学
北京百度网讯科技有限公司
华为技术有限公司
三六零科技集团有限公司
阿里巴巴（中国）有限公司
中国移动通信集团有限公司
中国人民大学
浙江蚂蚁小微金融服务集团股份有限公司
国际商业机器(中国)有限公司
北京天融信网络安全技术有限公司
联想（北京）有限公司
上海依图网络科技有限公司
深信服科技股份有限公司
深圳市腾讯计算机系统有限公司
北京三快在线科技有限公司（美团点评）
奇安信科技集团股份有限公司
陕西省网络与信息安全测评中心
北京猎户星空科技有限公司
中国科学院大学自动化研究所
四川大学
内蒙古自治区大数据发展管理局
维沃移动通信有限公司
北京大学
北京神州绿盟信息安全科技股份有限公司
阿里云计算有限公司
上海观安信息技术股份有限公司
OPPO广东移动通信有限公司
中国平安保险（集团）股份有限公司

人工智能安全标准化白皮书（2019版）

编写人员

杨建军　刘贤刚　王建民　胡　影　张宇光　苏　航　刘　焱
张　屹　李　实　郭　锐　朱红儒　王小璞　程海旭　上官晓丽
张　峰　落红卫　徐飞玉　谢安明　吴月升　王　龑　赵春昊
陈兴蜀　叶晓俊　金　涛　刘伯仲　武　杨　罗治兵　赫　然
全　鑫　苏永梓　吴子建　魏玉峰　郑新华　谢　江　贾　科
刘　行　严敏瑞　刘建鑫　何　源　于　乐　朱　军　李　依
白晓媛　杜杨洲　周　俊　李汝鑫　王海棠　曹晓琦　鲍旭华
张大江　江为强　常　玲　彭骏涛　宁　阳　付荣华　王江胜
王艳辉　赵晓娜　包沉浮　赵新强　公　静　马　杰　孙　卫
刘笑岑　李　祎　雷晓锋　于惊涛　卞松山　张红卫　黄汉川
杨　帆　李青山　夏玉明　韩　方　蔡　伟

目录 CONTENTS

目录 CONTENTS

目录 CONTENTS

一、人工智能概述

人工智能，是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统[1]。人工智能相关技术的研究目的是促使智能机器会听（如语音识别、机器翻译等）、会看（如图像识别、文字识别等）、会说（如语音合成、人机对话等）、会行动（如机器人、自动驾驶汽车等）、会思考（如人机对弈、定理证明等）、会学习（如机器学习、知识表示等）[2]。

1.1 人工智能迎来第三次发展浪潮

早在1950年，阿兰图灵在《计算机器与智能》中阐述了对人工智能的思考，并提出以图灵测试对机器智能进行测量。1956年，美国达特茅斯学院举行的人工智能研讨会上首次提出人工智能的概念：让机器能像人类一样认知、思考并学习，这标志着人工智能的开端。

人工智能在20世纪50年代末和80年代初两次进入发展高峰，但受制于技术、成本等因素先后进入低谷期（如图1-1所示）。近年来，随着大数据、云计算、互联网、物联网等信息技术的发展，泛在感知数据和图形处理器等计算平台推动以深度神经网络为代表的人工智能技术飞速发展[2]，人工智能在算法、算力和数据三大因素的共同驱动下迎来了第三次发展浪潮。

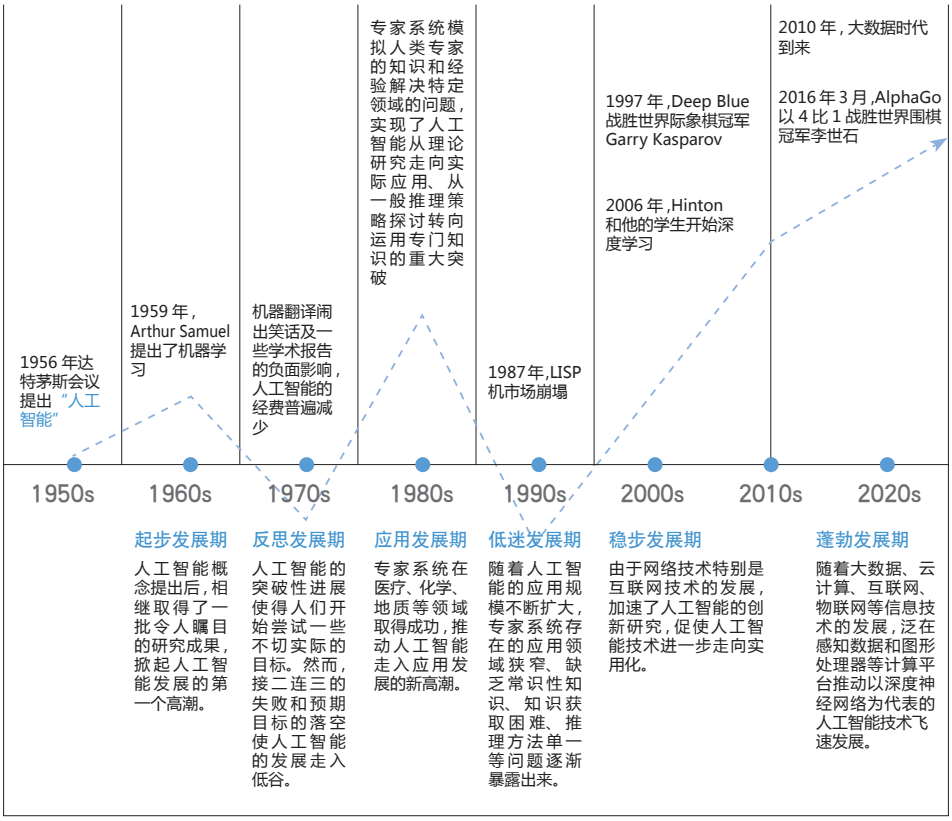


图1–1 人工智能发展历程图

1.2 人工智能技术与应用进展显著

人工智能技术不断发展，尤其是以深度学习为代表的机器学习算法，及以语音识别、自然语言处理、图像识别为代表的感知智能技术取得显著进步。专用人工智能即面向特定领域的人工智能，由于其具备任务单一、需求明确、应用边界清晰、领域知识丰富、建模相对简单等特征，陆续实现了单点突破，在计算机视觉、语音识别、机器翻译、人机博弈等方面可以接近、甚至超越人类水平。

与此同时，机器学习、知识图谱、自然语言处理等多种人工智能关

键技术从实验室走向应用市场（如图1-2所示）。**机器学习**主要研究计算机等功能单元，是通过模拟人类学习方式获取新知识或技能，或通过重组现有知识或技能来改善其性能的过程。深度学习作为机器学习研究中的一个新兴领域，由Hinton等人于2006年提出。**深度学习**又称为深度神经网络（层数超过3层的神经网络），是机器学习中一种基于对数据进行表征学习的方法。在传统机器学习中，手工设计特征对学习效果很重要，但是特征工程非常繁琐，而深度学习基于多层次神经网络，能够从大数据中自动学习特征，具有模型规模复杂、过程训练高效、结果训练准确等特点[3]。**知识图谱，**本质上是结构化的语义知识库，是一种由节点和边组成的图数据结构，以符号形式描述物理世界中的概念及其相互关系。**自然语言处理，**研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。**人机交互，**主要研究人和计算机之间的信息交换，包括人到计算机和计算机到人的两部分信息交换。**计算机视觉，**是使用计算机模仿人类视觉系统的科学，让计算机拥有类似人类提取、处理、理解和分析图像以及图像序列的能力。**生物特征识别，**是指通过个体生理特征或行为特征对个体身份进行识别认证的技术。**智能语音，**主要研究通过计算机等功能单元对人的语音所表示的信息进行感知、分析和合成。

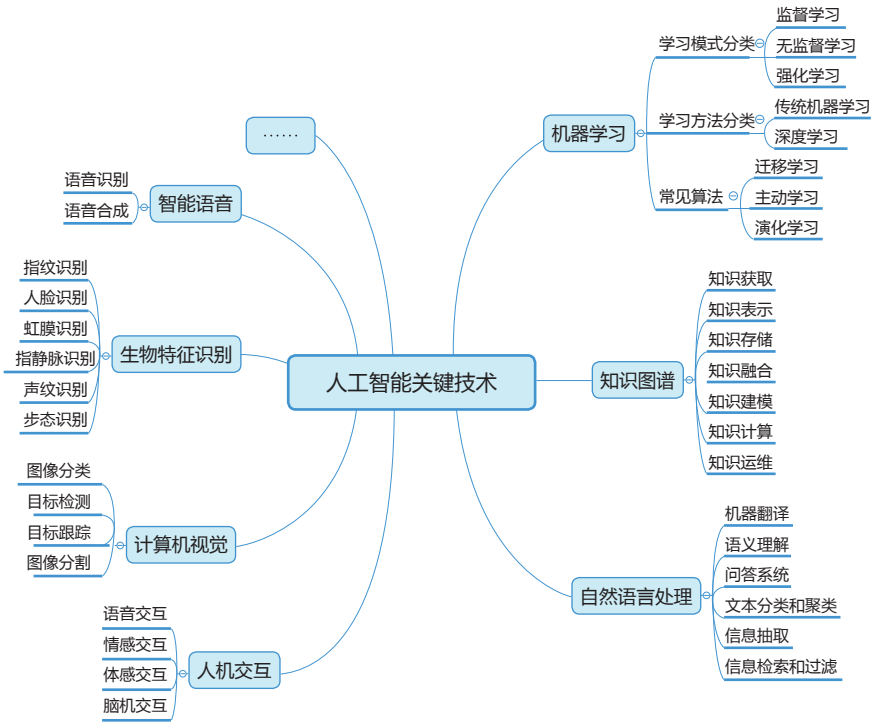


图1–2 人工智能关键技术

1.3 人工智能产业链初具规模

从全球范围来看，人工智能市场发展迅速、潜力巨大。我国作为全球最为重要的科技创新和应用主体，持续发力，促进人工智能产业快速发展。2018年，我国人工智能产业规模约为344亿元，人工智能领域融资额达796.9亿元[5]。截至2019年9月，我国人工智能相关公司总数已超过2500家[5]，主要从事计算机视觉、语音识别和语言技术处理等相关业务[6]。《新一代人工智能发展规划》更是提出到2030年国内人工智能核心产业规模要超过1万亿元，带动相关产业规模超过10万亿元，为人工智能产业持续快速增长指出宏伟目标。

当前，全球人工智能产业链已初具规模，形成多层产业结构（见图

1-3），其中基础层是人工智能的基础支撑，提供计算力、数据等基础资源；技术层是人工智能的技术体系，提供算法开发的软件框架、算法模型、关键技术；应用层实现人工智能应用，提供人工智能产品、服务、行业应用解决方案等。

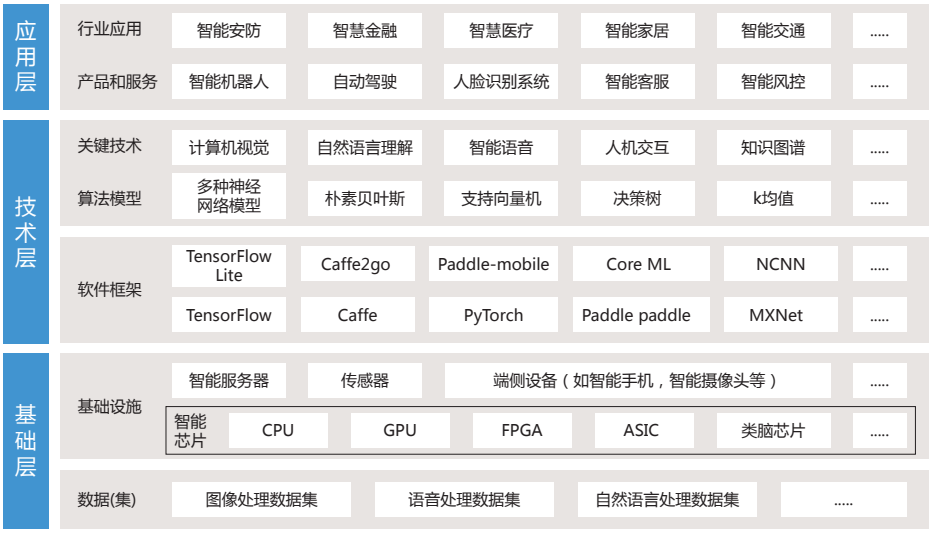


图1–3 人工智能产业链结构

人工智能产业链主要涉及智能基础设施厂商、智能信息及数据提供商、智能技术服务厂商、智能产品和服务提供商、智能应用解决方案提供商等。**智能基础设施厂商，**主要包括智能芯片、智能服务器、端侧设备等为人工智能提供基础计算能力支持的厂商及传感器等硬件厂商。**智能信息及数据提供商，**主要包括数据集提供商，及提供人工智能数据采集、标注、分析、处理相关服务的厂商。**智能技术服务厂商，**依托基础设施和大量数据提供智能技术服务，主要包括提供人工智能的软件框架或技术平台、提供人工智能算法模型和关键技术咨询、提供人工智能在线服务等。**智能产品和服务提供商，**即提供智能机器人、智能运载工具、智能终端、生物特征识别产品、自然语言理解产品、计算机视觉产品等人工智能产品

和服务的厂商[7]，这些产品和服务可以软硬件产品、云服务、API服务等形式存在。**智能应用解决方案提供商，**即人工智能在智能金融、智能制造、智能交通、智能家居、智能医疗、智能金融等垂直行业和应用场景的解决方案提供者。

1.4 我国人工智能应用场景广阔

近年来，我国陆续出台多项政策推动人工智能产业发展，多角度促进人工智能与经济社会深度融合发展。工信部印发了《促进新一代人工智能产业发展三年行动计划（2018-2020年）》，中央深改委会议审议通过了《关于促进人工智能和实体经济深度融合的指导意见》，科技部印发了《国家新一代人工智能创新发展试验区建设工作指引》。截止2019年，全国已建立15个国家级人工智能开放平台，21个省市地区政府出台了人工智能产业相关政策[5]。

在国家和地方政策扶持、数据资源丰富等多因素的驱动下，我国广阔的人工智能应用市场成为发展优势。从垂直行业来看（见图1-4），人工智能在安防、金融、零售、医疗、政务、交通、制造、家居等行业领域得到应用。相关报告显示，当前机器人、安防、医疗成为热门应用场景，智能军事、智能写作、无人船等领域相对处于起步阶段[5]。人工智能在金融领域应用最为深入，在零售行业各环节多点开花，在医疗行业智能应用发展迅速，在政务和安防领域发展前景广阔，在制造业领域有待进一步开发应用潜力[6]。

图1–4 人工智能行业应用示意

1.5 人工智能总体发展水平仍处于起步阶段

目前，人工智能在大数据、算力和深度学习算法的驱动下迎来了新一代发展浪潮，但如考虑到人工智能的研究目的是“探索智能本质”，人类尚无法研制出具有类人智能的智能机器。因此，人工智能总体发展水平仍处于起步阶段，主要表现在三个方面：

一是通用人工智能研究与应用依然任重道远[2]。当前专用人工智能领域虽然已取得突破性进展，但是专用智能主要是弱人工智能，即缺乏自主意识，不能真正实现概念抽象、推理决策和解决问题。通用人工智能，也称为强人工智能或类人智能，通常是指达到人类水平的、能够自适应地应对外界环境挑战的、具有自我意识和思维能力的人工智能。比如，人的大脑就是一个通用的智能系统，能举一反三、融会贯通，可处理视觉、听觉、判断、推理、学习、思考、规划、设计等各类问题。通用智能目前还处于起步阶段，距离人类智能仍有许多差距。

二是人工智能技术距离“很好用”还存在诸多瓶颈。虽然在信息感知和机器学习方面进展显著，但是在概念抽象和规划决策方面刚刚起步。以深度学习为主的人工智能，更适合处理具有丰富的数据或知识、完全信息、确定性信息、静态、单领域和单任务的应用场景[9]，还不能解决通用人工智能问题。可以说，人工智能正处于从“不能实用”到“可以实用”的技术拐点，距离“很好用”还存在可解释性、泛化、能耗、可靠性、鲁棒性等诸多瓶颈[10]。

三是人工智能产业投资愈加理性，复杂场景应用还需要时间。2018年第二季度以来全球人工智能领域投资热度逐渐下降，2019年第一季度全球人工智能融资规模126亿美元，环比下降7.3%。其中我国人工智能领域融资金额30亿美元，同比下降55.8%[11]。在解决实际复杂问题方面，经过大量量身定制的系统能够在围棋、象棋等特定领域挑战中胜出，但是还不能满足具有不确定性、动态的、多领域的复杂应用场景需求。

从另一个角度来看，产业投资降温和技术应用难题有助于人们冷静思考如何研究和应用人工智能。因此，需要充分尊重人工智能技术发展规律和技术应用的成熟度，加强人工智能基础科学和关键技术研究，挖掘人工智能具体应用场景的痛点和难点[18]，构建我国人工智能上下游产业链。以此同时，也应重视人工智能技术发展应用可能带来的安全风险和挑战，降低人工智能在复杂环境、极端条件下的潜在威胁，促进人工智能产业健康、良性、有序发展。

二、人工智能安全法规政策和标准化现状

随着人工智能技术和产业的不断发展，很多国家和地区纷纷制定了人工智能相关的法律法规和政策，以推动人工智能健康有序和安全可控发展，并在人工智能伦理道德、人工智能系统安全、机器人、自动驾驶、隐私保护等方向进行了探索与实践。

2.1 人工智能安全法律法规和政策

2.1.1 国际国外情况

（一）联合国：聚焦人身安全和伦理道德，自动驾驶、机器人、人工智能犯罪等领域逐步深入

目前，联合国对人工智能安全的研究主要聚焦于人身安全、伦理道德、潜在威胁和挑战等方面，关注人工智能对人身安全、社会安全和经济发展的影响和挑战，目前已发布了自动驾驶、智能机器人等领域的相关法律法规和研究成果，相关研究正逐步深入。

2016年，联合国欧洲经济委员会通过修正案修改了《维也纳道路交通公约》（以下简称“《公约》”）。2017年9月，联合国教科文组织与世界科学知识与技术伦理委员会联合发布了《机器人伦理报告》，指出机器人的制造和使用促进了人工智能的进步，并讨论了这些进步所带来的社会与伦理道德问题。

2017年，在荷兰和海牙市政府的支持下，联合国在荷兰建立人工智能和机器人中心，以跟进人工智能和机器人技术的最新发展。该办事处也将联合联合国区域间犯罪和司法研究所（UNICRI）共同处理与犯罪相联系

的人工智能和机器人带来的安全影响和风险。

（二）美国：关注人工智能设计安全，采用标准规范和验证评估减少恶意攻击风险

2019年2月，美国总统签署行政令，启动“美国人工智能倡议”。该倡议提出应在人工智能研发、数据资源共享、标准规范制定、人力资源培养和国际合作五个领域重点发力。其中，标准规范制定的目标是确保技术标准最大限度减少恶意攻击可利用的漏洞，促进公众对人工智能创新技术的信任。

为响应美国人工智能倡议，2019年6月美国对《国家人工智能研究与发展战略计划》进行了更新，在2016年版本的基础上提出长期投资人工智能研究、应对伦理和法律社会影响、确保人工智能系统安全、开发共享的公共数据集和环境、通过标准评估技术等八个战略重点。

人工智能道德、法律和社会问题方面。提出通过设计提高公平性、透明度和问责制，建立道德的人工智能，为道德人工智能设计架构。设计包含道德推理的人工智能架构，如采用操作人工智能和评估监视器分开的双层监视器体系结构，或选择安全工程确保开发的人工智能行为是安全且对人类无害，或使用集合理论原则与逻辑约束相结合来制定道德体系结构。美国国防高级研究计划局（Defense Advanced Research Projects Agency，DARPA）正在开展“可解释人工智能（XAI）”计划，旨在创建一套机器学习技术，在同时保持高水平的学习性能的同时，生成更多可解释的人工智能系统。

创建健壮且可信赖的人工智能系统方面。采取提高可解释性和透明性、建立信任、加强验证、防范攻击、实现长期人工智能安全和价值调整等措施。例如加强对人工智能系统的验证评估，确保系统符合正式规范且满足用户的操作要求。研究人工智能系统的自我监控架构，用于检查系统与设计人员的原始目标行为的一致性。2019年2月，DARPA宣布开展另一

项计划“确保人工智能抗欺骗可靠性（GARD）”，旨在开发新一代防御技术，抵抗针对机器学习模型的对抗欺骗攻击。

构建人工智能公共数据资源方面。采取开发和提供各种数据集、制定培训和测试资源、开发开源软件库和工具包等措施。同时会考虑数据安全共享问题，研究数据安全共享技术和隐私保护技术。例如VA Data Commons正在创建世界最大的链接医学基因组数据集。

标准和基准测试方面。提出制定广泛的人工智能标准、建立技术基准、增加人工智能测试平台的可用性、让社区参与标准和基准测试。其中在人工智能标准方面，提出要针对软件工程、性能、度量、人身安全、可用性、互操作性、安全性、隐私、可追溯性建立人工智能标准体系。

（三）欧盟：重视人工智能伦理道德，GDPR对人工智能带来挑战

2017年，欧洲议会曾通过一项立法决议，提出要制定“机器人宪章”，推动人工智能和机器人民事立法。2018年4月，欧盟委员会发布《欧盟人工智能战略》，通过提高技术和产业能力、应对社会经济变革、建立适当的伦理和法律框架三大支柱，来确立欧盟人工智能价值观。

2018年5月，欧盟GDPR正式生效，其中涉及人工智能的主要有：GDPR要求人工智能的算法具有一定的可解释性，这对于“黑箱”人工智能系统来说可能具有挑战性。同时，GDPR第22条对包括画像在内的自动化决策提出了要求：如果自动化决策产生的法律效力涉及数据主体，或对数据主体有类似的重要影响，则数据主体应有权不成为此决策的对象；如果自动化决策是为了履行数据主体和控制者的合约必须做出的决策，且经过数据主体明示同意，数据控制者应当实施适当措施以保护数据主体的权利、自由和合法权益，并至少保证数据主体具有对自动化决策进行人为干预、个人表达自己观点并拒绝该决策的权利。

2019年4月8日，欧盟委员会发布了由人工智能高级专家组编制的《人工智能道德准则》，列出了人工智能可信赖的七大原则，以确保人工智能

应用符合道德，技术足够稳健可靠，从而发挥其最大的优势并将风险降到最低。其中，可信赖人工智能有两个组成部分：一是应尊重基本人权、规章制度、核心原则及价值观；二是应在技术上安全可靠，避免因技术不足而造成无意的伤害。

（四）德国：积极应对人工智能伦理道德风险，提出关于自动驾驶的首项道德伦理标准

2017年3月，德国24家企业组建“德国人工智能协会”，为行业利益发声，其中包括妥善应对伦理风险等负面影响。德国数据伦理委员会负责为人工智能发展制定道德规范和行为守则。所有基于算法和人工智能的产品、服务需通过审查，尤其是避免出现歧视、诈骗等不法现象。

德国将自动驾驶伦理道德作为规范人工智能发展的核心领域之一。2017年5月12日，德国通过首部针对自动驾驶汽车的法案，对《道路交通法》进行了修订，首次将自动驾驶汽车测试的相关法律纳入其中。法案的目的是保障驾驶者的人身安全，这是德国向无人驾驶迈出的重要一步。2018年5月，德国政府推出关于自动驾驶技术的首项道德伦理标准，该准则将会让自动驾驶车辆针对事故场景作出优先级的判断，并加入到系统的自我学习中，例如人类的安全始终优先于动物以及其他财产等。

2018年7月，德国联邦政府内阁通过了《联邦政府人工智能战略要点》文件，旨在推动德国人工智能研发和应用达到全球领先水平，以负责任的方式促进人工智能的使用，造福社会，并释放新的增值潜力。该文件确立了德国发展人工智能的目标以及在研究、转化、人才培养、数据使用、法律保障、标准、国际合作等优先行动领域的措施，例如采取政府和科研数据开放、国家企业间数据合作、欧洲数据区、扩大医疗卫生行业数据系统互操作性等措施使数据可用能用，保障人工智能系统的透明度、可追溯性和可验证性等。

（五）英国：关注机器人及自治系统的监管，建立数据伦理与创

新中心为政府提供咨询

2016年10月，英国下议院的科学和技术委员会发布了一份关于人工智能和机器人技术的报告，对“机器人技术及自治化系统”（简称RAS）的监管进行了研究。

2018年4月，英国政府发布《人工智能行业新政》，旨在推动英国成为全球人工智能领导者。该文件包括国内外科技公司投资计划、扩建阿兰图灵研究所、创立图灵奖学金以及启动数据伦理与创新中心等内容。其中，数据伦理和创新中心，是一个由英国政府设立的独立咨询机构，可为政府机构和行业提供建议，以支持负责任的技术创新并帮助建立强大、可信赖的治理体系。2019年该中心的主要工作是分析数据驱动技术带来的机遇和风险，包括算法偏见策略审查、人工智能晴雨表，及针对人工智能和保险、智能扬声器和深度造假等主题进行研究等。

2018年4月，英国议会下属的人工智能特别委员会发布报告《人工智能在英国：准备、志向与能力？》，报告认为当前不需要对人工智能进行统一的专门监管，各个行业的监管机构可以根据实际情况对监管做出适应性调整。报告呼吁英国政府制定国家层面的人工智能准则，为人工智能研发和利用设定基本的伦理原则，并探索相关标准和最佳实践等，以便实现行业自律。报告在一些重点问题的建议为：

在最大化公共数据的价值方面，报告提出要区分数据和个人数据，建议通过数据信托、开放公共数据、开放银行数据机制等措施促进数据访问和共享。

在实现可理解、可信赖的人工智能方面，建议避免在特定重大领域采用“黑盒”算法，鼓励研制可解释性的人工智能系统，在安全攸关的特定场景中要求使用更加技术透明的人工智能系统。

在应对算法歧视方面，建议研究训练数据和算法的审查和测试机制，需要采取更多措施确保数据真正具有代表性，能够代表多元化的人群，并

且不会进一步加剧或固化社会不公平。

此外，在自动驾驶方面，英国在2017年2月出台《汽车技术和航空法案》，规定在自动驾驶汽车道路测试发生事故时，可通过简化保险流程，帮助保险人和保险公司获得赔偿。英国也将在2021年全面允许自动驾驶汽车合法上路。

（五）日本：成立人工智能伦理委员会，积极开展人工智能伦理道德研究

2014年12月，日本人工智能学会成立了“伦理委员会”，探讨机器人、人工智能与社会伦理观的联系。2016年6月，伦理委员会提出人工智能研究人员应该遵守的指针草案。该草案强调“存在无关故意与否，人工智能成为有害之物的可能性”，草案规定无论直接还是间接，均不得基于加害意图使用人工智能。在无意施加了危害时，需要修复损失，在发现恶意使用人工智能时采取防止措施。研究者须尽全力促使人们能够平等利用人工智能，并负有向社会解释人工智能局限性和问题点的责任。

2015年1月，日本经济产业省将委员会讨论的成果进行汇总编制了《日本机器人战略：愿景、战略、行动计划》（又称为《新机器人战略》），将机器人的发展与推进作为未来经济发展的重要增长点，制定了详细的“五年行动计划”，将围绕制造业、服务业、农林水产业、医疗护理业、基础设施建设及防灾等主要应用领域，展开机器人技术开发、标准化、示范考核、人才培养和法规调整等具体行动。

2017年3月，日本人工智能技术战略委员会发布《人工智能技术战略》报告，阐述了日本政府为人工智能产业化发展所制定的路线图，包括三个阶段：在各领域发展数据驱动人工智能技术应用（2020年完成一二阶段过渡）；在多领域开发人工智能技术的公共事业（2025-2030年完成二三阶段过渡）；连通各领域建立人工智能生态系统。

2.1.2 国内情况

我国已发布了一系列的人工智能相关政策法规，围绕促进产业技术发展出台了相关政策文件，包括《新一代人工智能发展规划》（以下简称《发展规划》）、《促进新一代人工智能产业发展三年行动计划（2018-2020年）》（以下简称《行动计划》）、《“互联网+”人工智能三年行动实施方案》、《关于促进人工智能和实体经济深度融合的指导意见》和《国家新一代人工智能创新发展试验区建设工作指引》等。这些文件中均提出了人工智能安全和伦理等方面的要求，**主要关注人工智能伦理道德、安全监管、评估评价、监测预警等方面，加强人工智能技术在网络安全的深度应用**。

《发展规划》提出要“制定促进人工智能发展的法律法规和伦理规范”，包括：重点围绕自动驾驶、服务机器人等应用基础较好的细分领域，加快研究制定相关安全管理法规；开展与人工智能应用相关的民事与刑事责任确认等法律问题研究，建立追溯和问责制度；开展人工智能行为科学和伦理等问题研究，建立伦理道德多层次判断结构及人机协作的伦理框架；制定人工智能产品研发设计人员的道德规范和行为守则，加强对人工智能潜在危害与收益的评估，构建人工智能复杂场景下突发事件的解决方案等方面内容。

《发展规划》也指出要“建立人工智能安全监管和评估体系”，包括：加强人工智能对国家安全和保密领域影响的研究与评估，完善人、技、物、管配套的安全防护体系，构建人工智能安全监测预警机制；加强人工智能网络安全技术研发，强化人工智能产品和系统网络安全防护；构建动态的人工智能研发应用评估评价机制等内容。

《行动计划》提出要建立“网络安全保障体系”，包括针对智能网联汽车、智能家居等人工智能重点产品或行业应用，开展漏洞挖掘、安全测试、威胁预警、攻击检测、应急处置等安全技术攻关，推动人工智能先进

技术在网络安全领域的深度应用，加快漏洞库、风险库、案例集等共享资源建设等内容。

在国家人工智能发展战略的指引下，国家相关部门在无人机、自动驾驶、金融等细分领域出台了相应的规范性文件：

（一）无人机。国家民航局先后发布了《关于民用无人机管理有关问题的暂行规定》、《民用无人机空中交通管理办法》、《轻小型无人机运行规定(试行)》、《民用无人机驾驶员管理规定》、《民用无人驾驶航空器经营性飞行活动管理办法（暂行）》、《民用无人机驾驶员管理规定》等规范性文件，对民用无人机飞行活动、民用无人机驾驶员等相关安全问题作出了明确规定。

（二）自动驾驶。工业和信息化部、公安部、交通运输部联合制定发布《智能网联汽车道路测试管理规范（试行）》（下称“《管理规范》”）。《管理规范》适用于在中国境内公共道路上进行的智能网联汽车自动驾驶测试。北京市发布《北京市自动驾驶车辆道路测试管理实施细则（试行）》及相关文件，确定33条、共计105公里开放道路用于测试。上海市发布《上海市智能网联汽车道路测试管理办法（试行）》，划定第一阶段5.6公里开放测试道路，并发放第一批测试号牌。重庆、保定、深圳也相继发布相应的道路测试管理细则或征求意见，支持智能网联汽车开展公共道路测试。

（三）金融。中国人民银行、证监会等四部委联合发布《关于规范金融机构资产管理业务的指导意见》，按其要求，针对自身特点披露智能投顾的算法缺陷也是业务主体的分内之责。中国人民银行发布《金融科技发展规划》，规划指出要加快制定完善人工智能、大数据、云计算等在金融业应用的技术与安全规范；研究制定人工智能金融应用监管规则，强化智能化金融工具安全认证，确保把人工智能金融应用规制在安全可控范围内。

2.2 主要标准化组织人工智能安全工作情况

2.2.1 ISO/IEC JTC1

2017年10月ISO/IEC JTC1在俄罗斯召开会议，决定新成立人工智能的分委员会SC42，负责人工智能标准化工作。SC 42已成立5个工作组，包括基础标准（WG1）、大数据（WG2）、可信赖（WG3）、用例与应用（WG4）、人工智能系统计算方法和计算特征工作组（WG5），此外SC42也包含人工智能管理系统标准咨询组（AG1）、智能系统工程咨询组（AHG3）等。

SC42 WG3可信赖组重点关注人工智能可信赖和伦理问题，已开展人工智能可信度、鲁棒性评估、算法偏见、伦理等标准研制工作，主要标准项目包括：

1）ISO/IEC TR 24027《信息技术 人工智能 人工智能系统中的偏差与人工智能辅助决策》，由美国NIST提出，主要研究人工智能系统与人工智能辅助决策系统中的算法偏见。

2）ISO/IEC PDTR 24028《信息技术 人工智能 人工智能可信度概述》，主要研究了人工智能可信赖的内涵，分析了人工智能系统的典型工程问题和典型相关威胁和风险，提出了对应的解决方案。该标准将可信赖度定义为人工智能的可依赖度和可靠程度，从透明度、可验证性、可解释性、可控性等角度提出了建立人工智能系统可信赖度的方法。

3）ISO/IEC TR 24029–1《人工智能 神经网络鲁棒性评估第1部分：概述》，由法国提出，主要在人工智能鲁棒性研究项目基础上，提出交叉验证、形式化验证、后验验证等多种形式评估神经网络的鲁棒性。TR 24029–2《人工智能 神经网络鲁棒性评估第2部分：形式化方法》在今年10月的日本会议上也已申请立项。

4）ISO/IEC 23894《信息技术 人工智能 风险管理》，梳理了人工智

能的风险，给出了人工智能风险管理的流程和方法。

5）TR《**信息技术 人工智能 伦理和社会关注概述**》，主要从伦理和社会关注方面对人工智能进行研究。

除了SC42外，ISO/IEC JTC1/SC27信息安全技术分委会，在其WG5身份管理和隐私保护技术工作组，已立项研究项目《**人工智能对隐私的影响**》，研究人工智能对隐私产生的影响。ISO/IEC JTC1/SC7软件和系统工程分委会，也在研制ISO/IEC/IEEE 29119-11《**软件和系统工程—软件测试—人工智能系统测试**》，旨在对人工智能系统测试进行规范。

2.2.2 ITU–T

2017年和2018年，ITU-T组织了 “AI for Good Global”峰会，此次峰会重点关注确保人工智能技术可信、安全和包容性发展的战略，以及公平获利的权利。ITU-T主要致力于解决智慧医疗、智能汽车、垃圾内容治理、生物特征识别等人工智能应用中的安全问题。在ITU-T中，SG17安全研究组和SG16多媒体研究组开展了人工智能安全相关标准研制。其中，ITU-T SG17已经计划开展人工智能用于安全以及人工智能安全的研究、讨论和相关标准化项目。ITU-T SG1安全标准工作组下设的Q9“远程生物特征识别问题组”和Q10“身份管理架构和机制问题组”负责ITU-T生物特征识别标准化工作。其中，Q9关注生物特征数据的隐私保护、可靠性和安全性等方面的各种挑战。

2.2.3 IEEE

IEEE开展了多项人工智能伦理道德研究，发布了多项人工智能伦理标准和研究报告。早在2017年底，IEEE发布了《以伦理为基准的设计：人工智能及自主系统中将人类福祉摆在优先地位的愿景（第二版）》报告，该报告收集了250多名在全球从事人工智能、法律和伦理、哲学、政策相

关工作的专家对人工智能及自主系统领域的问题见解及建议，目前该报告已更新为第二版。

IEEE正在研制IEEE P7000系列标准，用于规范人工智能系统道德规范问题，主要标准介绍如下：

1）IEEE P7000《**在系统设计中处理伦理问题的模型过程**》：该标准建立了一个过程模型，工程师和技术人员可以在系统启动、分析和设计的各个阶段处理伦理问题。预期的过程要求包括新IT产品开发、计算机伦理和IT系统设计、价值敏感设计以及利益相关者参与道德IT系统设计的管理和工程视图。

2）IEEE P7001《**自治系统的透明度**》：针对自治系统运营的透明性问题，为自治系统开发过程中透明性自评估提供指导，帮助用户了解系统做出某些决定的原因，并提出提高透明度的机制（如需要传感器安全存储、内部状态数据等）。

3）IEEE P7002《**数据隐私处理**》：指出如何对收集个人信息的系统和软件的伦理问题进行管理，将规范系统/软件工程生命周期过程中管理隐私问题的实践，也可用于对隐私实践进行合规性评估（隐私影响评估）。

4）IEEE P7003《**算法偏差注意事项**》：本标准提供了在创建算法时消除负偏差问题的步骤，还将包括基准测试程序和选择验证数据集的规范，适用于自主或智能系统的开发人员避免其代码中的负偏差。当使用主观的或不正确的数据解释（如错误的因果关系）时，可能会产生负偏差。

5）IEEE P7004《**儿童和学生数据治理标准**》：该标准定义了在任何教育或制度环境中如何访问，收集，共享和删除与儿童和学生有关的数据，为处理儿童和学生数据的教育机构或组织提供了透明度和问责制的流程和认证。

6）IEEE P7005《**透明雇主数据治理标准**》：提供以道德方式存储、

保护和使用员工数据的指南和认证，希望为员工在安全可靠的环境中分享他们的信息以及雇主如何与员工进行合作提供清晰和建议。

7）IEEE P7006《个人数据人工智能代理标准》：涉及到关于机器自动作出决定的问题，描述了创建和授权访问个人化人工智能所需的技术要素，包括由个人控制的输入、学习、伦理、规则和价值。允许个人为其数据创建个人“条款和条件”，代理人将为人们提供一种管理和控制其在数字世界中的身份的方式。

8）IEEE P7007《伦理驱动的机器人和自动化系统的本体标准》：建立了一组具有不同抽象级别的本体，包含概念、定义和相互关系，这些定义和关系将使机器人技术和自动化系统能够根据世界范围的道德和道德理论进行开发。

9）IEEE P7008《机器人、智能与自主系统中伦理驱动的助推标准》：机器人、智能或自治系统所展示的“助推”被定义为旨在影响用户行为或情感的公开或隐藏的建议或操纵。该标准确定了典型微动的定义（当前正在使用或可以创建），包含建立和确保道德驱动的机器人、智能和自治系统方法论所必需的概念、功能和利益。

10）IEEE P7009《自主和半自主系统的失效安全设计标准》：自治和半自治系统，在有意或无意的故障后仍可运行会对用户，社会和环境造成不利影响和损害。本标准为在自治和半自治系统中开发、实施和使用有效的故障安全机制，建立了特定方法和工具的实用技术基准，以终止不成功或失败的情况。

11）IEEE P7010《合乎伦理的人工智能与自主系统的福祉度量标准》，本标准建立与直接受智能和自治系统影响的人为因素有关的健康指标，为这些系统处理的主观和客观数据建立基线以实现改善人类福祉的目的。

12）IEEE P7011《新闻信源识别和评级过程标准》：该标准的目的

是通过提供一个易于理解的评级开放系统，以便对在线新闻提供者和多媒体新闻提供者的在线部分进行评级，来应对假新闻未经控制的泛滥带来的负面影响。

13）IEEE P7012《机器可读个人隐私条款标准》：该标准给出了提供个人隐私条款的方式，以及机器如何阅读和同意这些条款。

14）IEEE P7013《人脸自动分析技术的收录与应用标准》：研究表明用于自动面部分析的人工智能容易受到偏见的影响。该标准提供了表型和人口统计定义，技术人员和审核员可以使用这些定义来评估用于训练和基准算法性能的面部数据的多样性，建立准确性报告和数据多样性规则以进行自动面部分析。

2.2.4 NIST

2019年8月，美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）发布了关于政府如何制定人工智能技术标准和相关工具的指导意见，该指南概述了多项有助于美国政府推动负责任地使用人工智能的举措，并列出了一些指导原则，这些原则将为未来的技术标准提供指导。

指南强调，需要开发有助于各机构更好地研究和评估人工智能系统质量的技术工具。这些工具包括标准化的测试机制和强大的绩效指标，可让政府更好地了解各个系统，并确定如何制定有效的标准。NIST建议专注于理解人工智能可信度的研究，并将这些指标纳入未来的标准，也建议在监管或采购中引用的人工智能标准保持灵活性，以适应人工智能技术的快速发展；制定度量标准以评估人工智能系统的可信赖属性；研究告知风险、监控和缓解风险等人工智能风险管理；研究对人工智能的设计、开发和使用的信任需求和方法；通过人工智能挑战问题和测试平台促进创造性的问题解决等。

2.2.5 TC260

人工智能安全标准，是与人工智能安全、伦理、隐私保护等相关的标准规范。从广义来说，人工智能安全标准涉及人工智能的算法模型、数据、基础设施、产品和应用相关的安全标准。目前，全国信息安全标准化技术委员会（简称“信安标委”或TC260）的人工智能安全相关标准**主要集中在生物特征识别、智慧家居等人工智能应用安全领域，及与数据安全、个人信息保护相关的支撑领域，尚未有正式立项的人工智能自身安全或基础共性的安全标准**。

（一）人工智能基础共性标准

2018年TC260立项标准研究项目**《人工智能安全标准研究》**，主要研究人工智能安全风险、人工智能安全政策和标准现状、人工智能安全标准需求和安全标准体系等内容，本白皮书的许多材料来自该课题的成果输出。

2019年立项**《信息安全技术 人工智能应用安全指南》**标准研究项目，将研究人工智能的安全属性和原则、安全风险、安全管理及在需求、设计、开发训练、验证评估、运行等阶段的安全工程实践指南，适用于人工智能开发者、运营管理者、用户以及第三方等组织在保障人工智能系统工程安全时作为参考。

（二）生物特征识别安全标准

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《信息安全技术 基于可信环境的生物特征识别身份鉴别协议》、《信息安全技术 指纹识别系统技术要求》、《信息安全技术 网络人脸识别认证系统安全技术要求》、《信息安全技术 生物特征识别信息的保护要求》等标准。

GB/T 20979–2019《信息安全技术 虹膜识别系统技术要求》：规定了用虹膜识别技术为身份鉴别提供支持的虹膜识别系统的技术要求。本标

准适用于按信息安全等级保护的要求所进行的虹膜识别系统的设计与实现，对虹膜识别系统的测试、管理也可参照使用。

GB/T 36651–2018《信息安全技术 基于可信环境的生物特征识别身份鉴别协议》：规定了基于可信环境的生物特征识别身份鉴别协议，包括协议框架、协议流程、协议要求以及协议接口等内容。本标准适用于生物特征识别身份鉴别服务协议的开发、测试和评估。

GB/T 37076–2018《信息安全技术 指纹识别系统技术要求》：对指纹识别系统的安全威胁、安全目的进行了分析，提出指纹识别系统的安全技术要求，规范指纹识别技术在信息安全领域的应用。

《信息安全技术 网络人脸识别认证系统安全技术要求》：规定了安全防范视频监控人脸识别系统的基本构成、功能要求、性能要求及测试方法。本标准适用于以安全防范为目的的视频监控人脸识别系统的方案设计、项目验收以及相关的产品开发。其他领域的视频监控人脸识别系统可参考使用。

《信息安全技术 生物特征识别信息的保护要求》：研究制定生物特征识别信息的安全保护要求，包括生物特征识别系统的威胁和对策，生物特征信息和身份主体之间安全绑定的安全要求，应用模型以及隐私保护要求等。

（三）自动驾驶安全标准

2017年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，2019年立项标准制定项目《信息安全技术 车载网络设备信息安全技术要求》和研究项目《汽车电子芯片安全技术要求》。严格来说，这几项标准属于汽车电子范畴，还不属于自动驾驶范畴。

（四）智慧家居安全标准

2018年TC260立项标准《信息安全技术 智能家居安全通用技术要求》，2019年立项标准《信息安全技术 智能门锁安全技术要求和测试评

价方法》。

《信息安全技术 智能家居安全通用技术要求》：规定了智能家居通用安全技术要求，包括智能家居整体框架、智能家居安全模型以及智能家居终端安全要求、智能家居网关安全要求、网络安全要求和应用服务平台安全要求，适用于智能家居产品的安全设计和实现，智能家居的安全测试和管理也可参照使用。

《信息安全技术 智能门锁安全技术要求和测试评价方法》：给出了智能门锁的安全技术要求和测试评价方法，其中智能门锁是指通过识别指纹、指静脉、虹膜、人脸等人体生物特征以及智能卡、无线遥控编码、静态密码、临时密码等信息，控制执行机构实施启闭的门锁。

（五）数据安全和个人信息保护标准

TC260的大数据安全标准特别工作组自2016年成立以来，在数据安全和个人信息保护方向已发布6项国家标准，在研标准10项，研究项目18项。

在个人信息保护方向，主要聚焦于个人信息保护要求、去标识技术、App收集个人信息、隐私工程、影响评估、告知同意、云服务等内容，已发布GB/T 35273《信息安全技术 个人信息安全规范》、GB/T 37964《信息安全技术 个人信息去标识化指南》2项标准，在研5项标准，2项标准研究项目。

在数据安全方向，主要围绕数据安全能力、数据交易服务、出境评估、政务数据共享、健康医疗数据安全、电信数据安全等内容，已发布GB/T35274《信息安全技术 大数据服务安全能力要求》、GB/T 37932《信息安全技术 数据交易服务安全要求》、GB/T 37973《信息安全技术 大数据安全管理指南》、GB/T 37988《信息安全技术 数据安全能力成熟度模型》4项标准，在研5项标准，16项标准研究项目。

此外，国家标准化管理委员会于2018年1月正式成立国家人工智能标

准化总体组，承担人工智能标准化工作的统筹协调和规划布局，负责开展人工智能国际国内标准化工作。目前，国家人工智能标准化总体组已发布《人工智能标准化白皮书2018》、《人工智能伦理风险分析报告》等成果，正在研究人工智能术语、人工智能伦理风险评估等标准。

2.2.6 其他标准化组织

中国通信标准化协会已开展汽车电子、智能家居等方面标准研究工作，目前已发布YDB 201-2018《智能家居终端设备安全能力技术要求》、T/CSHIA 001-2018《智能家居网络系统安全技术要求》等标准。但相关研究工作仍以人工智能在具体应用场景的情况为主，在研标准包括《人工智能产品、应用及服务安全评估指南》、《人工智能服务平台安全》、《人工智能终端产品标准体系研究》、《移动智能终端人工智能能力及应用个人信息保护技术要求及评估方法》、《移动智能终端人脸识别安全技术要求及测试评估方法》等。

中国人工智能开源软件发展联盟是从事人工智能开源软件相关工作的社会组织，该联盟已研制机器翻译、智能助理等产品或服务评估标准、深度学习算法的可靠性评估标准，主要包括T/CESA 1039—2019《信息技术 人工智能 机器翻译能力等级评估》、T/CESA 1038—2019《信息技术 人工智能 智能助理能力等级评估》、T/CESA 1026—2018《人工智能 深度学习算法评估规范》等。T/CESA 1026—2018《人工智能 深度学习算法评估规范》提出了人工智能深度学习算法的评估指标体系、评估流程，以及需求阶段评估、设计阶段评估、实现阶段评估和运行阶段评估等内容，能够指导深度学习算法开发方、用户方以及第三方等相关组织对深度学习算法的可靠性开展评估工作。

2.3 人工智能伦理道德工作情况

人工智能的复杂性决定了其涉及技术、伦理、法律、道德等多个领域。为了保障人工智能健康发展，需要建立相应的伦理道德框架。在人工智能伦理道德方面，国内外研究成果比较丰富，其中“阿西洛马人工智能原则”和IEEE组织倡议的人工智能伦理标准成为国际上影响最广的人工智能伦理研究成果。而除了广泛达成的共识之外，多个国家和机构也发布了各自的相关准则[15]。

（一）阿西洛马人工智能原则

“阿西洛马人工智能原则”是2017年1月在阿西洛马召开的“有益的人工智能”（Beneficial AI）会议上提出，其倡导的伦理和价值原则包括：安全性、失败的透明性、审判的透明性、负责、与人类价值观保持一致、保护隐私、尊重自由、分享利益、共同繁荣、人类控制、非颠覆以及禁止人工智能装备竞赛等。

（二）IEEE

2017年3月，IEEE在《IEEE 机器人与自动化》杂志发表了名为“旨在推进人工智能和自治系统的伦理设计的IEEE 全球倡议书”，倡议建立人工智能伦理的设计原则和标准，帮助人们避免对人工智能产生恐惧和盲目崇拜，从而推动人工智能的创新，其提出了以下五个原则：1）人权：确保它们不侵犯国际公认的人权；2）福祉：在它们的设计和使用中优先考虑人类福祉的指标；3）问责：确保它们的设计者和操作者负责任且可问责；4）透明：确保它们以透明的方式运行；5）慎用：将滥用的风险降到最低。

（三）美国

美国公共政策委员会于2017 年1 月12 日发布了《算法透明和可责性声明》提出了以下七项准则：1）充分认识；2）救济；3）可责性；4）可解释；5）数据来源保护；6）可审查性；7）验证和测试。

（四）欧盟

2019年4月8日，欧盟委员会发布了由人工智能高级专家组编制的《人工智能道德准则》，列出了人工智能可信赖的七大原则，包括：1）人的能动性和监督能力；2）安全性；3）隐私数据管理；4）透明度；5）包容性；6）社会福祉；7）问责机制。

（四）日本

日本人工智能学会(JSAI)发布了《日本人工智能学会伦理准则》，要求日本人工智能学会会员应当遵循并实践以下准则：1）贡献人类；2）遵守法律法规；3）尊重隐私；4）公正；5）安全；6）秉直行事；7）可责性与社会责任；8）社会沟通和自我发展；9）人工智能伦理准则。

（五）英国

2018年4月，英国议会下属的人工智能特别委员会发布报告《人工智能在英国：准备、志向与能力？》，提出包含五方面内容的准则：1）人工智能应为人类共同利益和福祉服务；2）人工智能应遵循可理解性和公平性原则；3）人工智能不应用于削弱个人、家庭乃至社区的数据权利或隐私；4）所有公民都有权接受相关教育，以便能在精神、情感和经济上适应人工智能的发展；5）人工智能绝不应被赋予任何伤害、毁灭或欺骗人类的自主能力。

（六）加拿大

在加拿大发布的《可靠的人工智能草案蒙特利尔宣言》提出了七种价值，并指出它们都是人工智能发展过程中应当遵守的道德原则：福祉、自主、正义、隐私、知识、民主和责任。

（七）中国

2019年2月25日，科技部宣布成立国家新一代人工智能治理专业委员会，以进一步加强人工智能相关法律、伦理、标准和社会问题研究，深入参与人工智能相关治理的国际交流合作。2019年6月19日，该委员会发布

《新一代人工智能治理原则——发展负责任的人工智能》，提出人工智能发展应遵循八项原则：1）和谐友好；2）公平公正；3）包容共享；4）尊重隐私；5）安全可控；6）共担责任；7）开放协作；8）敏捷治理。

2019年4月，国家人工智能标准化总体组发布了《人工智能伦理风险分析报告》。报告提出两项人工智能伦理准则[15]，一是人类根本利益原则，指人工智能应以实现人类根本利益为终极目标；二是责任原则，指在人工智能相关的技术开发和应用两方面都建立明确的责任体系。在责任原则下，在人工智能技术开发方面应遵循透明度原则；在人工智能技术应用方面则应当遵循权责一致原则。

此外，机器人作为有代表性的人工智能产品，在机器人原则与伦理标准方面，日本、韩国、英国、欧洲和联合国教科文组织等相继推出了多项伦理原则、规范、指南和标准。

三、人工智能安全风险分析与内涵

人工智能应用正在改变人类经济社会的发展轨迹，给人民的生产生活带来巨大改变。但是，人工智能也给全社会带来不容忽视的风险挑战。人工智能安全风险是指安全威胁利用人工智能资产的脆弱性，引发人工智能安全事件或对相关方造成影响的可能性。要利用好人工智能技术就要全面清晰地了解新的攻击威胁、人工智能安全隐患及对相关方造成的影响。

3.1 新的攻击威胁

人工智能系统作为采用人工智能技术的信息系统，除了会遭受拒绝服务等传统网络攻击威胁外，也会面临针对人工智能系统的一些特定攻击，这些攻击特别影响使用机器学习的系统[13]。

（一）攻击方法

一是对抗样本攻击，是指在输入样本中添加细微的、通常无法识别的干扰，导致模型以高置信度给出一个错误的输出。研究表明深度学习系统容易受到精心设计的对抗样本的影响，可能导致系统出现误判或漏判等错误结果。对抗样本攻击也可来自物理世界，通过精心构造的交通标志对自动驾驶进行攻击。比如，Eykholt等人[12]的研究表明一个经过稍加修改的实体停车标志，能够使得一个实时的目标检测系统将其误识别为限速标志，从而可能造成交通事故。

攻击者利用精心构造的对抗样本，也可发起模仿攻击、逃避攻击等欺骗攻击[14]。**模仿攻击**通过对受害者样本的模仿，达到获取受害者权限的目的，目前主要出现在基于机器学习的图像识别系统和语音识别系统中。**逃避攻击**是早期针对机器学习的攻击形式，比如垃圾邮件检测系统、PDF文

件中的恶意程序检测系统等。通过产生一些可以成功逃避安全系统检测的对抗样本，实现对系统的恶意攻击。

二是数据投毒，主要是在训练数据中加入精心构造的异常数据，破坏原有的训练数据的概率分布，导致模型在某些条件会产生分类或聚类错误。由于数据投毒攻击需要攻击者接触训练数据，通常针对在线学习场景（即模型利用在线数据不断学习更新模型），或者需要定期重新训练进行模型更新的系统，这类攻击比较有效，典型场景如推荐系统、自适应生物识别系统、垃圾邮件检测系统等。正确过滤训练数据可以帮助检测和过滤异常数据，从而最大程度地减少可能的数据投毒攻击。

三是模型窃取，是指向目标模型发送大量预测查询，使用接收到的响应来训练另一个功能相同或类似的模型，或采用逆向攻击技术获取模型的参数及训练数据。针对云模式部署的模型，攻击者通常利用机器学习系统提供的一些应用程序编程接口（API）来获取系统模型的初步信息，进而通过这些初步信息对模型进行逆向分析，从而获取模型内部的训练数据和运行时采集的数据[4]。针对私有部署到用户的移动设备或数据中心的服务器上的模型，攻击者通过逆向等传统安全技术，可以把模型文件直接还原出来使用。

四是人工智能系统攻击。对机器学习系统的典型攻击是影响数据机密性及数据和计算完整性的攻击，还有其他攻击形式导致拒绝服务、信息泄露或无效计算。例如，对机器学习系统的控制流攻击可能会破坏或规避机器学习模型推断或导致无效的训练。机器学习系统使用的复杂设备模型（如硬件加速器）大多是半虚拟化或仿真的，可能遭受设备欺骗，运行时内存重新映射攻击及中间人设备等攻击。

（二）攻击影响

深度学习易遭受数据投毒、逃避攻击、模仿攻击、模型窃取和对抗样本攻击。

一是模型的训练、测试和推断过程中均可能遭受攻击。数据投毒攻击主要针对训练过程进行攻击，对抗样本攻击可针对训练、测试和推断过程进行攻击，模型窃取攻击主要针对推断过程进行攻击，逃避攻击和模仿攻击主要针对测试和推断过程。

二是攻击危害数据和模型的机密性、完整性和可用性。上述攻击通常导致模型决策失误、数据泄漏等后果。数据投毒攻击会破坏训练数据集，主要影响数据完整性和模型可用性。模型窃取攻击主要影响数据机密性、模型机密性和隐私。对抗样本、逃避攻击、模仿攻击不会破坏训练数据集，主要影响模型完整性和可用性。

3.2 人工智能安全隐患

通常，人工智能资产主要由数据、算法模型、基础设施、产品服务及行业应用组成，如图3-1所示。一个训练好的模型可以被理解为一个函数y=f(x)。深度学习训练好的模型通常包括两个部分，一个是对网络结构的描述或者称为对网络结构的定义，另一个是每层网络的具体参数值[4]。

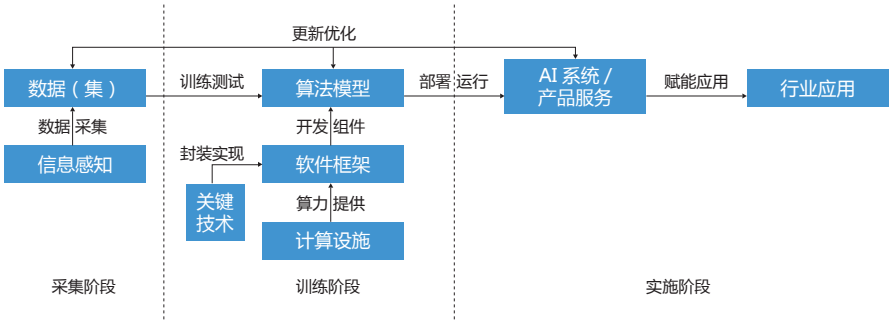


图3–1 人工智能资产示意图

3.2.1 算法模型安全隐患

算法模型，是人工智能系统的核心，而算法模型中的安全隐患则可能

给人工智能系统带来致命的安全后果。

（一）算法模型潜藏鲁棒性平衡、数据依赖等缺陷

一是模型准确性与鲁棒性难以权衡。人工智能算法模型普遍依赖于概率、统计模型构建，在准确性和鲁棒性之间存在权衡和博弈。Eykholt等人[12]的研究表明，针对在对抗样本攻击下的鲁棒性，准确度越高的模型的普遍鲁棒性越差，且分类错误率的对数和模型鲁棒性存在线性关系。

二是数据集对模型准确性影响大。目前人工智能仍处于海量数据驱动知识学习的阶段，数据集的数量和质量是决定模型质量的关键因素之一。模型应用可能出现预料之外的情况，而训练数据可能难以覆盖该类情况，导致与预期不符甚至伤害性的结果。正常的环境变化也可能产生数据集噪声，对模型的可靠性造成威胁，比如仿射变换、光照强度、角度、对比度变化会对视觉模型的预测产生不可预期的影响。

三是面临可靠性挑战。实时性较高的应用场景（如自动驾驶）要求算法模型随时可用，如果当数据进入人工智能核心模块前受到定向干扰，将会导致即时错误判断。

（二）算法可能潜藏偏见或歧视，导致结果偏差或处理不当

偏见歧视，是指由于算法的设计者或开发人员对事物的认知存在主观上的某种偏见，或者不经意使用了带有偏差的训练数据集等原因，造成模型准确性下降或分类错误，甚至在模型使用时产生了带有歧视性的结果。训练数据集的偏差通常是由于不正确的应用或未考虑统计方法和规则，比如当主观而非客观地选择数据或选择了非随机数据，就会产生选择偏差，当假设条件可能被相关信息解释所证实时，则会产生确认偏差。如果偏见已存在于算法之中，经深度学习后这种偏见便可能在算法中得到进一步加强。如果将算法应用在犯罪评估、信用贷款、雇佣评估等关切人身利益的场合，其产生的歧视可能会严重危害个人权益。

（三）人工智能算法决策的“黑箱”特征，存在结果可解释性和透明性问题

深度学习在很多复杂任务上取得了前所未有的进展，但是深度学习系统通常拥有数以百万甚至十亿计的参数，开发人员难以用可解释的方式对一个复杂的神经网络进行标注，成为了一个名副其实的“黑箱”。

一是基于神经网络的人工智能算法具有“涌现性”和“自主性”，容易造成算法黑箱[16]。涌现性，即智能是一种由算法底层的简单规则生成的复杂行为，算法行为不是边界清晰的单个行为而是集体行为的演化，其行为效果既不由“某一”行为所决定，亦不由其前提完全决定。自主性，当前弱人工智能的自主性主要是指智能行为的自组织性，深度学习算法可以在没有程序员的干预下从海量无标注的大数据中自我学习、自我进化。

二是重要行业人工智能应用面临可解释性挑战。人工智能在金融、医疗、交通等攸关人身财产安全的重点行业领域应用时，人类对算法的安全感、信赖感、认同度可能取决于算法的透明性和可理解性。此外，除了人工智能模型本身在技术上的不透明性外，在数据、数据处理活动方面也可能存在不透明性。

3.2.2 数据安全与隐私保护隐患

数据，是人工智能的基础资源，机器学习需要数量大、种类多、质量高的数据进行训练，从不同工程阶段来看涉及采集的海量原始数据、训练和测试数据集、应用系统实际输入数据（现场数据）等。人工智能系统的数据生命周期如图3-2所示。

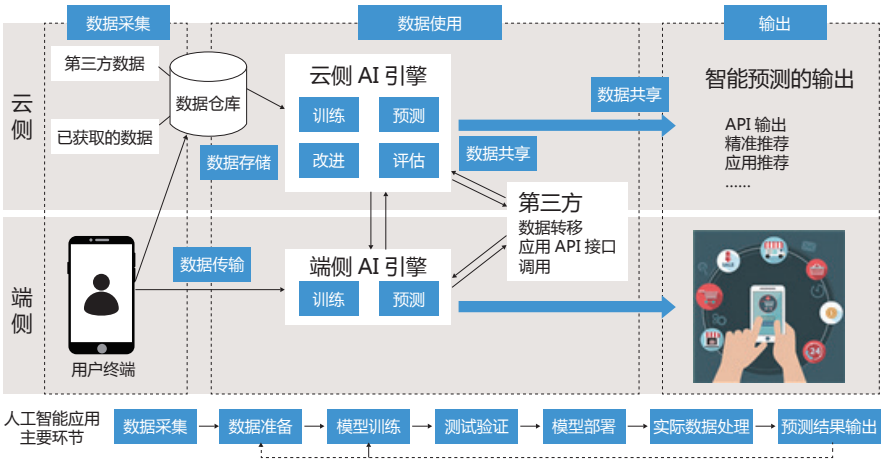


图3–2 人工智能系统的数据生命周期

（一）数据采集安全隐患

在数据采集阶段，人工智能系统通过用户提供、自动采集、间接获取等方式采集大量训练数据和应用数据，用于训练模型或推断预测，常见数据采集阶段主要问题包括：

一是过度采集数据。由于模型训练需要采集大量、多样的数据，如何确保数据采集遵循必要原则、针对使用目的明确数据采集范围是个难点。在人工智能应用场景中，为了优化改进产品、识别特定目标等目的，采集终端可能会采集大量的环境数据或用户行为日志，例如自动驾驶为了避免撞到行人会收集足够的环境信息来判断是否行人。

二是数据采集与用户授权不一致。除了直接从用户采集数据外，通常还存在从网上公开数据源、商务采购等渠道获取训练数据的情况。如何保证这些场景下的用户授权同意面临一定挑战，例如公开数据源通常限定仅被用于科研，如果将其作为商业目的使用会面临数据使用与用户授权范围不一致的风险。

三是个人敏感信息采集合规问题。当前发展较好的计算机视觉、语音

识别等应用，通常需要采集人脸、声纹等敏感的个人生物特征，而生物特征等个人敏感信息采集存在法律合规风险，例如2019年8月瑞典数据监管机构对当地一所高中开出20万瑞典克朗的GDPR罚单，理由是学校采用人脸识别系统记录学生的出勤率。

四是数据质量问题。人工智能模型精度受限于训练数据、应用数据的质量，而训练数据集规模不足、数据集的多样性和均衡性不足、数据集的标注质量低、数据投毒攻击、数据噪声等都将影响训练数据的质量[8]。

五是用户选择退出权难以保障。由于人工智能系统通常由训练好的模型部署而成，用户对于模型训练时所用的数据很难做到选择退出。

（二）数据使用安全隐患

数据使用阶段，即数据分析和处理，涉及模型训练过程和部署运行过程，包括数据准备、模型训练、测试验证、模型部署、实际数据处理、预测结果输出等。其中数据准备主要对采集的原始数据进行预处理和标注等操作，产生用于训练的数据集。

一是匿名化数据被重识别问题。在数据预处理阶段，通常使用数据预处理技术来提升数据质量，这个过程会把外部获取的数据和用户自己的数据进行合并，可能导致已经被匿名化的数据再次被识别。

二是数据标注安全隐患和合规问题。受限于数据标注成本，大多数公司委托数据标注外包公司和自主标注相结合的方式进行数据标注。由于数据标注人员能够直接接触原始数据和数据集，如果数据安全管理不规范，可能存在内部人员盗取数据、数据未授权访问、训练数据集污染、数据泄漏等风险。此外，目前数据标注大多依靠人工标注，人工处理个人敏感信息时可能会遇到信息使用超出用户授权范围的隐私合规风险，例如智能语音助手被曝出员工监听分析事件。

三是自动化决策隐私合规问题。利用个人信息训练的人工智能系统，通常需要考虑自动化决策的合规问题。欧盟GDPR及29条工作组相关文件

对于自动化决策有相关要求，需要防止自动化决策对弱势群体的歧视影响，同时要求人工智能系统适当阐释数据处理的意义和可能产生的结果，但机器学习算法的黑箱特征，使得数据处理过程缺乏可解释性。

（三）其他阶段的数据安全隐患

数据存储安全隐患。人工智能系统的数据，通常存储在云端的数据库、数仓等存储系统，或以文件形式存储在端侧设备。数据存储安全隐患主要体现在数据、模型的存储媒介安全，如果存储系统存在安全漏洞或模型存储文件被破坏都可能会造成数据泄漏。例如2019年2月深网视界被曝泄漏超过250万人的人脸识别信息，主要由于未对内部MongoDB数据库进行密码保护。

数据共享安全隐患。在数据采集和数据标注环节，许多人工智能公司会委托第三方公司或采用众包方式实现海量数据的采集和标注，数据链路中所涉及的多方主体的数据保护能力参差不齐，可能带来数据泄漏和滥用隐患。在人工智能系统运行阶段，存在很多向第三方分享和披露数据的情况，例如机器学习算法或模型训练由第三方完成，需要和第三方的人工智能API进行数据交互。

数据传输安全隐患。人工智能系统通常部署在云侧和端侧，云边端之间存在大量数据传输，传统数据传输存在的安全隐患都可能发生。

3.2.3 基础设施安全隐患

基础设施，是人工智能产品和应用普遍依赖的软硬件，如软件框架、计算设施、智能传感器等。其中，软件框架是通用算法模型组件的工程实现，为人工智能应用开发提供集成软件包和算法调用接口。基础设施为人工智能提供计算力资源，通常来源于智能芯片、智能服务器或端侧设备的边缘算力等。基础设施安全风险是人工智能面临的基础性风险，主要包括：

一是开源安全风险。当前人工智能技术和产业的快速发展，很大程度

上得益于主流人工智能软件、框架、依赖库等必要实验和生产工具的开源化，越来越多创业者能够依赖开源成果进行人工智能研究。开源社区在功能优化、框架设计等方面对人工智能的发展起到了关键作用，但往往忽视了其成果的安全风险。

二是软件框架安全风险。近年来，国内网络安全企业屡次发现TensorFlow、Caffe等机器学习相关软件框架、工具及依赖库的安全漏洞，这些漏洞可能被用于网络攻击，给人工智能应用带来新的威胁和挑战。

三是传统软硬件安全风险。人工智能基础设施由软件和硬件组成，也面临着传统的软、硬件安全风险，需要关注服务接口安全、软硬件安全、服务可用性等问题。

四是系统复杂度和不确定性风险。在许多情况下，人工智能系统被设计为在复杂环境中运行，具有大量潜在状态，无法对其进行详尽检查或测试。系统可能面临在设计过程中从未考虑过的情况。

五是系统行为难以预测。对于在部署后学习的人工智能系统，系统的行为可能主要由在无监督条件下学习的情况决定。在这种情况下，可能难以预测系统的行为。

六是人机交互安全风险。在许多情况下，人工智能系统的性能受人类交互的影响很大，不同人群可能有不同的反应特点，人类反应的变化可能会影响系统的安全性。

3.2.4 应用安全隐患

产品应用，是指按照人工智能技术对信息进行收集、存储、传输、交换、处理的硬件、软件、系统和服务，如智能机器人、自动驾驶等。行业应用则是人工智能产品和服务在行业领域的应用，如智能制造、智慧医疗、智能交通等。产品服务和行业应用的安全隐患主要表现在：

人工智能应用是依托数据、算法模型、基础设施构建而成，算法模

型、数据安全与隐私保护、基础设施的安全隐患仍然会存在，**并且呈现出人工智能应用攻击面更大、隐私保护风险更突出的特点**。

自动驾驶。由于增加了连接控制功能、IT后端系统和其他外部信息源之间的新接口，大幅提升了网络攻击面，使得自动驾驶面临物理调试接口、内部微处理器、运载终端、操作系统、通信协议、云平台等方面的脆弱性风险。

生物特征。在数据采集阶段可能面临呈现攻击、重放攻击、非法篡改等攻击威胁。生物特征存储阶段，主要是对生物特征数据库的攻击威胁。生物特征比对和决策阶段，存在比对结果篡改、决策阈值篡改和爬山攻击等安全威胁。生物特征识别模块间传输，存在对生物特征数据的非法窃听、重放攻击、中间人攻击等威胁。

智能音箱。存在硬件安全、操作系统、应用层安全、网络通信安全、人工智能安全、个人信息保护六方面脆弱性。例如，对于开放的物理端口或接口，攻击者可利用接口、存储芯片的不安全性，如直接拆解音箱硬件芯片，在Flash芯片中植入后门，用于监听获取智能音箱的控制权，篡改操作系统或窃取个人数据。

3.2.5 人工智能滥用

人工智能滥用，包含两层含义：一是不当或恶意利用人工智能技术引发安全威胁和挑战；二是利用人工智能技术后造成不可控安全风险。

一方面，人工智能在欺诈、违法不良信息传播、密码破解等攻击手段的应用，给传统安全检测带来了新的挑战。具体来讲，**一是网络攻击自动化趋势明显，**在网络领域，需要大量高技能劳动力的攻击活动（如APT攻击等）已逐步实现高度自动化。**二是不良信息传播更加隐蔽，**不法分子利用人工智能技术，使得各类不良信息的传播更加具有针对性和隐蔽性，给维护网络安全带来了巨大的隐患和挑战。**三是越来越多被应用于欺诈等**

违法犯罪中，2017年，我国浙江、湖北等地发生多起犯罪分子利用语音合成技术假扮受害人亲属实施诈骗的案件，造成严重后果和恶劣社会影响。**四是口令破解概率提升，**利用人工智能技术破解登录验证码的效果越来越好、且难以防范。2018年西北大学团队基于人工智能技术建立了一套验证码求解器，仅利用500个目标验证码优化求解器，便可使求解器在0.05秒之内攻破验证码。

另一方面，随着人工智能创新技术与各领域的交叉融合，在对各领域进行有益推动的同时，人工智能滥用问题逐渐凸显。ISO/IEC PDTR 24028将人工智能滥用分成三个层次：Misuse即过度依赖人工智能会导致无法预料的负面结果；Disuse即对人工智能的依赖不足带来的负面结果；Abuse即在建立人工智能系统时而未充分尊重最终用户利益。由于**创新技术应用边界难以控制**，可能引发滥用风险，如利用人工智能技术模仿人类，如换脸、手写伪造、人声伪造、聊天机器人等，除引发伦理道德风险外，还可能加速技术在黑灰色地带的应用，模糊技术应用的合理边界，加剧人工智能滥用风险。

3.3 安全影响

人工智能作为正在高速发展的新兴产业，其收益和风险并存。尤其随着人工智能在国防、医疗、交通、金融等重要行业领域的深入应用，如果出现严重安全事件或被不当利用，可能会对国家安全、社会伦理、网络安全、人身安全与个人隐私等造成影响。其中，网络安全主要是对网络空间安全（如信息安全、系统安全等）造成影响，个人隐私则是对个人信息保护权益产生影响。本节主要描述对国家安全、社会伦理和人身安全方面的影响。

一是国家安全影响。人工智能可用于构建新型军事打击力量，对国防安全造成威胁。自主系统和机器人在军事上的应用，将加快战斗速度、提升战斗能力，如生产具有自动识别目标和精准打击能力的人工智能武器、

通过生成对抗性网络来制造军事相关的伪装和诱饵、人工智能系统间通过电磁对抗和机器学习帮助改进无线电频谱分配等。**利用人工智能对目标用户进行信息定制传播，可达到社会舆论动员目的。**通过搜集用户行为数据，采用机器学习对用户进行政治倾向等画像分析，为不同倾向的用户推送其期望的内容，也可通过学习和模拟真实的人的言论来影响人们对事情的判断，一旦被恶意利用可能造成大范围影响。**人工智能在情报分析上的大量应用，增加了国家重要数据的泄露风险。**人工智能技术在情报收集和分析方面有很多用途，情报工作者可以从监控、社交媒体等渠道获取越来越多的数据，通过人工智能技术对海量数据进行挖掘分析，可以获得许多重要敏感数据。

二是社会伦理挑战。“机器换人”对中低技术要求的劳动力就业造成影响，长远会加剧社会分化和不平等现象。工业机器人和各种智能技术的大规模使用，使从事劳动密集型、重复型、高度流程化等行业的工人面临失业威胁。虽然也有一些研究报告对过度的失业担忧提出质疑，但从长远来看，人工智能会加剧社会分化和不平等现象，尤其对于受教育程度较低的人群，人工智能的普及会让他们的竞争力大幅降低。**对人工智能技术的依赖会对现有社会伦理造成冲击，影响现有人际观念甚至人类的交往方式。**例如智能伴侣机器人依托个人数据分析，能够更加了解个体心理，贴近用户需求，对人类极度体贴和恭顺，这可能降低人们在现实生活中的社交需求。**人工智能技术的高速发展对相对稳定的成文法律体系造成冲击，法律规制的滞后和缺漏难以避免，可能触发难以追责的法律困境。**例如目前人工智能正逐渐应用于医疗、交通行业，但如果人工智能诊断出现医疗误判、自动驾驶出现交通事故，应当由谁承担事故责任？如何区分人工智能模型的设计者、使用者的监护责任和人工智能系统自身的责任？既然人工智能有能力代替人类进行决策与行动，是否应当在法律上赋予人工智能一定的主体权利？相关部门、学术界和产业界需要深入的思考和讨论上述

问题。

三是人身安全风险。人工智能在攸关人身安全的应用领域，可能由于漏洞缺陷或恶意攻击等原因损害人身安全。随着人工智能与物联网的深入结合，智能产品日益应用到人们的家居、医疗、交通等攸关人身安全的领域，一旦这些智能产品（如智能医疗设备、无人汽车等）遭受网络攻击或存在漏洞缺陷，可能危害人身安全。**人工智能在开发武器等攻击领域的应用，如果不加约束将对人身安全构成极大威胁。**人工智能技术可能被用于开发武器，借助人脸识别、自动控制等技术开发的人工智能武器，如“杀人蜂”，可以实现全自动攻击目标。如果赋予人工智能武器自行选择并杀害人类的能力，将给我们的人身安全与自由构成极大威胁。

3.4 人工智能安全属性和内涵

针对人工智能面临的对抗样本、数据投毒、模型窃取等新型攻击威胁，人工智能的算法模型、数据、基础设施和产品应用主要面临算法偏见、算法黑箱、算法缺陷、数据安全、隐私保护、软硬件安全、滥用、伦理道德等安全隐患。

为了防范人工智能的新型攻击威胁和安全隐患，需要对传统网络安全的保密性、完整性、可用性、可控性和不可否认性等安全属性进行扩展。比如：解决算法偏见需要坚持公平性，针对算法黑箱需要加强可解释性或透明性，应对算法缺陷需要提高鲁棒性，针对滥用问题要重视可控性，面向伦理道德问题要做到以人为本，而数据安全、隐私保护、软硬件安全这些保护原则与以前类似。

综上所述，本白皮书给出人工智能安全的原则、属性和内涵为：

（一）人工智能安全原则

1）以人为本原则（Human Orientation）。是指人工智能的研发和应用应以人类向善、人类福祉为目的，保障人类尊严、基本权利和自由。

2）权责一致原则（Parity of Authority and Responsibility）。建立机制确保人工智能的设计者和操作者能对其结果负责，如准确记录、可审计性、最小化负面影响、权衡和补救等。

3）分类分级原则（Classification）：考虑到人工智能总体发展还处于起步阶段，可针对不同人工智能技术发展的成熟度，不同应用领域的安全需求，对人工智能的能力水平和特定功能建立分类分级的不同准则。

（二）人工智能安全属性

人工智能作为还未成熟的创新技术，为了保障其在重要行业领域深入应用时的安全，不仅需要保障人工智能资产的保密性、完整性、可用性等传统安全属性，也需要考虑鲁棒性、透明性、公平性等其他属性目标。

1）保密性（Confidentiality）：确保人工智能系统在生命周期任一环节（如采集、训练、推断等），算法模型和数据不被泄漏给未授权者。如防范模型窃取攻击。

2）完整性（Integrity）：确保人工智能系统在生命周期任一环节（如采集、训练、推断等），算法模型、数据、基础设施和产品应用不被植入、篡改、替换和伪造。如防范对抗样本攻击、数据投毒攻击。

3）可用性（Availability）：确保对人工智能算法模型、数据、基础设施、产品应用等的使用不会被不合理拒绝。可用性包括可恢复性，即系统在事件发生后迅速恢复运行状态的能力。

4）可控性（Controllability）：是指对人工智能资产的控制能力，防止人工智能被有意或无意的滥用。可控性包括可验证性（verifiability）、可预测性（predictability），可验证性是指人工智能系统应留存记录，能够对算法模型或系统的有效性进行测试验证。

5）鲁棒性（Robustness）：指人工智能面对非正常干扰或输入的健壮性。对人工智能系统而言，鲁棒性主要用于描述人工智能系统在受到外部干扰或处于恶劣环境条件等情况下维持其性能水平的能力。鲁棒性要求人

工智能系统采取可靠的预防性措施来防范风险，即尽量减少无意和意外伤害，并防止不可接受的伤害。

6）透明性（Transparency）：提供了对人工智能系统的功能、组件和过程的可见性。透明性并不一定要求公开其算法源代码或数据，而是根据人工智能应用的安全级别不同，透明性可有不同的实现级别和表现程度。透明性通常包括可解释性（Explicability）、可追溯性（Traceability），让用户了解人工智能中的决策过程和因果关系。可解释性是指在人工智能场景下，算法特征空间和语义空间的映射关系，使得算法能够实现站在人的角度理解机器。

7）公平性（Fairness）：指人工智能系统在开发过程中应当建立多样化的设计团队，采取多种措施确保数据真正具有代表性，能够代表多元化的人群，避免人工智能出现偏见、歧视性结果。

8）隐私（Privacy）：按照目的明确、选择同意、最少够用、公开透明、主体参与等个人信息保护原则，保护公民的个人信息。

（三）人工智能安全内涵

本白皮书认为人工智能安全仍然属于网络安全的一部分，依据《网络安全法》对网络安全的定义，将人工智能安全定义为：人工智能安全是指通过采取必要措施，防范对人工智能系统的攻击、侵入、干扰、破坏和非法使用以及意外事故，使人工智能系统处于稳定可靠运行的状态，以及遵循人工智能以人为本、权责一致等安全原则，保障人工智能算法模型、数据、系统和产品应用的完整性、保密性、可用性、鲁棒性、透明性、公平性和隐私的能力。

四、人工智能安全标准体系

4.1 人工智能安全标准化需求分析

目前，国内人工智能安全相关标准主要集中在生物特征识别、自动驾驶等部分领域的应用安全标准，以及大数据安全、个人信息保护等支撑类安全标准，而与人工智能自身安全或基础共性直接相关的基础安全标准还比较少。本文基于人工智能安全属性和内涵，结合当前人工智能面临的安全风险，参考国内外已有人工智能安全政策、标准研制方向，从人工智能算法模型安全、数据安全与隐私保护、基础设施安全、产品和应用安全、测试评估等维度进行了标准化需求分析。

结合人工智能安全风险分析结果和当前标准化现状和《人工智能标准化白皮书（2018）》中的人工智能模块划分，人工智能本身安全在以下方面存在标准化需求：

（一）人工智能算法模型安全标准化需求

考虑到人工智能算法模型面临鲁棒性、对抗样本攻击等方面的安全挑战，此外，SC42也已开展人工智能可信度、神经网络鲁棒性评估等方面标准研制。建议从人工智能算法模型安全需求出发，充分考虑我国应用的人工智能算法模型在鲁棒性、可信度方面的要求，研究人工智能算法模型安全指标，研制算法模型安全评估要求和算法模型可信赖类标准。

（二）人工智能数据安全与个人信息保护标准化需求

人工智能数据的完整性、安全性和个人信息保护能力是保障人工智能安全的重要前提，国内外已开展人工智能数据安全与隐私保护相关标准、技术研究，建议针对突出数据安全与隐私保护风险开展标准化研究工作，**一是针对人工智能数据集面临的数据投毒、逆向攻击、模型窃取等突出问**

题，围绕人工智能数据生命周期，开展数据集防护、算法模型保护、抗逆向攻击等方面的人工智能安全标准化工作。**二是平衡隐私保护和人工智能分析效果，**防范逆向工程、隐私滥用等安全风险，开展人工智能隐私保护要求及技术类标准研究工作。

（三）人工智能基础设施安全标准化需求

人工智能系统包括云侧、边缘侧、端侧和网络传输等部分，人工智能基础设施面临软件框架漏洞、传统软硬件安全等方面风险，除服务接口安全、软硬件安全、服务可用性等传统网络安全需求外，建议结合人工智能特有安全需求和特殊系统安全需求，针对人工智能的基础组件、系统和平台等基础设施，如开源算法框架、代码安全、系统安全工程等，研制人工智能信息系统的安全标准。

（四）人工智能产品和应用安全标准化需求

人工智能涉及数据、算法、基础设施等多个维度，产品和应用范围很广，产品和应用具有复杂度高、受攻击面广、安全能力不同的特点，但仍具有安全共性需求。建议优先针对产业发展较成熟、安全需求迫切、标准不完善的智能产品和应用考虑标准化需求，分析梳理国家和人身安全、伦理、智能化攻击等场景下的安全风险，研究不同产品和应用的基础共性和特异性安全需求。可优先从智能门锁、智能音箱、智能客服等人工智能产品选取标准化对象，研制产品和应用指南类、评估类标准。

此外，人工智能依托数据和算法模型、基础设施实现，具有组成体系复杂、风险维度多样、供应链复杂、安全运营要求高的特点，建议面向从事人工智能研究、应用的主体及人工智能产品和应用，研制人工智能安全风险管理、供应链安全管理、安全运营等类型标准。

（五）人工智能测试评估安全标准化需求

人工智能的复杂性使其面临算法模型安全、数据安全与隐私风险、基础设施安全等类型的风险和挑战，建议充分兼容已有支撑性安全标准，设

计人工智能安全测试评估指标，优先针对算法鲁棒性、人工智能系统应用可信赖、隐私保护、数据集安全、应用安全等主题开展测试评估类标准研制工作。

4.2 人工智能安全标准与其他领域标准的关系

近年来人工智能的飞速发展，离不开大数据、云计算等基础设施的有力支撑。形成智能化服务也离不开基础设施的支撑，因此，为保障人工智能安全，不应仅考虑人工智能本身的安全，还应综合考虑数据安全、算法模型安全、基础设施安全、网络安全等方面基础性安全威胁和挑战，充分兼容原有大数据、云计算标准及标准体系，面向有人工智能特点的领域开展标准研制工作，特别要研制人工智能基础性标准、人工智能场景下有特定安全需求的安全标准。此外，在生物特征识别、态势感知等较成熟的技术和应用场景下，人工智能扮演着越来越重要的角色，应当在标准化工作中兼顾此类已有安全标准，综合考虑人工智能该类场景下面临的安全风险，提出标准研制计划。

4.3 人工智能安全标准体系

如图4-1所示，人工智能安全标准体系结构包括基础，数据、算法和模型，技术和系统，管理和服务，测试评估，产品和应用等六个部分，主要反映标准体系各部分的组成关系。

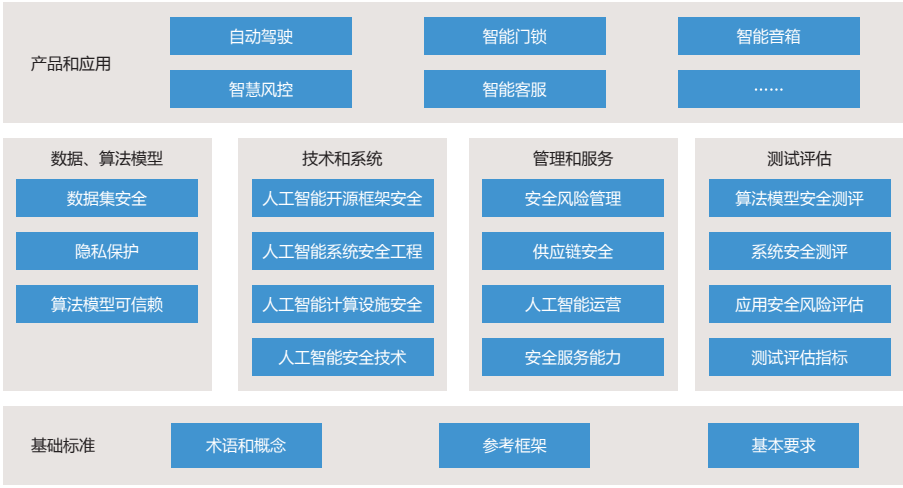


图4–1 人工智能安全标准体系

4.3.1 人工智能基础性安全标准

人工智能基础性安全标准包括人工智能概念和术语、安全参考架构、基本安全要求等。

——人工智能安全概念和术语是在人工智能安全方面进行技术交流的基础语言，规范术语定义和术语之间的关系，有助于准确理解和表达技术内容，方便技术交流和研究。该类标准需充分考虑ISO、ITU-T、我国国家人工智能标准化总体组等国内外标准化组织已发布的人工智能概念和术语的规范性定义。

——人工智能安全参考架构是理解和进一步研究人工智能安全的基础，可通过对人工智能角色进行安全分析，提出人工智能安全模型，规范人工智能安全体系结构，帮助准确理解人工智能安全保障包含的结构层次、功能要素及其关系。

——人工智能基本安全要求标准主要是为响应人工智能安全风险、法规政策要求，提出人工智能基本安全原则和要求，为人工智能安全标准体

系提供基础性支撑，可指导相关方开展人工智能安全建设，对数据保护、算法安全、内部信息系统的设计、开发和实现提出要求，为人工智能安全实践落地提供技术要求，切实保护人工智能安全。

4.3.2 人工智能数据、算法和模型安全标准

数据、算法和模型安全标准是针对人工智能数据、算法和模型中突出安全风险提出的标准，包括数据集安全、隐私保护、算法模型可信赖等。

——数据集安全类标准主要围绕人工智能数据的生命周期，保障数据标注过程安全、数据质量，指导人工智能数据集的安全管理和防护，降低人工智能数据集安全风险。

——隐私保护类标准基于人工智能开发、运行、维护等阶段面临的隐私风险，从隐私采集、利用、存储、共享等环节制定人工智能隐私保护安全标准，重点防范因隐私数据过度采集、逆向工程、隐私数据滥用等造成的隐私数据安全风险，该类标准应充分兼容TC260已有个人信息保护标准，重点解决人工智能场景下典型隐私保护问题。

——算法模型可信赖类标准主要围绕算法模型鲁棒性、安全防护、可解释性和算法偏见等安全需求，解决算法在自然运行时的鲁棒性和稳定性问题，提出面向极端情况下的可恢复性要求及实践指引，通过实现人工智能算法模型的可信赖，切实保障人工智能安全。

4.3.3 人工智能技术和系统安全标准

技术和系统类标准用于保障人工智能开源框架安全和人工智能系统安全工程。

——人工智能开源框架安全类标准针对人工智能服务器侧、客户端侧、边缘侧等计算、运行框架提出安全要求，除开源框架软件安全、接口安全、传统软件平台安全要求外，应提出针对人工智能开源框架的特定安

全要求，保障人工智能应用在训练、运行等环节的底层支撑系统安全。

——人工智能系统安全工程类标准针对安全需求分析、设计、开发、测试评估、运维等环节的安全需求，从数据保护、模型安全、代码安全等方面，针对隐私保护、模型安全等突出风险，提出人工智能应用安全开发要求和指南，研制安全工程实施指南。

——人工智能计算设施安全类标准针对智能芯片、智能服务器等计算设施的安全需求，提出人工智能计算设施安全要求和指南类标准。

——人工智能安全技术类标准针对人工智能安全保护和检测技术，如基于隐私保护的机器学习、数据偏见检测、换脸检测、对抗样本防御、联邦学习等，制定人工智能安全技术类标准。

4.3.4 人工智能管理和服务安全标准

人工智能安全管理和服务类标准主要是为保障人工智能管理和服务安全，主要包括安全风险管理、供应链安全、人工智能安全运营等。

——人工智能安全风险管理类标准主要从风险管理角度出发，应对人工智能数据、算法模型、技术和系统、管理和服务、产品和应用等多维度的安全风险，提出技术、人员、管理等安全要求和实践指南，引导降低人工智能整体安全风险。

——人工智能供应链安全类标准主要从供应链安全管理出发，梳理典型产品、服务和角色的供应链安全管理需求，参考已有ICT供应链安全管理标准研制思路，提出人工智能供应链安全管理实践指南，切实保障人工智能生产要素的供应安全。

——人工智能安全运营类标准主要针对人工智能服务上线、提交或正式运行后的安全运营问题，基于业界典型实践案例，从人员安全、运营安全、应急响应等角度提出实践指引，降低人工智能业务连续性安全风险。

——人工智能安全服务能力类标准主要针对人工智能服务提供者对外

提供人工智能服务时，所需具备的技术和管理能力要求进行规范。

4.3.5 人工智能测试评估安全标准

测试评估类标准主要从人工智能算法、人工智能数据、人工智能技术和系统、人工智能应用等方面分析安全测试评估要点，提炼人工智能安全测试评估指标，分析应用成熟、安全需求迫切的产品和应用的安全测试要点，主要提出人工智能算法模型、系统安全、应用风险、测试评估指标等基础性测评标准。包括但不限于

——人工智能算法模型安全测评类标准主要围绕人工智能算法是否满足安全要求开展。

——人工智能系统安全测评类标准主要围绕人工智能系统运行是否满足安全要求开展。

——人工智能应用安全风险评估类标准主要围绕人工智能应用是否满足安全要求开展。

——人工智能安全测试评估指标类标准主要根据人工智能安全要求及具体对象安全需求，提炼人工智能安全测试评价指标，为开展人工智能安全测评奠定基础。

4.3.6 人工智能产品和应用安全标准

产品和应用类标准主要是为保障人工智能技术、服务和产品在具体应用场景下的安全，可面向自动驾驶、智能门锁、智能音响、智慧风控、智慧客服等应用成熟、使用广泛或安全需求迫切的领域进行标准研制。在标准研制中，需充分兼容人工智能通用安全要求，统筹考虑产品和应用中的特异性、急迫性、代表性人工智能安全风险。

五、人工智能安全标准化工作建议

（一）重视完善人工智能安全标准体系

建议开展人工智能标准化工作，统筹规划人工智能安全标准体系，加强人工智能安全基础标准研究，深化人工智能应用安全标准工作。**一是统筹规划人工智能安全标准体系。**为确保人工智能安全标准研制工作有序推进，建议调研分析国内人工智能安全标准化需求，优先开展人工智能安全标准体系研究，标准体系应覆盖人工智能的基础、平台、技术、产品、应用等多个对象的安全需求，并能明确与大数据安全、个人信息保护、云计算安全、物联网安全等相关标准的关系。**二是抓紧研究和落实人工智能伦理原则。**围绕人工智能算法歧视和算法偏见等突出问题，分析梳理各场景下人工智能伦理需求，研制提炼人工智能伦理原则，指导人工智能相关标准落实原则要求。

（二）加快开展重点领域标准研制工作

人工智能具有涉及面广、应用场景复杂、安全需求类型多的特点，建议建立人工智能安全标准化工作推进计划，按照“急用先行、安全事件驱动”的思路研制人工智能安全标准，加速开展重点领域标准研制工作，有序推进人工智能安全标准化工作不断深入。**一是加强人工智能安全基础标准研究。**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。建议加强人工智能基础安全标准研究，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，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，掌握人工智能算法的安全威胁和保护需求，明确算法的通用安全原则和要求，强化

人工智能算法模型的安全性和鲁棒性。规范人工智能算法模型、智能产品的安全要求和测评方法，解决人工智能面临的数据质量、数据集安全等问题。**二是深化人工智能应用安全标准工作。**人工智能正越来越多的与各应用领域融合，应当开始研究人工智能产品和应用安全标准。建议在人工智能应用标准之前优先考虑智能产品的安全标准，基于全国信息安全标准化技术委员会已开展的智能门锁、智能家居等领域标准研制基础，抽取人工智能安全标准化特征和要求。在下一步工作中，优先针对存在标准化需求急切、应用较成熟、安全需求迫切、应用广泛，或较敏感的领域，开展人工智能产品和应用安全标准研制工作，完善已有标准的人工智能安全要求。**三是按照“充分研究，急用先行，安全事件推动”的思路进行标准研制，**建议优先立项安全需求迫切、应用成熟的安全标准，优先立项《人工智能安全参考框架》、《人工智能数据集安全》、《人工智能数据标注安全》、《机器学习算法模型可信赖》、《人工智能开源框架安全》、《人工智能应用安全指南》、《人工智能安全服务能力要求》等标准，同步开展人工智能基础性标准研究工作，研究应用安全风险评估类标准及智能制造、智能网联汽车等重点人工智能产品和服务类安全标准，逐步推进其他领域人工智能安全标准研究工作。

（三）大力推广人工智能安全标准应用实践

为提升人工智能安全标准的有效性和可操作性，解决人工智能安全突出风险，探索人工智能安全重难点问题标准化路径，建议深入开展人工智能安全标准的应用实践工作。**一是完善人工智能安全标准试点机制，**选取若干试点企业，开展标准适用性和实施效果评价，在应用实践中建立“实践跟踪、问题发现、经验总结、完善标准、反哺下一步标准化工作”的工作思路，推动人工智能安全标准化工作高速、高质量发展。**二是完善人工智能安全标准研究、宣贯及应用推广机制，**组织高校、科研院所和企业共同突破人工智能安全标准化工作难点，发挥各类单位优势，建立“产学研

用”一体化的人工智能安全标准研究、宣贯及应用机制，促进人工智能产业良性发展。

（四）切实加强人工智能安全标准化人才培养

人才是开展人工智能安全标准化工作的基石，建议建立健全多层次、多类型的人工智能安全人才培养机制。**一是培养人工智能安全专业人才，**建立面向专业技术、标准制定、宣贯培训、测试评估等方面培养方案。**二是鼓励高校、科研院所、企业建立合作，**探索人工智能安全综合型人才培养路径。**三是加强对人工智能安全及标准化项目的支持力度，**优化科研资源配比、管理和考核机制，确保相关领域人才集中力量解决重点人工智能安全问题。

（五）积极参与国际人工智能安全标准化工作

ISO、IEEE等国际组织已组织开展多项人工智能安全标准化研究工作，已在部分领域取得一定标准化成果，建议充分消化、吸收国际国外已有标准化工作成果，结合我国人工智能安全需求，探索具有我国特色的人工智能安全标准化工作路径。**一是紧密跟踪研究国内外人工智能安全标准化工作动态和发展趋势，**形成人工智能安全国际标准化研究成果，吸收国外标准研制经验，推动我国更好开展人工智能安全标准化工作。**二是不断提升我国在人工智能安全领域的国际标准影响力，**大力支持我国单位和专家参与国际标准化工作，加强人工智能安全标准提案研究，鼓励我国专家在国际标准化组织任职及担任国际标准项目编辑。**三是充分发挥我国国际标准化交流和合作机制，**结合我国人工智能产业丰富应用场景，开展人工智能安全重难点领域标准合作交流机制，借助国际、国外力量丰富我国人工智能安全标准化工作成果。

（六）尽快建立人工智能高安全风险预警机制

针对存在高危安全风险的人工智能技术、产品和应用，建议研究提出人工智能安全高危风险预警机制。**一是建立人工智能高危安全风险目录，**

梳理安全风险突出、产生安全问题后影响巨大的人工智能技术、产品和应用，并对目录中的风险条目进行分类分级。**二是建立人工智能高安全风险预警机制，**结合技术、应用和产品特点，提出风险预警方案。**三是研究制定人工智能高危安全风险管理标准，**以标准为出发点，综合技术、管理、评估等手段，从风险识别、分析、处理等方面提出人工智能高危安全风险管理方案。

（七）有效提升人工智能安全监管支撑能力

标准化工作能有力支撑人工智能安全监管落地，建议制定人工智能安全标准的指标评估体系。**一是建立健全人工智能的监管体系，制定配套标准，**政府应以标准为有力抓手，建立贯穿人工智能开发、设计、数据采集和市场应用全周期的监管体系，防止人工智能被非法利用或用于偏离既定目的的领域。**二是建议加快研究人工智能供应链安全管理机制，**研制人工智能供应链安全配套标准，提出针对电信、能源、交通、电力、金融等行业的人工智能供应链采购要求，推动相关标准在重点领域进行试点，为我国党政部门和重点行业进行人工智能供应链安全风险管理提供有益参考，为企业加强人工智能供应链管理提供实践指南。

附录A

人工智能相关安全标准

A.1 TC260人工智能安全标准研究项目

表A-1 国内人工智能安全标准研究项目

序号	标准内容	标准类型
1	**《人工智能安全标准研究》** 由全国信息安全标准化技术委员会研制。该项目是国内第一个国家人工智能安全标准研究项目。本项目通过调研国内外人工智能安全相关的政策、标准和产业现状，分析人工智能面临的安全威胁和风险挑战，梳理人工智能各应用领域安全案例，提炼人工智能安全标准化需求，研究人工智能安全标准体系。	研究
2	**《人工智能应用安全指南》** 由全国信息安全标准化技术委员会研制。项目旨在以人工智能应用为切入点，分析人工智能应用安全，为提出人工智能安全应用相关标准奠定基础。	研究

A.2 TC260人工智能安全相关标准

表A-2 国内人工智能安全标准

序号	标准内容	标准类型
1	**GB/T 20979–2019《信息安全技术 虹膜识别系统技术要求》** 由全国信息安全标准化技术委员会提出。标准规定了用虹膜识别技术为身份鉴别提供支持的虹膜识别系统的技术要求。	修订
2	**GB/T 36651–2018《信息安全技术 基于可信环境的生物特征识别身份鉴别协议框架》** 由全国信息安全标准化技术委员会提出。标准规定了基于可信环境的生物特征识别身份鉴别协议，包括协议框架、协议流程、协议要求以及协议接口等内容。	制定
3	**GB/T 37076–2018《信息安全技术 指纹识别系统技术要求》** 由全国信息安全标准化技术委员会提出。标准对指纹识别系统的安全威胁、安全目的进行了分析，规避指纹识别系统的潜在安全风险，提出指纹识别系统的安全技术要求，规范指纹识别技术在信息安全领域的应用。	制定
4	**《信息安全技术 汽车电子系统网络安全指南》** 由全国信息安全标准化技术委员会研制。通过吸收采纳工业界、学术界中的实践经验，为汽车电子系统的网络安全活动提供实践指导。	制定
5	**《信息安全技术 车载网络设备信息安全技术要求》** 由全国信息安全标准化技术委员会研制。旨在提出解决智能网联汽车行业中关于车载网络设备信息安全技术要求标准问题。建立科学、统一的车载网络设备信息安全技术要求标准。	制定
6	**《信息安全技术 智能家居安全通用技术要求》** 由全国信息安全标准化技术委员会研制。规定了智能家居通用安全技术要求，包括智能家居整体框架、智能家居安全模型以及智能家居终端安全要求、智能家居网关安全要求、网络安全要求和应用服务平台安全要求，适用于智能家居产品的安全设计和实现，智能家居的安全测试和管理也可参照使用。	制定
7	**《信息安全技术 智能门锁安全技术要求和测试评价方法》** 由全国信息安全标准化技术委员会研制。目标是针对智能门锁的信息安全技术要求和测试评价方法予以规定，解决特斯拉线圈攻击、生物识别信息仿冒、远程控制风险等智能门锁安全的新问题，使各研发单位在产品应用设计之初就对产品的信息安全设计与开发进行规范化考虑，以全面提升产品的安全性，促进行业的健康有序发展，保障包括智能门锁系统在内的网络空间安全，保障人民群众生命与财产安全。	制定

A.3 ISO/IEC JTC1/SC42人工智能安全相关的标准

表A-3 SC42人工智能标准研制工作情况表

工作组名称	召集人	工作情况
WG1基础工作组	加拿大	ISO/IEC 22989 《人工智能概念和术语》 ISO/IEC 23053 《运用机器学习的人工智能系统框架》
WG2大数据工作组	美国	ISO/IEC 20546《信息技术 大数据 概述和术语》（已发布） ISO/IEC 20547-2 《信息技术 大数据参考框架 第2部分：用例及衍生需求》（已发布） ISO/IEC 20547-5 《信息技术 大数据参考框架 第5部分：路线图》（已发布） ISO/IEC 20547-1 《信息技术 大数据参考框架 第1部分：框架和应用进程》 ISO/IEC 20547-3 《信息技术 大数据参考框架 第3部分：参考架构》
WG3可信赖工作组	爱尔兰	ISO/IEC TR 24027《信息技术 人工智能 人工智能系统和人工智能辅助决策的偏见》 ISO/IEC TR 24028《信息技术 人工智能 人工智能可信度概述》 ISO/IEC TR 24029-1《信息技术 人工智能 评估神经网络鲁棒性第1部分：概述》 ISO/IEC 23894《信息技术 人工智能 风险管理》 TR《信息技术 人工智能 伦理和社会关注概述》
WG 4用例和应用工作组	日本	TR 《信息技术 人工智能 用例》
WG 5计算方法与人工智能系统特征工作组	中国	TR《人工智能计算方法与系统标准化研究》

附录B

人工智能应用安全实践案例

（排名不分先后）

B.1 百度人工智能安全实践

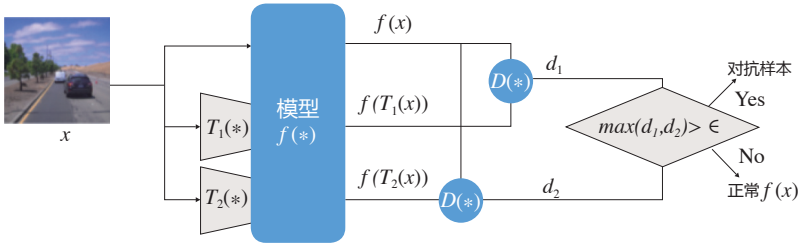
对抗样本攻击已经从实验室环境真正进入了网络对抗实战，增加了对个人隐私、财产安全、交通安全和公共安全的威胁，制约了人工智能在各行业的可信健康发展。针对对抗样本的产生涉及深度神经网络的可解释性等基础问题，目前还没有彻底的解决方法。百度在人工智能落地的过程中，逐步形成了从安全验证、模型加固、对抗样本检测到模型鲁棒性形式化验证的整体解决方案——AdvBox。

1）安全验证：旨在验证模型的安全，通常从是否对环境变化和对抗样本敏感两个维度去验证模型的安全性，验证形式包括白盒验证和黑盒验证。对于深度神经网络结构和参数完全公开的模型，AdvBox构造了针对性的对抗样本数据集，检验模型在不同扰动下的对抗表现；在只能获取有限模型信息时，AdvBox通过试探模型对不同输入的预测结果进行验证。以各类公有云服务商提供的人脸识别API为例，可以将不同的对抗样本上传并记录预测结果，以逐渐增强模拟攻击效果。此外，由于对抗样本往往具有迁移性，即一个模型所构造的对抗样本往往也能欺骗其他的黑盒模型，AdvBox能够利用对抗样本的迁移性进行模型安全性测试。用户能够利用AdvBox模拟针对模型的攻击，在不访问目标模型、仅在本地对于构造的类似模型进行攻击的情况下，利用产生的对抗样本欺骗目标模型。

2）模型加固：在发现人工智能模型的安全缺陷后，AdvBox会采取对抗训练、输入数据预处理和生成对抗网络（GAN, Generative Adversarial

Network）等技术手段对模型进行加固保护。其中，对抗训练是指在训练集中加入对抗样本，使得模型对对抗样本有更好的抵御能力。此外，AdvBox还通过数据预处理、修改模型的激活函数或者损失函数、网络特征压缩等方法，减少扰动对模型预测准确率的影响。最后，AdvBox基于GAN训练神经网络模型，以增强模型的泛化能力。加固后的模型能够有效降低黑盒、白盒攻击的成功率。

3）对抗样本检测：越来越多的人工智能服务通过API的形式对外提供，为抵御暴露在公网的人工智能API受到的对抗样本攻击风险，AdvBox在模型加固的基础上，通过对抗样本检测技术检测输入数据的合法性。综合使用的检测方法包括基于局部本征维数、模型可解释性方法、连续帧预测结果一致性对比等。在基于深度学习的检测任务中，AdvBox一旦检测到对抗样本，将立即通过上报异常的方法阻止该样本绕过。这对于诸如恶意软件检测、互联网内容合规审查等应用场景具有重要意义，黑灰产从业者总是尝试各类办法绕过恶意软件检测和内容审查，对抗样本也会成为他们的工具之一，所以，具备检测对抗样本的模型非常重要。具体而言，对抗样本检测流程如图B-1所示，x是输入，函数T是对输入所进行的变换，函数D是对预测结果不一致性的度量。如果进行不同的变化后，预测结果的不一致性超过了阈值，则判定为对抗样本。



图B–1 对抗样本检测流程

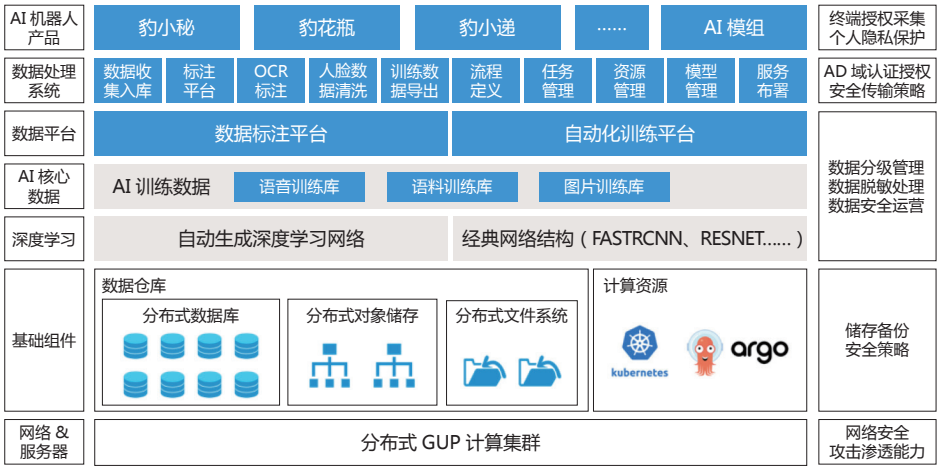
4）模型鲁棒性形式化验证：鉴于深度学习模型的超高维特性和复杂结构，如果使用样本测试的方法进行安全验证，我们无法确定选出的有限

样本是否能够完全覆盖所有代表性场景，无法有效证明模型的安全性；此外，遍历验证也同样无法有效证明模型的安全性。AdvBox利用形式化验证方法进行模型安全防护，具体而言，从构造对抗性样本上讲，AdvBox会找出所需扰动的值域和下界，以下界数值的大小为度量来判定模型对恶意攻击的安全性。在极端场景下，使得模型自动失效，并启用备选方案避免人工智能误判对人身财产或者公众造成伤害。

B.2 猎户星空人工智能安全实践

人工智能训练数据是所有人工智能技术公司的核心数据资产，训练数据的安全是确保人工智能模型效果符合预期的核心基础。语音识别、语音合成、视觉识别、自然语音处理等都需要大量训练数据支撑，人工智能数据包括通用格式的语音数据文件、对话语料文本文件、人脸图片数据或物品图片数据等多种类型，数据量普遍较大。若未对人工智能数据进行妥善管理，容易导致人工智能数据管理混乱、人工智能数据集安全性不高、隐私泄露、人工智能模型训练效果不佳等问题。

为了加强人工智能核心数据资产的安全管理，同时要保障人工智能训练过程的高效和稳定，很难将传统的数据加密和解密方式应用于人工智能数据安全保护方案中。猎户星空自主研发了人工智能自动化训练平台，通过统一训练平台，将训练任务和训练数据隔离，实现人工智能训练过程自动化，训练结果的可视化；将训练数据隐藏于隔离的存储介质，能够确保大数据资产的安全，也能优化GPU计算资源的调度管理，实现资源最大化使用。



图B–2 猎户星空人工智能自动化训练平台架构图

如图B-2所示，基于猎户星空人工智能业务特性和数据权属，猎户星空参照大数据相关安全标准设计了人工智能自动化训练平台的系统架构。

首先，在业务模式设计上，人工智能训练服务平台所使用的GPU服务集群，采用了统一域账户授权管理。基于分布式文件存储方式，按照语音、视觉、自然语言理解的不同训练需求进行分组管理，研发人员发起训练任务时，只需选择训练目标，设置数据范围，选择训练类型，系统即可根据预置信息进行计算资源分配和相关训练数据挂载。该训练平台确保了1）业务整体对数据授权边界的合理清晰，2）数据的处理逻辑基于可用不可见的安全原则，3）数据的应用产出基于数据价值而不是裸数据输出。

其次，自动化训练平台基于数据业务链路构建了全面的数据管控体系，包括数据加工前、数据加工中、数据加工后、数据合规等方面的数据安全管控。其中，在数据合规层，平台参考了GB/T 35273-2017《信息安全技术 个人信息安全规范》、GB/T 35274-2017《信息安全技术 大数据服务安全能力要求》、GB/T 31168-2014《信息安全技术云计算服务安全能力要求》和 ISO 27001 系列标准等国内外数据安全标准规范的要求，实现了个人隐私信息保护、云服务安全，保障了大数据服务的安全性。

再次，根据标准的大数据安全分级要求，自动化训练平台采用AD域认证授权方式，确保员工只有经过授权才能登录系统；将所有训练任务的数据范围、数据规模经过数据分级管理，根据岗位角色和职责进行授权；并利用堡垒机对员工的所有操作行为进行留痕记录。

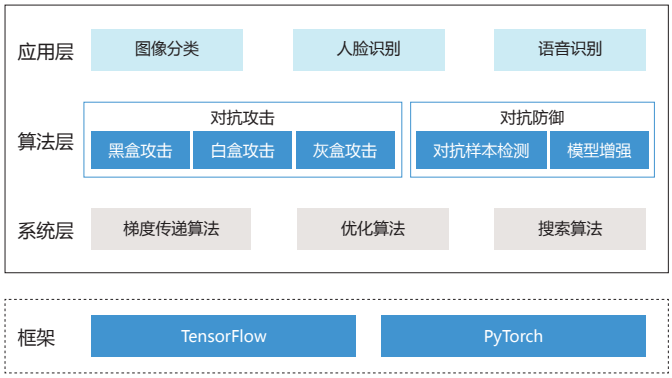
最后，自动化训练平台对基础设施安全、存储安全、系统安全、应用安全、平台网络服务安全进行了全方位的保护。基于《大数据安全标准化白皮书》，猎户星空建立了人工智能数据安全制度、原则、策略、管理方案以及实施细则，能确保人工智能数据在资源整合、共享、发布、交换等过程中的安全性。

B.3 清华大学人工智能安全实践

人工智能在飞速推动技术革命和产业进步的同时，其安全风险往往被人忽视。研究发现，许多在数据集上表现良好的算法非常容易被人眼不可见的对抗样本所欺骗，导致人工智能系统判断失准。如何构建一个可以支持多种不同对抗样本攻防的安全性平台引起了相关研究者的重视。利用这些平台可以使得研究者更方便的实现对抗攻击和防御，从而研发更加鲁棒和安全的深度学习模型，为人工智能的对抗性研究提供模型的表现基准。但是，目前的平台大多只支持对抗攻击算法的研究，缺乏对防御算法的支持，多数平台也不支持对损失函数、对抗检测等功能的灵活定义，并且对模型鲁棒性评估和比较也缺乏统一的标准和对比指标。

清华大学针对人工智能对抗性攻防的典型问题，从不同的应用场景、不同的威胁场景等方面，研发了RealSafe算法平台，从系统、算法和应用三个层次涵盖了标准程序库。RealSafe是开源标准程序库，业界可以免费对其进行非商业用途使用，为我国人工智能安全的理论算法研发和标准制定提供平台支持。其中，系统层可实现不同对抗攻防算法的共性通用模块，包括模型结构、损失函数设定等不同的模块支持；算法层从对抗攻击和对抗防御两方面，支持主流的对抗攻击和防御方法的高效实现；应用层从图像、视频、语音、网络数据等不同层面，支持相关应用的对抗攻防验证。对抗攻防平台的研制将大幅度降低相关模型的研发和使用门槛，并可以通过通用算法模块的研制降低新模型的开发成本，为各类人工智能模型的安全性、各类针对人工智能模型的攻防算法的性能提供统一的评价标准。

RealSafe平台针对目前人工智能方法安全的标准化需求和已有平台的不足，能够抽取不同算法的共性模块，支持对不同安全应用的灵活设定，提高人工智能攻防算法的开发和研究效率，评测主流人工智能方法的安全性，可以在很大程度上满足未来人工智能安全标准化测试的需要。如图B-3，本平台包括：



图B–3 人工智能安全平台架构

1）攻防算法性能基准：基于定义好的针对不同的应用场景的模型接口，实现针对不同的威胁场景的攻击与防御算法，进而提供对于具体的人工智能模型鲁棒性的评估基准，降低评估新模型鲁棒性、攻击新模型的开发成本。已有的人工智能安全平台对于典型的攻防算法支持不全面，因而无法提供一个客观全面的评估基准，本平台针对人工智能模型的对抗安全问题，建立完善的指标体系，实现对人工智能算法安全性的评估。

2）算法平台分为应用层、系统层和算法层三个层次，具体包括：

系统层：借助合理的抽象向系统层提供统一的基本算法接口，包括：支持不同的机器学习框架的梯度传递算法，支持的机器学习框架包括 TensorFlow 、PyTorch 等；多种优化算法，包括基于随机梯度下降的算法、拟牛顿法等；多种搜索算法：例如一些黑盒攻击方法中使用的基于随机游走的对抗样本搜索。

算法层：提供典型的针对多种威胁模型的对抗攻击算法和对抗防御算法的支持。对抗攻击部分包括：(1) 黑盒和灰盒攻击：包括基于决策边界攻击、基于迁移的攻击等方法，(2) 白盒攻击：包括多种基于梯度信息的攻击方法。对抗防御部分包括基于对抗样本检测的防御、基于模型增强的防御等。

应用层：提供对于图像分类、人脸识别等典型应用场景的典型模型的支持；提供接入系统层对抗攻防模块的模型接口，以降低向框架中添加新的应用模型的开发成本；提供模型安全性及攻防算法性能评估接口。

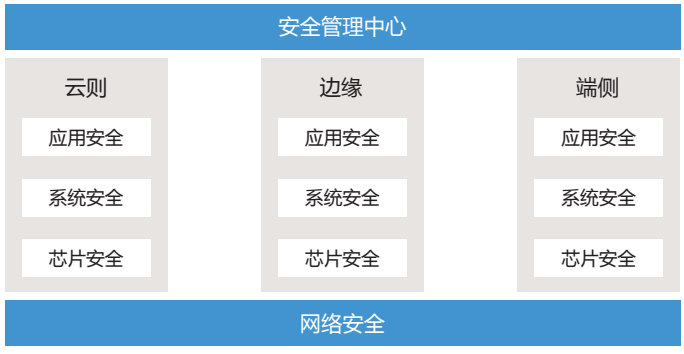
B.4 依图人工智能安全应用实践

针对典型人工智能安全问题，依图提出了人工智能系统安全技术的相关要求与解决方案。依图将人工智能系统分为了设备要素和网络要素两大类，其中设备要素指存储、处理或应用人工智能相关信息的计算机产品，包括芯片、操作系统、应用软件、其他相关软硬件等；网络要素指传输、交换或共享人工智能相关信息的计算机产品或应用，包括网络传输等，如表B-1所示为各组成要素与对应的安全要求。

表B-1 人工智能系统组成要素对应安全要求

要素	组成	隐私保护	算法公正	透明监管
人工智能设备	芯片	✓	✓	✓
	操作系统	✓	—	—
	应用	✓	✓	✓
	其它	✓	✓	✓
网络	网络传输	✓	—	—

从实际应用场景来看，如图B-4所示，依图将人工智能系统分为了云侧、边缘侧、端侧以及网络传输四个部分，并在这四个部分的基础上建立了安全管理中心，建立了人工智能系统的安全体系框架。



图B–4 人工智能系统安全框架

以人脸识别为例，人脸识别信息系统包括云侧的私有云存储及海量数据模型计算、边缘侧的识别算法、端侧的数据采集，以及云、边、端三侧间传输数据的网络链路等组成要素。具体来讲，

1）建立“安全委员会”，确保人工智能应用从“生产”到“应用”的全链路安全

依图通过“安全委员会”严格把控和执行各个方面的安全策略，最大程度确保人工智能应用的信息安全策略和各项措施能够落实到位。

2）依图“求索”芯片采用全链路安全设计，从各个层面确保芯片的使用安全可靠

依图“求索”神经网络芯片从算法研发到固件发布烧制，形成了完整的安全技术方案及管理规范，确保芯片的安全可靠。从技术上来看，

芯片内设置加密区域：“求索”芯片内预制了加密区域，通过唯一标识保证系统不被盗用，避免服务器克隆风险。

使用TrustZone技术：芯片硬件和软件资源划分为安全域和非安全域。通过该技术确保在安全域内进行关键运算过程与核心密钥交换。

数据模型加密：提供对算法核心数据模型的加密功能。保证模型数据只能被可信任的应用程序访问调用。保证数据模型文件不会遗失。

自研传输加密算法：采用自研加密算法，保证数据传输安全可靠。

此外，芯片还采用了固件防读取，防物理攻击等安全设计方案。

3）在人工智能算法中加入安全设计，确保避免“算法歧视”、“模型泄露”、“样本攻击”、“数据损坏”等问题

在算法设计方面，通过技术和管理手段，确保算法的可解释性，避免产生算法歧视。在人工智能应用中，利用多维方向提升复杂度、结合噪声噪点训练等方法防特征逆向攻击，确保算法模型的安全不泄露。此外，在算法模型训练的过程中，设计相关流程和工具，通过在算法模型中添加适当标签，在实际应用中结合标签与输入样本的对比，发现攻击者对输入样本的改

动，定位污染参数，修正或拒绝污染样本，在一定程度上防止恶意的对抗样本攻击。此外，还可以利用多种目前理论算力不可破解的对称及非对称加密算法和多重数字签名技术，进行数据加密和数据传输路径加密，通过保证数据的完整性，不可篡改性，实现以数据为中心的安全。依图对人工智能应用加强了审计设计，确保审计的全面、规范、独立、安全。

4）面向数据标注规范、标注数据校验、数据输入输出校验等流程，建立了的技术和制度防护等方面的数据治理机制。

在训练数据安全性方面，着重防止入侵者污染数据，一方面通过自动化工具对标注数据进行基本校验，并在训练模型的输入输出过程中再次进行偏差校验，避免污染数据进一步污染人工智能模型。另一方面，在人员管理中开展安全意识培训，制定和实施安全操作规程。

5）部署了局域网环境下“非密码安全登录”等功能的基础环境安全组件、加密通信和自研加密算法。

通过该基础环境安全组件，在集群登录中仅能通过主节点登录，无法直接登录集群内子节点。在端口管理中，该基础环境组件能够对应用的端口进行“白名单”管理和权限控制。在系统安全中，依图通过加强操作审计、实现审计模块、嵌入自主开发的操作系统等手段，确保了系统的安全性。在网络安全方面，设计了面向“云”、“边”、“端”的终端身份鉴别机制，在传输通道部署了自研加密算法。

B.5 IBM人工智能安全实践

人工智能意义重大，正越来越深刻地影响人类决策。为此，IBM提出人工智能三原则[1]，包括：

1）重在目的：IBM开发和应用的人工智能的目的是增强人类的智力和能力。人工智能应该继续由人类控制。

2）强调透明：人们对人工智能使用的信任是至关重要的。信任的基础是透明。需要知道人工智能是如何达成可靠、公平和可解释。

3）普及技能：人工智能将广泛应用于人类生活的方方面面，只有普及人工智能技能才能使大众充分了解、信任和利用人工智能[2]。

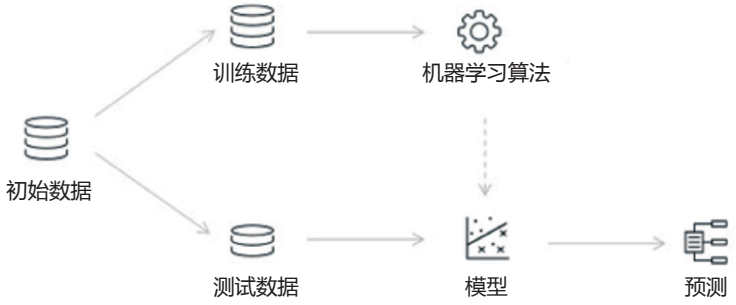
为达成人工智能的可靠、公平、可解释并贯彻在人工智能的整个生命周期，IBM 研发并开源了一系列可信人工智能的关键技术，其中，“人工智能公平360工具箱”（AIF360）[3]可用于检测和缓解机器学习模型中的偏见；对抗性鲁棒性360工具箱（ART）[4]可用于快速制作和分析机器学习模型的攻击和防御方法；人工智能可解释360（AIX360）[5]可用于支持机器学习模型和算法的可解释性。这一系列开源技术有助于促进可信人工智能的创新和应用。为满足企业对技术整合支持、保障及服务的需求，IBM将这一系列技术和IBM多年的人工智能实践集成到了企业版IBM Watson OpenScale[6]中，以满足企业需求。以AIF360为例，AIF360体现了IBM可信人工智能的透明性。

AIF360可帮助检测、缓解或消除机器学习模型中的偏见。AIF360包含一套全面的数据集、模型、指标以及检测、缓解或消除机器学习模型中的偏见的算法。

[1] http://ibm.biz/IBMAIPrinciples
[2] http://ibm.biz/IBMMITAILAB
[3] http://ibm.biz/ibmaif360
[4] http://ibm.biz/ibm_art
[5] http://ibm.biz/ibmaix360
[6] http://ibm.biz/ibmOpenScale

1）偏见与机器学习

AIF360设计了检查数据集、机器学习模型、机器学习算法等是否具有偏见的公平性指标和偏见缓解器。公平性指标可用于检查机器学习工作流中的偏见。偏见缓解器可用于克服工作流程中的偏见，以产生更公平的结果。



图B–5 AIF360工作流

如图A-5所示，偏见可以在工作流中的任何一个环节进入系统。训练数据集、算法、测试数据集都可能引发偏见，具体而言，（1）在训练数据集中，训练数据可能偏向于特定类型的实例。（2）在算法建模方面，算法可能生成针对输入中特定特征加权的模型。（3）在测试数据集方面，测试数据集对正确答案的期望可能有偏见。

2）检测和减轻偏见

偏见检测过程从一个初始数据集开始，采用随机分割算法将初始数据集分割为训练数据集和测试数据集。

首先，加载初始数据集时，设置受保护的属性，为该属性设置特权组和非特权组。在测试中比较特权组和非特权组的有利结果百分比，从后者中减去前者，若结果为负值则表示该属性对弱势群体不利，该结果即为检测到的偏差。

为在训练数据集中减少这种偏差，AIF360提供了多种用来检测偏见的指标及偏见缓解算法。例如“重新加权算法”能将适当的权重应用于训练

数据集中的不同组，以使训练数据集在敏感属性方面免于区分。通过这样转换后生成的新训练数据集会在特权和非特权组的年龄属性上获得更公平的结果。

然后，使用与测试原始训练数据集相同的办法，检查转换后的新训练数据集在消除偏差方面的效果。在已有数据集已获得偏差为0的结果，0差异证明了偏见缓解方法的有效性，说明有偏见倾向的数据集偏见被消除了。

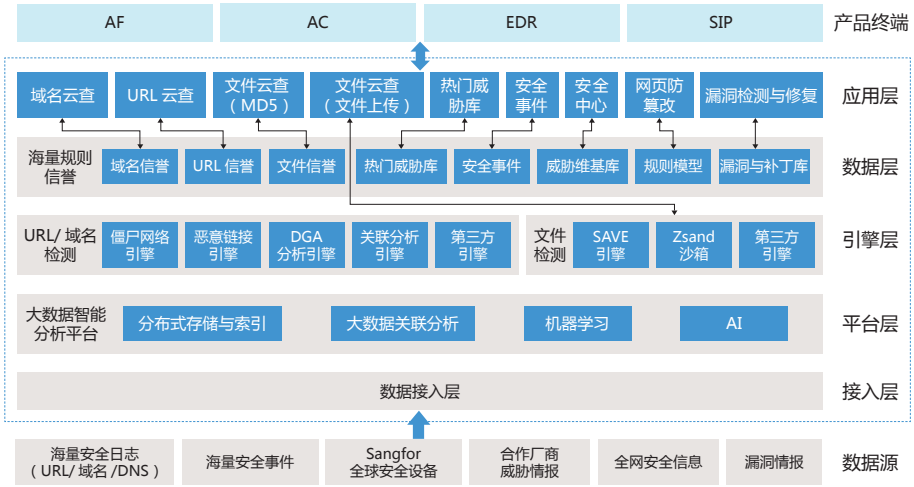
B.6 深信服人工智能安全实践

人工智能技术赋能网络安全领域，有助于厂商、企业、个人有效提升应对网络欺诈、恶意攻击等网络安全问题的能力。由于人工智能相对于传统的技术，具备泛化能力强，能力退化慢，占用宽带及内存少等优势，使人工智能成为网络安全防护的突破口。深信服安全云脑具备威胁情报和未知威胁检测能力。其中，

1）威胁情报中心汇聚了深信服在线安全设备、第三方安全厂商、安全社区等渠道的海量情报数据，基于大数据和云计算技术，利用各类数据分析及人工智能算法，形成了威胁情报中心。

2）未知威胁检测中心能够利用云端沙箱、人工智能杀毒、人工智能检测引擎、蜜罐技术、URL鉴定引擎等威胁检测引擎，结合资深安全分析师人工分析手段，对未知可疑威胁进行鉴定识别；

如图B-6所示，从物理架构上，安全云脑分为接入层、平台层、引擎层、数据层和应用层。



图B–6 安全云脑架构

攻击者针对人工智能的攻击手法越来越呈现出多样化且不断演进的特点，使得现有安全云脑的人工智能模型可能出现失效、误判等情况。为保障安全云脑的可靠性，应及时更新模型以应对攻击者对模型的攻击。然而，更新模型可能面临几方面安全风险，1）仅基于用户侧的数据更新，无法保证数据的无偏性与完备性，可能使得用户侧模型可靠性下降；2）如果基于少量且局部的数据更新，容易被攻击者精心伪造的数据欺骗，破坏真实数据的分布，使得模型做出错误判断；3）基于所有的真实数据更新，如果样本分布不均衡，可能导致模型对数量少的种类预测准确性不足。

因此，深信服为安全云脑建立了统一的数据收集和处理机制，保证数据来源的多样性与可靠性，在云端进行模型的更新训练并将新的模型下发到所有安全设备，提高人工智能模型的抗攻击能力、准确性等。此外，在应用过程中，深信服提升了安全云脑人工智能模型的可解释性、抗攻击能力、准确性等，具体来讲：

可解释性：通过结合人工智能与启发式规则，发挥安全专家与数据专家的能力，数据专家会对模型持续更新，提升人工智能的泛化能力，安全专家会对检测出的变种进行规则提取和更新，从而增强规则的直观性。通过人工智能+规则的闭环迭代，能够增强可解释性。

抗攻击能力：模型在线更新基于海量的信誉库以及多维度情报源，结合关联分析、智能算法构建的深度分析系统，通过对各类数据层层筛选，并进行全方位的分析，确保数据来源的多样性及可靠性，最终输出高准确度的判定结果。

鲁棒性：a、模型会持续进化，会周期性进行自我测试，一旦准确度低于某个阀值或者全新的不在模型覆盖的样本出现，则触发模型在云端进行训练，以确保模型的鲁棒性。b、关联分析，如图A-7所示，该引擎主要是基于深度学习、图计算等算法构建，并辅助安全专家的分析，形成人机

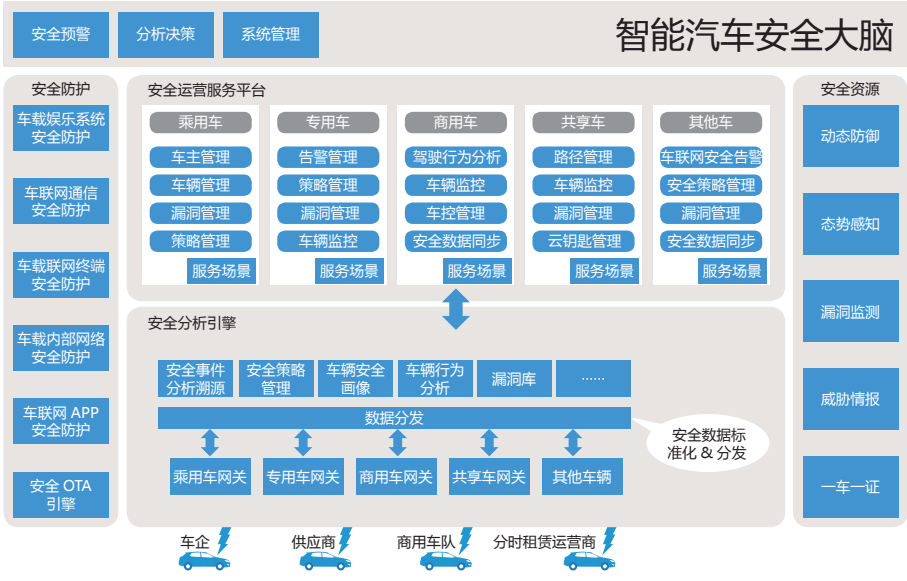
共智。核心关联模块使用深度学习算法分析DNS的流量特征、共同出现模式等，家族聚类模块基于图计算社区关联分析相关技术，通过聚类发现未知家族。举证模块结合多种维度信息对前面两个模块的结果进行举证，从而增强结果的鲁棒性。

图 B–7 关联分析引擎架构图

B.7 360 人工智能安全实践

智能汽车是智能交通的重要资产，随着其智能化、网联化的发展，给智能交通带来了安全风险。面对日益严峻的安全挑战，360通过基于人工智能分析的360“安全大脑”，打造智能交通安全动态防御体系，全面保障智能交通安全。

智能汽车安全大脑总体框架包括车联网安全分析引擎、车联网安全运营服务平台，同时通过安全防护和安全资源服务，提供安全预警、分析决策、系统管理等安全运营支撑，如图B-8所示：



图B–8 智能汽车安全大脑总体框架图

1）安全分析引擎

将车辆数据标准化后，安全分析引擎对车辆进行安全画像、行为分析、策略管理等。

2）安全资源

包含密码证书、安全存储等基础安全资源，支持SM2、SM3、SM4等

国密算法通信，同时提供漏洞监测、威胁情报等资源。

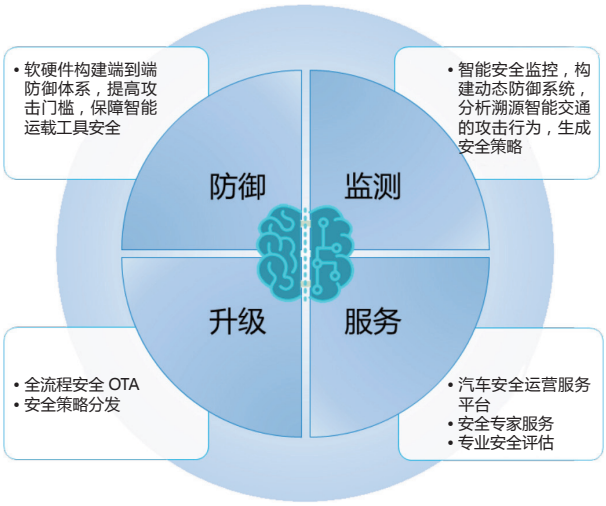
3）安全防护

包含车载娱乐系统安全防护、车载联网终端安全防护、车联网通信安全防护、车载内部网络安全防护、车联网APP安全防护等。及时下发安全补丁提前保护其他车辆，保障整体智能交通的运行安全。

4）运营服务平台

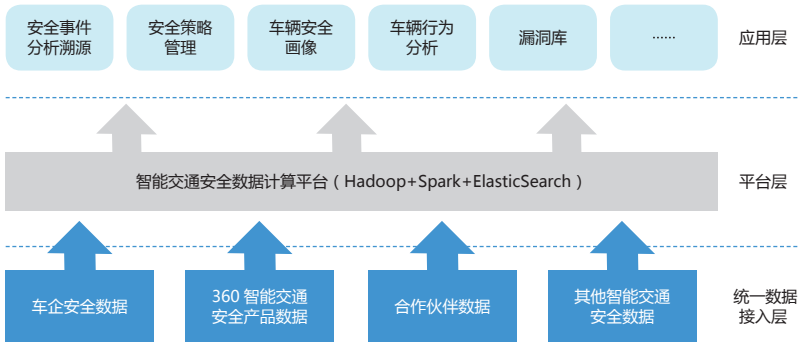
根据乘用车、专用车、商用车、共享车等场景提供针对性服务，针对不同场景下智能交通安全工作重点，通过资源整合，优化人工智能算法，提供漏洞管理、安全数据同步、安全策略管理等功能，能够提升智能交通安全运营能力。

智能汽车安全大脑核心能力包括防御、监测、升级、服务四部分，如图B-9所示。



图B–9 智能汽车安全大脑核心能力示意图

智能汽车安全大脑包括统一数据接入层、安全数据计算平台层和安全分析应用层三个方面，如图B-10所示。



图B–10 智能汽车安全大脑体系框架图

1）统一数据接入层

通过一套标准的数据接入接口，负责接入车厂的车联网安全数据、360车联网安全产品数据以及合作伙伴数据，将安全数据转发至安全数据计算平台层，实现与车厂平台高效实时的数据通道，满足不同车厂、不同车型、不同地域的智能网联汽车数据接入要求。

2) 安全数据计算平台层

车联网安全存储中心和计算中心可实现海量车联网数据的计算和分析挖掘。从数据处理能力来看，一方面由大规模Hadoop集群和Spark集群提供计算能力，另一方面通过搭建GPU集群支持基于人工智能技术的车辆画像和车辆行为数据的分析挖掘。

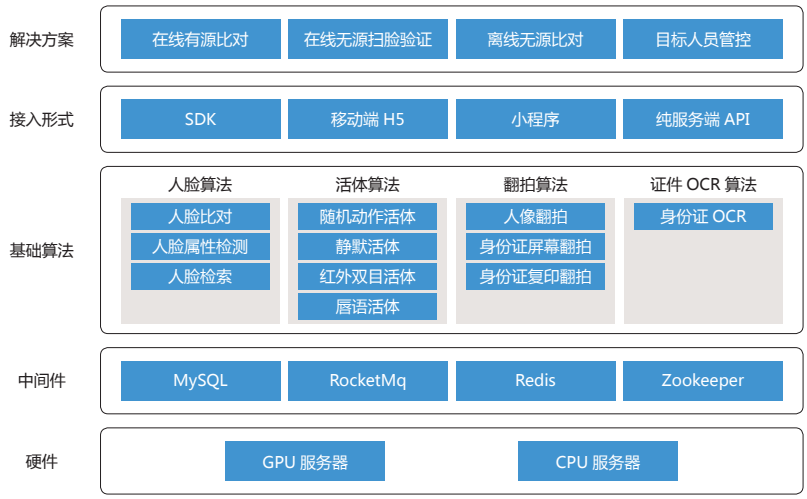
3）安全分析应用层

通过分析车辆行为数据，对车辆进行安全画像、行为分析，为安全运营提供汽车威胁情报支持，主要用于建立汽车信息安全事件的分析、溯源及响应机制，应用人工智能技术，实现汽车信息安全感知预警、动态防御和安全策略更新。此外，应用层还将建立汽车信息安全漏洞库，通过持续的监测发现联网汽车安全漏洞，监测潜在的汽车破解行为。

B.8 阿里巴巴人工智能安全实践

（一）实人认证

为解决传统网络身份认证手段无法解决身份冒用、身份合法性等带来的诸多安全威胁和风险，保障电商、社交、新零售等业务的用户身份真实有效，精准识别虚假身份，阿里巴巴利用人工智能技术构建了实人认证系统。实人认证系统技术框架如图B-11所示。

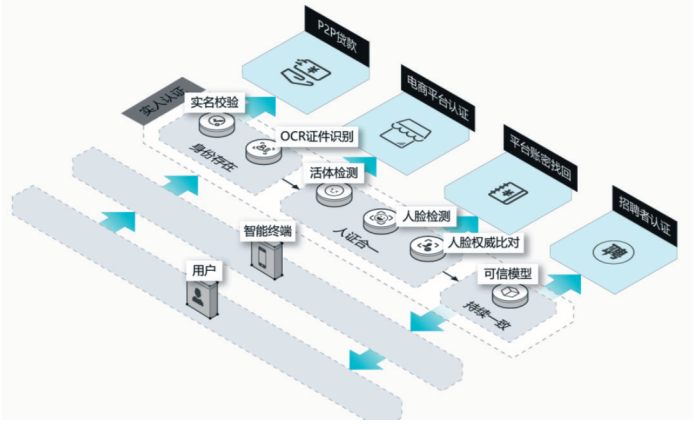


图B–11 实人认证系统系统技术框架

实人认证系统业务架构如图B-12所示。实人认证系统能够利用活体检测、人脸比对等生物识别技术、证件OCR识别技术等方法来提供实名校验和生物识别等服务，其中：

1）实名校验：采用证件OCR识别技术，自动识别并读取姓名、身份证号、有效期等信息，并与权威数据库进行验证比较。该技术综合识别率达到99%以上。

2）生物识别：通过视频获取照片进行验证识别对象是否为本人。为



图B–12 实人认证系统业务架构

防范非活体检测攻击，可通过互动操作达到鉴别真人的目的。在部分人脸识别使用受限的场景，基于人体的身份和动作识别能够提供有效补充。

（二）对抗样本攻防——问答型对抗验证码

针对人工智能模型本身的对抗攻击技术带来了诸多人工智能安全问题及风险，阿里巴巴结合业务需求对于对抗样本攻击防御技术总结如图B-13所示，可以分为对抗攻击、安全评估和对抗防御三个阶段。

图B–13 对抗样本攻击防御技术总结

基于上述技术体系，阿里巴巴在业务中进行了问答型对抗验证码以及安全评估和人工智能防火墙等诸多应用实践。

近年来，针对最常见的字符验证码，不法分子通过各种手段收集大量图像后，用机器学习技术进行OCR（光学字符识别）模型的训练，实现对验证码的自动识别，正确率可达80%以上，盗取用户账号、恶意注册薅羊毛等一系列犯罪行为都由此产生。

阿里安全在2018年底推出的新一代人工智能验证码产品正式上线，淘宝、天猫等业务场景均已使用。该产品结合知识图谱和对抗样本攻击技术，实现了在不影响用户体验的条件下，极大降低了被黑灰产进行机器破解的可能性。

1）知识图谱

基于结构化知识图谱建立的丰富的常识问答库可有效避免攻击，如图B-14所示。所谓知识图谱是由一些实体、实体属性以及实体之间的关系构成，比如大象和老虎的体重对比等。该知识图谱可生成亿级别的题库，用户回答时长9秒左右，而回答一次通过率为90%。绝大部分普通用户回答毫无难度，但通过机器来自动回答极为困难。

B–14 基于知识图谱的常识问答库案例

2）对抗样本

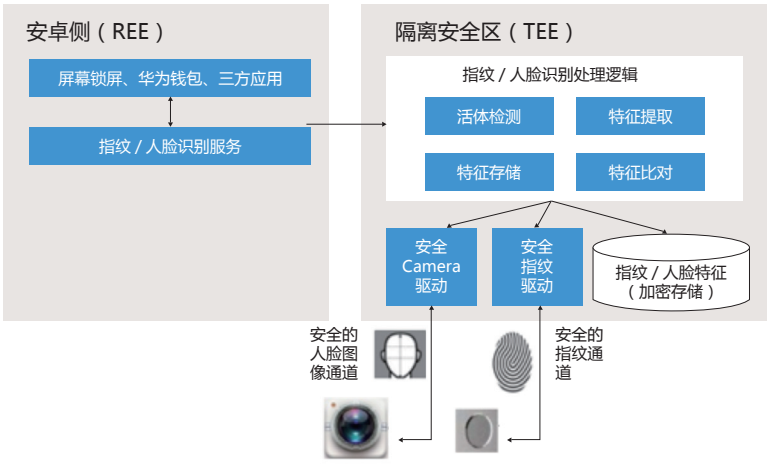
由于人机验证问答题的答案以图片形式展示给用户，黑灰产若要破解该类验证码，需要使用人工智能技术自动识别图片文字。阿里巴巴应用人工智能研究领域最新的对抗样本技术对原始图像有针对性的加入干扰，不影响肉眼识别，但会显著降低人工智能模型的识别率，从而防范打码平台的破解，同时保持用户体验。

B.9 华为人工智能安全实践

生物特征识别，是基于生物特征信息进行身份识别的一种技术，人工智能技术在生物特征识别中被广泛使用。目前，生物特征识别主要被应用于指纹/人脸解锁及指纹/人脸支付等场景，用以解决物理场景下的身份鉴别等安全问题。

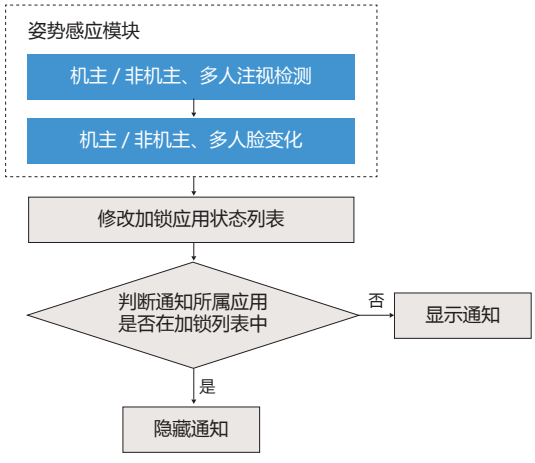
生物特征信息普遍具有较强的隐私属性，为保障生物特征信息的安全性，华为手机在芯片级隔离的可信执行环境（TEE，Trusted Execution Environment）中进行将指纹/人脸图像采集、特征提取、特征比对以及特征存储等过程，指纹/人脸特征数据通过TEE的安全存储或重放保护内存块（RPMB，Replay Protected Memory Block）进行加密存储，采用内置安全芯片实现对指纹/人脸特征数据的加/解密操作，确保个人敏感数据不出TEE。

指纹/人脸等生物特征信息数据通过安全的采集设备和安全通道被采集并传输到安全隔离区内，与密钥等敏感数据一起只存在于TEE中。同时，指纹/人脸验证（如：活体检测，特征提取，特征比对）、密码验证等操作也全部在TEE中完成，外部无法获取内置安全芯片内的指纹/人脸特征信息和加密密钥，确保人脸数据不会泄露，保护了用户敏感数据和相关业务的安全性。在安卓的人脸识别框架中，框架只负责处理指纹/人脸的认证发起和认证结果等数据，不涉及指纹/人脸数据本身。安卓的第三方应用也无法获取到用户的指纹/人脸数据，更无法将用户指纹/人脸数据传出终端设备，该机制进一步保护了用户的个人隐私和敏感数据的安全性，具体框架如图B-15所示。



图B–15 生物特征数据安全存储/处理框架

对于手机用户而言，除生物特征信息外泄的情况外，周围人对用户手机屏幕显示内容的偷窥也可能引发用户隐私和个人数据泄露风险。华为结合人工智能技术提出了防偷窥和用户隐私保护的解决方案。手机利用人工智能技术可以检测到是否存在机主/非机主、多人注视屏幕的情况，当手机识别出有陌生人或多人注视屏幕时，会将手机屏幕显示的私密信息自动隐藏，以便保护用户的隐私信息，具体实现逻辑流程如图B-16所示：



图B–16 基于人工智能技术的消息隐藏逻辑流程图

通过姿势感应模块，手机可以对机主/非机主、多人注视的情况进行识别，当识别出有机主/非机主、多人脸变化的情况时，则表明存在他人偷窥情况，此时，通过对加锁应用状态列表进行相应的修改，使得应用的加锁状态发生变化，判断屏幕显示的通知所属应用是否在加锁列表中，如果应用在列表中，则隐藏该应用的通知；否则，显示通知。简单来讲，如图B-17所示，当用户自己使用手机时，信息为展开状态；当手机检测到有陌生人/多人注视时，将通过“折叠”信息的方式保护用户隐私，确保应用信息只对用户自己可见。

图B–17 基于人工智能技术的消息隐藏功能实现效果展示

参考文献

[1] 国家人工智能标准化总体组.《人工智能标准化白皮书（2018）》，2018年1月.

[2] 谭铁牛.《人工智能的历史、现状和未来》.求是,2019年4月.

[3] 阿里云.《中国企业2020：人工智能应用实践与趋势》,2019年8月.

[4] 刘焱.《AI安全之对抗样本入门》.机械工业出版社.2019年6月.

[5] 国家工业信息安全发展研究中心.《2019年中国人工智能产业发展指数》,2019年9月.

[6] 德勤.《中国人工智能产业白皮书》,2018年11月.

[7] 清华大学.《人工智能芯片技术白皮书（2018）》,2018年12月.

[8] 上海观安信息技术股份有限公司.《人工智能数据安全风险与治理》.2019年9月.

[9] 张钹.《迈向第三代人工智能的新征程》,2019年9月.

[10] 谭铁牛.《人工智能：天使还是魔鬼》,2018年06月.

[11] 中国信息通信研究院.《人工智能发展白皮书—产业应用篇(2018年)》,2018年12月.

[12] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. CVPR 2018.

[13] ISO/IEC PDTR 24028 Information technology — Artificial intelligence (AI) — Overview of trustworthiness in artificial intelligence.

[14] 李盼,赵文涛,刘强,等.机器学习安全性问题及其防御技术研究综述[J].计算机科学与探索,2018(2):171-184.

[15] 国家人工智能标准化总体组.《人工智能伦理风险分析报告》,2019年4月.

[16] 刘劲杨《人工智能算法的复杂性特质及伦理挑战》,光明日报,2017年9月.

[17] 欧盟.《人工智能伦理准则》,2019年4月.

[18] 丛末.《笑谈中国AI发展态势，张钹、李德毅、张正友、肖京同台共议「AI五问」》,AI科技评论,2019年9月.

全国信息安全标准化技术委员会
大数据安全标准特别工作组