CAICT 中国信通院

集智 蓝皮报告

可信人工智能产业生态发展报告

（2022 年）

中国信息通信研究院华东分院

中国信息通信研究院云计算与大数据研究所

2022 年 9 月

版权声明

前　言

随着新一轮科技革命和产业变革的深入推进，人工智能呈现爆发式成长，广泛应用于日常生产、生活的方方面面，社会各界对可信品质的关注度也提升到了前所未有的高度。近年来，各界均在不断探索将更多的可信理念从基础能力、算法技术、应用场景和产品设备等不同层面进行融合实践，实现了人工智能在安全性、可靠性、可解释、可问责等一系列内在属性的可信赖程度逐步提升，为构建我国可信人工智能产业生态提供了有益参考。

为广泛吸纳产学研用各界的优秀经验，中国信息通信研究院联合京东探索研究院及政产学研多家单位共同编写《可信人工智能产业生态发展报告》，对人工智能产业融合可信要素的发展态势进行总体分析，研判发展趋势并提出措施建议，希望能为社会各界提供借鉴和参考。报告主要观点如下：

全球可信人工智能发展态势向好。人工智能监管向立法执法拓展，欧盟发布全球首部《人工智能法案》，美国推出《2022 算法问责法案》，中国人工智能地方立法相继落地；稳定性、隐私保护成为热点，可解释性、公平性等研究正逐步开展；标准层面，行业组织、企业和研究机构共同推进，全力打造可信 AI 标准体系。

可信人工智能产业生态加快形成。可信人工智能伦理、法律研究进一步深入，在硬件、技术、应用及支撑体系等层面蓬勃发展，形成兼顾稳定性、可解释性、隐私保护和公平性，涵盖基础硬件、技术平台、产品设备、应用场景等多元化产业生态。

未来，可信人工智能向着形成产业共识、突出理念落地、优化技术布局、注重动态平衡、强化多元主体发展。凝聚强化产业共识，进一步向具体实践迈进，可信人工智能一体化研究和技术发展加速创新，能力之间的动态平衡引发更多关注，并形成了多元化主体参与的可信人工智能生态。

建议从加强要素协同、前瞻布局研究、健全标准体系、强化可信流程管理、推动产业交流合作等方面推动深入落地。协调制度、技术、人员整体推进；前瞻布局技术研究，将可信理念融入全流程管理；健全标准评估体系，系统性推进更多领域可信落地；强化产业交流合作，共同打造可信产业生态朋友圈。

可信人工智能正处于飞速发展阶段，我们的认识有待不断深化，报告存在不足之处，烦请不吝指正。

目 录

图目录

表目录

一、可信人工智能发展背景

（一）基本内涵

可信人工智能并不是一个新名词，随着可信实践的深入，产业界对可信人工智能的认识也在不断深化。目前关于可信人工智能概念解释繁多，我们梳理了一些具有代表性的观点。业界除了关注透明度、隐私保护、责任、公平等方面，近年更强调“人类对可信人工智能认识及素养的提升”以及“对提升人工智能可信度的可持续性工具的提出”。

表 1 主流可信人工智能概念梳理

国家/组织	时间	文件/项目	内涵
美国	2019.10	《人工智能原则：国防部应用人工智能伦理建议》	1）负责任；2）公平性；3）可追溯性；4）可靠性；5）可控性
	2021.07	“可信及负责任人工智能”项目	1）准确性；2）可解释性；3）隐私；4）可靠性；5）稳健性；6）安全性；7）减少有害偏见
	2021.11	《人工智能指南道德标准》	人工智能系统开发、测试和审查符合最高公平性、问责制和透明度标准
欧盟	2017.01	关于机器人民事法规则的欧洲委员会建议	1）自由；2）隐私；3）正值和尊严；4）自决和不歧视；5）个人数据保护
	2019.04	《人工智能道德准则》	1）受人类监管；2）技术的稳健性和安全性；3）隐私和数据管理；4）透明度；5）多样性、非歧视性和公平性；6）社会和环境福祉；7）问责制
日本	2017.02	《人工智能学会伦理指针》	1）对人类的贡献；2）遵守法律规则；3）尊重他人隐私权；4）公正性；5）安全性；6）诚信行为；7）对社会的责任

	2017.07	《AI 发展纲领》	开发者参照技术的特性在可能的范围内，应努力排除包含 AI 系统学习数据在内，所形成的一切偏见和其他不当差别对待的措施；开发者应遵守国际人权法、国际人道法，随时注意 AI 系统是否有不符合人类价值的行为
	2019.03	《以人为中心的人工智能社会原则》	1）以人为中心；2）教育应用；3）隐私保护；4）安全保障；5）公平竞争；6）公平；7）问责和透明；8）创新
加拿大	2018.12	《可靠的人工智能草案蒙特利尔宣言》	1）福祉；2）自主；3）正义；4）隐私；5）知识；6）民主；7）责任
澳大利亚	2019.11	《澳大利亚人工智能伦理框架》	1）人类、社会和环境福祉；2）以人为本的价值观；3）公平；4）隐私保护和安全；5）可靠性和安全性；6）透明度和可解释性；7）可争论性；8）问责性
新西兰	2020.03	《新西兰值得信赖的人工智能的指导原则》	1）公平和正义；2）可靠性、安全性和私密性；3）透明度；4）人类的监督和责任；5）福利
韩国	2020.12	《国家人工智能伦理标准》	1）人权保障；2）隐私保护；3）尊重多样性；4）禁止侵权；5）社会宣传；6）主体合作；7）数据管理；8）明确责任；9）确保安全；10）透明度
中国	2017.11	S36 香山会议	中国科学院何积丰院士首次提出可信人工智能，包含人、信息、物理三大要素
	2019.08	《人工智能行业自律公约》	1）安全可控；2）透明可释；3）；保护隐私；4）明确责任；5）多元包容
	2019.09	《新一代人工智能伦理规范》	1）增进人类福祉；2）促进公平公正；3）保护隐私安全；4）确保可控可信；5）强化责任担当；6）提升伦理素养
	2021.07	《可信人工智能白皮书》	可信人工智能是从技术和工程实践的角度，落实伦理治理要求，实现创新发展和风险治理的有效平衡。可信包含可靠可控、透明可释、数据保护、明确责任、多元包容 5 项可信要素

生命未来研究机构	2017.02	《阿西洛马 AI 原则》	安全性、故障透明性、司法透明性、责任、价值观的调和、人类价值观、个人隐私权、自由和隐私、共享利益、共享繁荣、人类控制、非破坏
IEEE	2017.03	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	1）人权；2）福祉；3）问责；4）透明；5）慎用
G20	2019.06	G20 人工智能原则	1）包容性增长、可持续发展及人类福祉；2）以人为本的价值观和公平；3）透明度和可解释性；4）健壮性、信息安全性和物理安全性；5）问责制
世界卫生组织	2021.06	《健康领域人工智能伦理与治理指南》	1）保护自主权；2）促进人类安全和福祉；3）确保透明度；4）促进问责制；5）确保公平；6）促进具有响应性和可持续性的工具
联合国教科文组织	2021.11	《人工智能伦理问题建议书》	1）相称性和不损害；2）安全和安保；3）公平和非歧视；4）可持续性；5）隐私权和数据保护；6）透明度和可解释性；7）责任和问责；8）人类的监督和决定；9）认识和素养

资料来源：根据公开资料整理

针对人工智能产业现状，切实结合当前发展需求，我们认为《可信人工智能白皮书》中对“可信人工智能”的表述更为贴切，因此本报告将沿用中国信通院所提出的内涵，即“‘可信’反映了人工智能系统、产品和服务在安全性、可靠性、可解释、可问责等一系列内在属性的可信赖程度。”基于这一内涵，企业、院校和各类机构等各类参与者，在人工智能产业各个环节中，秉持可信理念，将可信贯穿研发、生产、经营等内部全流程，落实于算力、算法、数据等产业核心要素中，打造出全面融合可信要素的人工智能产业，构建形成了可信的人工智能产业生态。

（二）发展意义

人工智能作为新一轮科技革命和产业变革的重要驱动力量，广泛应用于医疗、金融、交通等领域，带来了巨大的经济效益与社会效益。据 IDC 相关数据，2021 年全球人工智能产业规模为 3619 亿美元，并预计在 2022 年同比增长 19.6%，超过到 4300 亿美元。然而随着人工智能应用的深入，其自身的技术缺陷以及带来的决策偏见、使用安全等问题引发了信任危机，可信成为关注焦点。技术上，算法脆弱易受攻击带来的危险性；黑箱模型导致算法不透明，使得人们无法直观理解决策背后的原因。应用上，训练数据中存在的偏见歧视导致公平性缺失；以人脸识别技术为代表的生物识别信息的频繁使用增加了隐私泄露的可能。伦理上，人工智能系统决策复杂，难以界定责任主体，带来伦理安全问题。构建可信人工智能成为缓解和消除这些担忧的必然选择。

学术界率先推开了可信人工智能的大门，在可信人工智能概念提出后，逐步推广可信共识。2020 年起，可信人工智能领域研究论文数量飞速增长，围绕鲁棒性、可解释性、隐私保护等方面的技术研究持续升温。随着产业界开始落地实践，针对人工智能产业化过程中的可信探索与实践不断成熟，融合可信要素的人工智能产业生态开始兴起。

可信已经成为人工智能产业发展必备要素，驱动人工智能规范发展。人工智能只有可信可靠，才能获得可持续性发展。对用户而言，可信要素将推动人工智能技术黑箱趋于透明，增强用户对人工

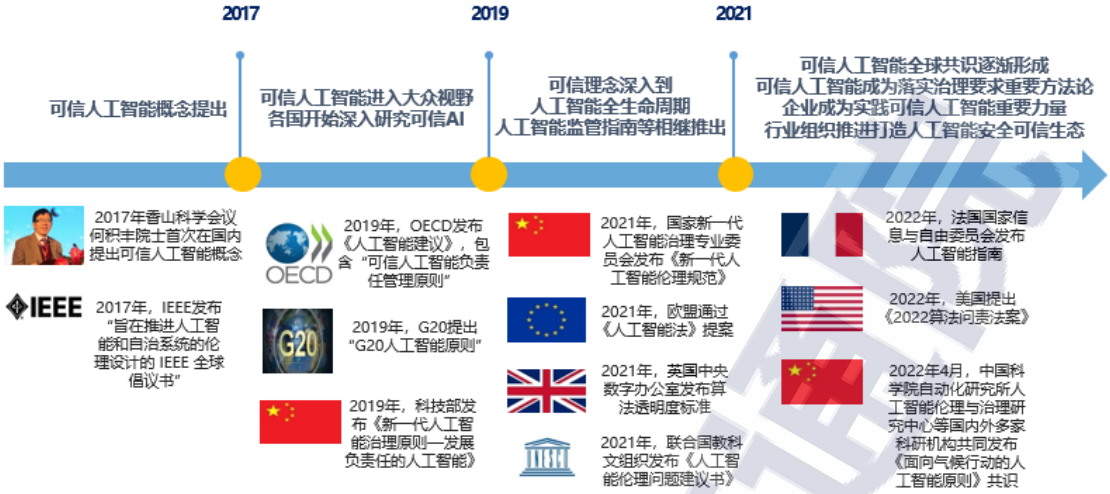
智能的信任感。对开发者而言，可解释的人工智能有助于全生命周期的企业管理，有助于履行内部报告和外部监管合规义务，确保应用和服务在最大程度上减少偏见。

可信作为传统领域数字化转型赋能因子，培育数字经济新兴增长点。随着人工智能产业对可信的探索与实践不断成熟，其代表的透明度高、可解释性强、确定性高的特性将与国防、法律、医疗等领域深度结合，缓解当前应用人工智能时所遇到的隐私保护、系统稳定等问题，以可信赋能数字化转型，甚至衍生出新的细分领域赛道，促进传统产业生态健康高质量发展，成为数字经济新的可持续增长点。

二、可信人工智能发展态势

（一）总体发展

可信人工智能处于快速发展中。2017 年可信人工智能的概念正式提出后，各国积极开展研究，随后可信理念深入到人工智能全生命周期，各项人工智能监管指南相继推出，2021 年后融合可信要素的人工智能产业生态开始构建。目前，企业已成为实践可信人工智能的主要力量，各大高校与行业组织也在积极同步推进打造人工智能可信生态环境。



资料来源：中国信息通信研究院整理

图 1 全球主要国家及组织可信人工智能发展动向

总体来看，可信人工智能的发展与各国人工智能产业基础紧密关联，随着人工智能产业化过程中可信要素的融入，企业日益成为可信人工智能实践主体。以微软、谷歌为代表的美国科技巨头通过内部开设人工智能伦理委员会，以及发布专注于人工智能可解释、公平性等伦理服务及工具，开展可信人工智能探索工作；以商汤、腾讯为代表的中国头部科技企业通过发布行业内人工智能可解释、伦理等领域专业报告，分享各自在可信 AI 技术、标准、服务等相关实践；以 SynSence 为代表的欧洲企业通过在自身产品中融合可信元素，以提高 AI 工具服务的可信度；以 GuardKnox 为代表的以色列企业致力于将安全融入到智慧交通领域，实现“零信任”；以三星为代表的日韩企业则注重人工智能行业应用中隐私安全的保障。

资料来源：中国信息通信研究院整理

图 2 全球可信人工智能产业生态实践图

在可信人工智能实践上，世界各国高度重视，在政策、技术、标准的研究制定上均采取了相关措施。**美国、欧盟、中国依托人工智能技术、人才、产业优势，在可信人工智能研究及治理实践上处于全球领先，成为全球可信人工智能领跑者；以日本、韩国、加拿大等为代表的人工智能第二发展梯队的追赶正在加速，试图通过构建人工智能伦理标准打造人工智能健康发展环境。**

政策发布上，全球持续探索人工智能立法，推动可信人工智能范式法制化。自 2021 年以来，从欧盟发布人工智能领域的第一份综合性法案《人工智能法案》，到美国推出《2022 算法问责法案》，再到中国深圳、上海等各地方相继推动人工智能立法条例。各国针对人工智能算法的监测、人工智能应用的审查的相关监管法规不断增加，人工智能治理已进入建章立制阶段。

技术研究上，提升人工智能系统稳定性、隐私保护技术占据可信人工智能技术研究主流，可解释性、公平性等技术研究紧随其后。当前以对抗训练、梯度屏蔽为代表的人工智能系统稳定性技术稳步发展，技术重点从数字域逐步向物理域扩展[1][2]，人工智能系统的稳定性测试技术成为科技巨头布局方向；以同态加密、多方安全计算、差分隐私等为代表的隐私安全技术发展迅速[3][4]，全球隐私计算专利数量迎来井喷；人工智能可解释性增强技术研究当前仍处于初期阶段，以谷歌、IBM、微软、腾讯为代表的科技巨头推出多个 AI 可解释性工具及服务；提升人工智能公平性主流方法分别从数据和技术两方面入手，通过构建完整异构数据集及引入公平决策量化指标算法，以减轻决策偏差。

标准研制上，行业组织成为可信人工智能标准重要推进者，涵盖多个可信人工智能领域。国际标准化组织和国际电工委员 ISO/IEC 布局最早，涉及 AI 系统偏差、风险管理、AI 系统质量模型、神经网络鲁棒性等；电气和电子工程师协会（IEEE）主要以隐私、可解释为突破点；中国人工智能产业发展联盟（AIIA）牵头制定了《可信 AI 操作指引》，成立人工智能治理与可信委员会，并持续开展可信 AI 测试工具征集和可信 AI 试评估等落地实践。自 2021 年起，中美两国加快了对可信人工智能领域标准研制工作的步伐，美国发布《人工智能风险管理框架概念文件》旨在降低人工智能系统风险；全国信标委人工智能分委会作为中国可信人工智能标准主

要研制单位，持续推动《人工智能 可信赖规范 第一部分：通用要求》等可信赖人工智能标准系列研究工作。

（二）政策规划

2017 年以来，全球各个国家（地区）在可信人工智能领域进行了广泛的政策部署，内容主要涉及伦理道德、隐私保护、负责任、公平性、安全性等层面，在实施路径与侧重点方面体现出了一定的特色与差异。

美国以政府和行业双轮驱动 AI 创新发展，维护人工智能领域的全球领导地位。美国以推动快速发展、降低创新门槛以及成本最小化为宗旨，政府和行业共同发力。一是制定基于性能的灵活性框架，权衡 AI 技术创新利弊，以适应 AI 应用程序的快速迭代和更新；二是设立行业准则、“安全港”、灵活监管、监管例外、监管豁免等内容，促进效益最大化的同时最小化潜在风险；三是制定可信 AI 标准指南为产业创新与发展提供详细路线图。

欧盟致力于构建 AI 信任生态系统与监管框架，确保成为数字化转型的领先者。一是以合法性、伦理性和鲁棒性为基准制定可信人工智能框架，提出了实现可信 AI 的方法与评价准则，为培育可信生态提供参照；二是通过发布《走向卓越与信任—欧盟人工智能监管新路径》，依托私营与公共投资的相互合作，为建立可信 AI 生态创造政策环境；三是规划制定《人工智能法》并定位全球最高标准，通过制订灵活完整的规则为可信应用生态提供法制保障。

我国协调发展与治理，凭借规范化治理确保人工智能可信赖。一是从规则、标准、评估、管控等层面的战略角度出发，协调并明确发展与治理的关系；二是通过成立新一代人工智能治理专业委员会，为人工智能治理框架和行动指南提供技术与规则支撑，积极引导全社会负责任地开展 AI 研发与应用活动；三是通过在个人信息保护、网络安全、数据安全等领域强化立法与执法力度，构建可信人工智能底层要素的坚固法律体系，进而确保各领域应用安全可靠。

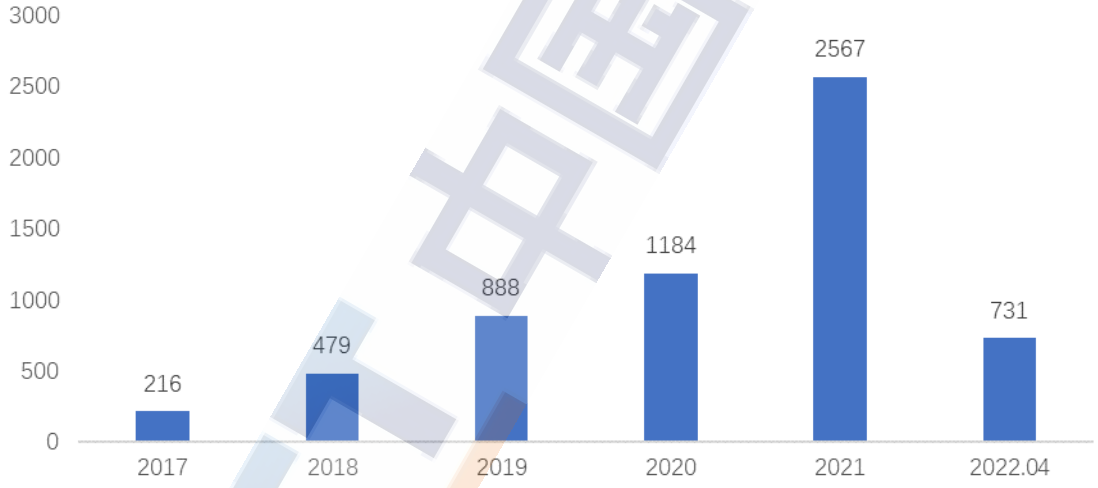
资料来源：根据公开资料整理

图 3 可信人工智能相关政策发展历程

（三）技术进展

当前对于可信人工智能技术聚焦在提升人工智能系统稳定性、可解释性、隐私保护、公平性等方面，这些技术构成了可信人工智能的基础支撑能力。我们结合当前产业与技术发展最新趋势及热点，对相关文献及专利检索分析，得出如下发现。

可信人工智能领域论文发表量在人工智能论文发表量占比逐步提升，接近 2%；美国、中国、英国是全球可信人工智能领域论文发表主要国家，占比超过全球 50%。可信人工智能逐渐进入研究者视野，各大科技巨头加速实践落地，可信人工智能领域论文数量迎来井喷，2021 年可信人工智能领域论文数量同比增长 116%，在全球人工智能领域论文数量中占比也由 2020 年的 0.8%提升到 1.7%左右。2003 年至 2022 年 4 月，全球可信人工智能领域相关论文数量论文共计 7059 篇。美国、中国、英国是可信人工智能领域论文发表的主要国家，三国发表的论文总数占全球论文总数 53%以上。



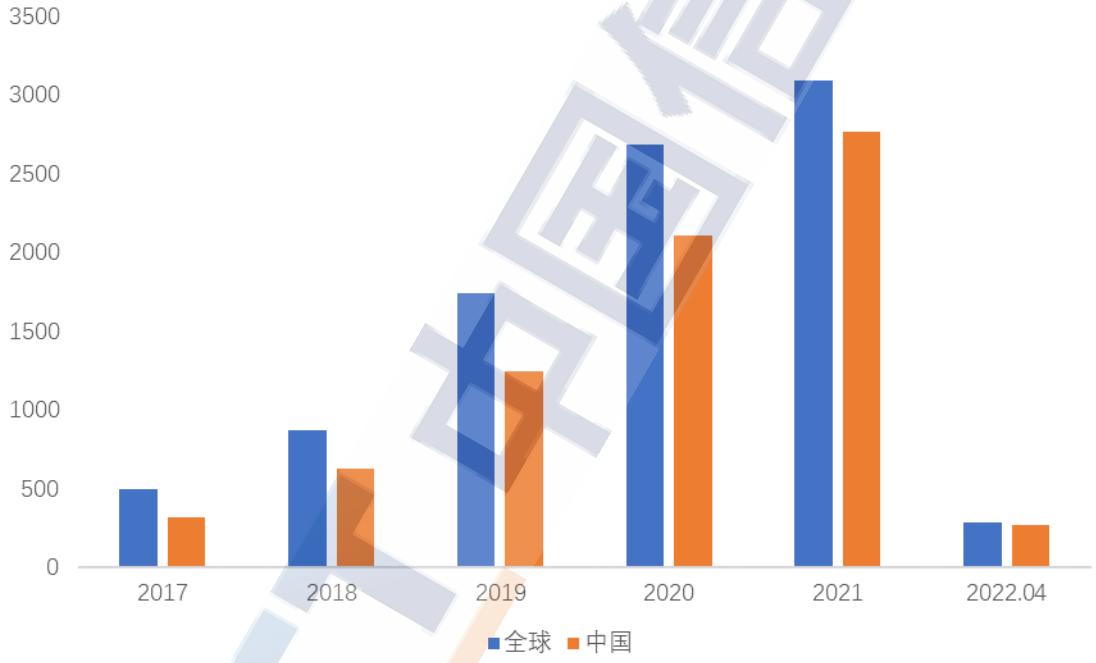
数据来源：Web of Science 官网

图 4 可信人工智能领域论文数量（2017-2022.4）[1]

可信人工智能领域，中国累计专利申请量与授权量居全球首位，在全球申请数占比持续增长。全球可信人工智能领域专利申请量快速增长，2017 年至 2022 年 4 月累计达 9174 件，我国申请量 7339 件，占比 80%。全球累计可信人工智能领域专利授权量达 1968

[1] 中国信息通信研究院根据 Web of Science 检索整理。

件，我国累计授权量 1432 件，占比达 73%。自 2018 年开始，全球可信人工智能相关专利申请数量迅速增长，这或许与全球各国开始重视并开展可信人工智能相关研究有关，其中 2022 年专利申请数量下降可能受部分专利申请流程存在滞后性的影响。我国在全球可信人工智能领域专利申请数占比处于持续增长，这可能与我国较强的人工智能技术基础、人工智能企业更加注重知识产权保护和我国大力倡导发展“可信赖人工智能”等因素有关。



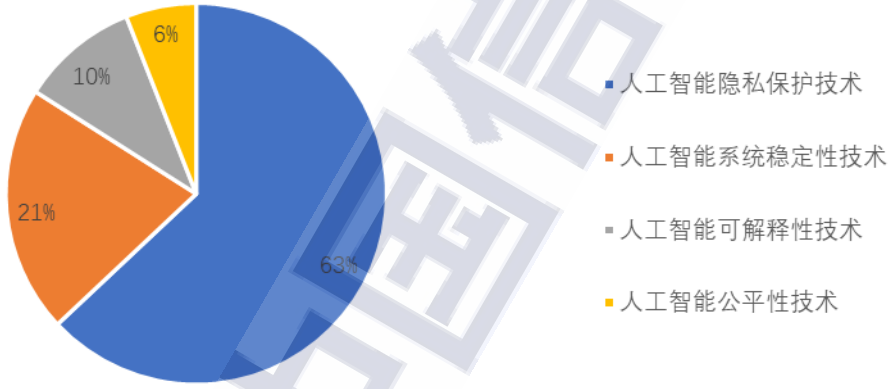
数据来源：incoPat

图 5 全球/中国可信人工智能领域专利申请量（2017-2022.4）[2]

专利技术分布上，针对人工智能隐私保护、提升人工智能系统稳定性领域专利占据主流，可解释性、公平性等领域专利稍显落后。从可信人工智能领域技术来看，主要集中于隐私保护方向与人工智能系统稳定性方向，隐私保护领域专利占比 63%，人工智能系

[2] 中国信息通信研究院根据 incoPat 检索整理。

统稳定性领域专利占比 21%，对于人工智能可解释性与公平性方向专利研究较少。这或许与可信人工智能产业应用成熟度有关，由于隐私保护与系统稳定性相关技术应用场景较为广泛，且成熟度高，因此这两个方向的专利申请呈集聚趋势；而人工智能可解释性与公平性由于相关实践应用落地不够，且相关技术研究仍处于初期阶段，因此针对这两个领域的专利申请相对较少。



数据来源：incoPat

图 6 可信人工智能领域专利申请量技术分布（2017-2022.4）[3]

（四）标准建设

自 2017 年以来，国际标准化组织与各国政府陆续布局可信人工智能标准。总体来说，当前对于可信人工智能的研究，主要涉及安全性、可靠性、公平性、透明性以及对人工智能的风险评估等。**综合各可信人工智能相关标准看，当前标准研究更多集中在隐私安全、伦理道德、风险评估，以及人工智能在金融、医疗等领域的可信应用；针对稳定性、透明度等领域，相关标准研究稍显不足。**

[3] 中国信息通信研究院根据 incoPat 检索整理。

表 2 主要国际组织可信人工智能标准进展

国家/组织	时间	标准	领域
ISO/IEC	2020.05	《信息技术 人工智能 人工智能可信度概述》	可信概念
	2021.03	《人工智能 神经网络鲁棒性评价 第 1 部分：概述》	稳定性
	2021.11	《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》	公平性
	在研	《信息技术 人工智能 风险管理》	风险管理
	在研	《信息技术 人工智能 伦理和社会关注概述》	伦理道德
	在研	《人工智能 神经网络鲁棒性评价 第 2 部分：常规方法选择方式》	稳定性
IEEE	在研	P7000 系列（15 项标准）	伦理道德、隐私安全、透明度、公平性、可解释
	在研	《可解释人工智能结构框架指南》	可解释
	在研	《自适应教学系统中人工智能伦理设计实践》	伦理道德
	2021.01	《金融服务可信数据和人工智能系统》	综合
ITU-T	在研	《技术报告：全同态加密技术为机器学习中的安全推理服务和数据聚合提供安全指导》	隐私安全
	在研	《技术报告：人工智能安全技术应用安全管理指南》	隐私安全

资料来源：中国信息通信研究院整理

国际标准组织以 ISO/IEC、IEEE 等为代表，在可信人工智能标准领域抢先布局。**国际标准化组织和国际电工委员** ISO/IEC 在可信人工智能领域布局最早，其下设的 WG3 工作组专门开展可信人工智能研究，并已布局 10 余项可信人工智能标准研究，涉及 AI 系统偏差、风险管理、AI 系统质量模型、神经网络鲁棒性等。**电气和电子工程师协会**（IEEE）主要以隐私、可解释为突破点，其下辖工作组开展了一系列人工智能伦理相关标准研究工作。其中，P7000 系列涉及伦理、透明度、隐私、安全机制等，是可信人工智能领域伦理

方面较为权威的标准，受到业界广泛关注。截至 2022 年 4 月，IEEE P7000 系列已至少开展 15 项标准研制工作。**国际电信联盟电信标准化部门**（ITU-T）下设多个人工智能工作组，其中 SG17 工作组将人工智能安全视为未来重要工作方向，因此致力于研究相关隐私安全标准，截至 2022 年 4 月，SG17 已有三份在研技术报告，涵盖机器学习安全应用、人工智能技术应用安全管理等方面。

美国、欧盟具备依托人工智能领域领先优势，加强在可信人工智能领域标准研究，构建可信人工智能生态。**美国国家标准和技术研究院**（NIST）隶属于美国商务部，在美国政府支持下进行人工智能研究工作。早在 2018 年 NIST 就启动了人工智能研究与标准基础项目，涉及人工智能系统安全性、可解释性、透明性等标准；2021 年 NIST 发布《人工智能风险管理框架概念文件》等报告，旨在建立可信赖负责任的人工智能框架，打造可信人工智能生态，维持美国在全球人工智能领域领导地位。**欧盟**从伦理向监管推进，抢占全球伦理规则主导权。从《人工智能道德准则草案》到《人工智能法》草案，欧盟希望通过抢占全球伦理规则主导权，构建“卓越生态系统”和“信任生态系统”，将欧洲建设成为全球人工智能研究和创新的“灯塔中心”。

中国科研机构及相关企业积极参与可信人工智能领域相关标准制定，涉及隐私安全、算法安全、风险管理、行业应用等多个方面，标准类型囊括国际标准、国家标准、行业标准等各类型，参与

者覆盖了中国信通院、电子标准院等各类研究机构，以及华为、百度、阿里等人工智能头部企业。

三、可信人工智能生态分析

可信已经在人工智能产业各个环节中落地，贯穿研发、生产、经营等内部全流程，打造出全面融合可信要素的人工智能产业，涵盖基础能力、理论技术、应用场景与产品设备等多元化模块。其中，**基础能力**提供基于不同业务需求的“最优可信设计”，构筑“安全、透明、可追踪、可计量”的数据体系，不断提升可信计算效能。**算法技术**在隐私保护、数据标注、精准识别、噪声处理等方面不断提升算法公平性、稳定性与可解释性水平[5][6][7]。**应用场景方面**目前已涌现出面向隐私安全保护、风险识别与控制、数据传输与共享等的诸多案例[8][9]，并已在金融、医药、教育、制造等领域实现商业化应用。**产品设备方面**已逐步渗透至医疗设备与器械、智能终端、智能驾驶、智能机器人、虚拟现实设备等领域，打造兼具稳定性、合规性、可解释性等特征的功能体系。

（一）基础能力

1. 平台系统

人工智能平台和系统通过搭建云边端协同模式，保证模型在不同的现实环境中都能部署和运营，帮助企业和开发者将精力聚焦在算法开发、模型验证和业务运营中，有效提升研发实施效率。近年来，随着机器学习和深度学习算法的深入实践，以及预训练模型、低代码/无代码接口、自助服务和生命周期自动化工具的快速发展，

与人工智能平台发展相互促进，更具有可信属性的人工智能平台系统迅速发展。

目前，**人工智能平台系统与可信理念的融合在数据处理、模型构建、部署和支撑服务等方面还面临不少挑战。**数据处理方面，数据接入、数据分析、数据管理和数据标注已经普遍实现，团队标注成为标配，数据预处理的自动化实现和无监督数据增强还需要持续探索。模型构建方面，虽然现有人工智能平台和系统普遍配置了丰富框架和算法，支持交互式、可视化、自动化多种开发模式和单机/分布式多种训练方式，但是模型评估建议能力和可解释性还需要增强，特别是面向文本、语音和视频场景的自动学习建模模版能力。此外，还需要提升基于规则的模型自动更新和模型在端云设备的协同部署能力，加强GPU虚拟化和池化，优化数据和模型的安全性。

资料来源：中国信息通信研究院整理

图 7 平台系统领域的典型可信应用

人工智能企业主要围绕业务生命周期，重点构建系列可信能力。**一是将内置模型保护于框架中，实现模型的安全、可信，**通过鲁棒性评测、对抗测评、对抗训练、模型加密等方法增强模型保护

能力，为人工智能模型安全性评估和增强提供支持。以对抗训练为例[10][11]，通过在输入上进行梯度上升，在参数上进行梯度下降，从而向增大损失的方向增加扰动。**二是搭建混合引擎架构，实现跨场景可信协同**，集成运用隐私评估、差分训练、联邦学习等多种技术方法[9][11][12][13][14][29]，通过数据安全交换协议有效利用多源数据，仅协同经过处理后的、不带有隐私信息的梯度和模型信息，在保证用户隐私数据保护的前提下实现跨场景协同。一些场景和平台中，也选择加入区块链技术实现全流程的可记录、可验证、可追溯、可审计，以证书授权实现双向认证，确保参与方身份真实性。**三是整合运用多种可解释技术，全面提升可解释性**，融合语义级可解释技术、可解释方法工具集等技术，建立适当可视化机制尝试评估和解释模型的中间状态，整合数据治理、资源管理和应用管理核心能力，大幅提高模型的可解释性，让用户更理解、信任并有效地使用模型。早在 2016 年起，谷歌、IBM、微软、腾讯等科技巨头就相继推出可解释性工具与服务，探索人工智能算法可解释化。**四是持续加强虚拟化和池化，提升运营维护和运营调度的可信能力**，将服务器等物理资源抽象成逻辑资源，通过区分优先次序并及时调度分配工作负载，让 CPU、内存、磁盘等硬件变成可以动态管理的“资源池”，并实现计算资源的隔离。

2.数据安全

数据是人工智能三大要素之一。人类把需要计算机识别和分辨的内容打上标签，让计算机不断地识别这些特征标签，从而让计算

机“学会”人类的理解和判断。只有经过大量的训练，人工智能算法才能总结出规律并顺利应用到新的样本上，因此，大量、多种类、标注精准的数据对人工智能训练效果极为重要。

可信人工智能在数据安全治理方面的应用主要集中在基于传统的人工方式难以处理的规模庞大、类型复杂的数据资产管理与分类分级，涉及安全、隐私计算、存证溯源、数据控制、计算处理等多种技术[15][16][17]。在《网络安全法》《数据安全法》《个人信息保护法》等一系列法律法规和政策文件出台后，金融、电信、工业等行业均已出台行业标准，形成**以法律法规和行业标准为引导加快推进可信进展的局面**。企业主要以训练样本合成、可信多位标签、存储治理追踪、智能巡检与兜底等方式，通过建设自动化的数据分类分级能力，确保分散在组织各处各层面的各类数据能够被及时发现和准确标注，实现**智能化的自动分类分级和安全保护**。

可信贯穿数据采集、标注、存储和巡检全过程。在**数据采集环节，以训练样本合成替代敏感数据采集。**由于部分数据涉及到身份证号码、住址等个人隐私信息，敏感程度很高，收集难度极大，因此在数据收集阶段，业内主要使用公开样本数据和自主合成样本的方法，一方面收集和使用相关公开赛事的数据样本，另一方面开发隐私数据训练样本自动合成算法来模拟真实样本数据。为解决数据量不足的困难，大多选择使用旋转、加噪等数据增强方法来扩充训练样本。**通过构建可信的多维标签体系，进行灵活的数据分类分级**，融合目标检测、光学字符识别、图像分类、人脸识别、文本校

验、风格识别等算法模型，结合多方信息联合判定校验，输出多维度标签，进一步提升隐私数据识别和治理的准确性，克服单一模型难以应对复杂的分级场景和不同治理需求、解释性较弱的问题。**严格执行分级加密存储和明暗水印追踪**，对于分类分级识别后的高敏感数据，通常需要经过加密后进行存储，基于区块链、数据沙箱等技术，实现长久稳固储存、全方位安全防护、安全共享；在文件分发流转过程中，添加对应的明暗水印以便数据泄露后展开追踪调查。**强化智能巡检与兜底，推动可信能力的持续建设与优化**，依赖安全运营专家人工验证需要耗费大量的人力，面对新增业务数据也很难及时发现异常，极易漏审、误审；通过智能巡检和兜底机制，将安全专家经验与机器学习能力相结合，并结合黑盒验证、红蓝演练等推动可信能力的持续建设与优化。

资料来源：中国信息通信研究院整理

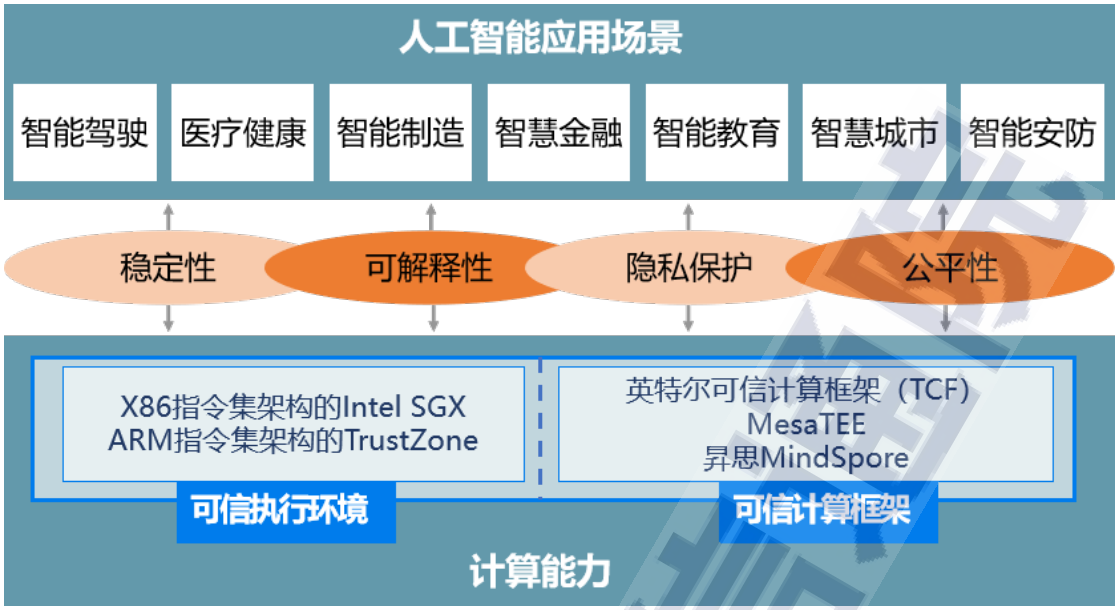
图 8 数据要素领域的典型可信应用

3. 计算能力

人工智能发展的关键要素是数据，而让数据发挥作用的关键则是算力。目前，各行各业都存在更多维度、更大深度的智能需求，

而在这背后需要更多的算力来为人工智能算法提供处理能力。算力已经成为助推经济发展的动力，据中国信通院测算，在算力中每投入1元，可带动3-4元经济产出，“算力正成为数字经济时代的重要驱动力”已成为共识。

在这种情况下，以人工智能芯片为代表的计算架构也承担了越来越多的敏感数据计算职能，对可信有极大的需求。例如，智能驾驶场景下，从感知系统进行数据收集，到车载计算平台实时处理，再到快速传输分析结果，每一个环节都离不开人工智能芯片的支持；由于智能驾驶场景较为复杂，一旦失效可能引发严重后果，对稳定性有很高要求。基于此，建立可信智能计算能力成为人工智能算力发展的一大趋势，欧盟委员会人工智能高级专家组（AI HLEG）于2019年和2021年发布的《可信人工智能政策投资建议》和《人工智能联合计划》就建议共同研发边缘端AI芯片，从基础软硬件方面支撑可信AI系统。



资料来源：中国信息通信研究院整理

图 9 计算能力领域的典型应用

可信执行环境（Trusted Execution Environment，TEE）**是人工智能芯片常见的可信技术**，是密码学与系统安全的结合，通过软硬件方法在中央处理器中构建安全区域，保护内部加载程序和数据的机密性与完整性，隔离的硬件设备提升了可信计算抵御攻击的能力，同时也可避免额外的通信过程以及公钥密码学中大量的计算开销。目前，主流 TEE 技术以 X86 指令集架构的 Intel SGX 和 ARM 指令集架构的 TrustZone 为代表，中国芯片厂商起步稍晚。**可信计算框架取得阶段性进展，构建系列可信能力**，英特尔可信计算框架（TCF）将密集运算和隐私数据处理工作转移至区块链下，以此解决区块链的可扩展性和隐私问题，并使用可信执行环境保证网络弹性和安全性；模型部署上，模型训练工具链迁移能力不足，对数据隔离条件下的训练、敏感数据的加密训练等都未能提供有效支持，需要通过开放模型生产工具链平台，向上对接训练数据协议、向下

具备规范模型安全、模型格式等能力。目前，全球首个内存安全的可信安全计算服务框架 MesaTEE 利用 Intel SGX 技术和 HMS 内存安全技术，兼顾云上数据代码完整性、保密性和内存安全带来的不可绕过性；全场景 AI 框架昇思 MindSpore 成为首个获得 CC EAL2+证书的人工智能框架。**通用的和可解释的智能芯片设计流程将成为下一个发展重点。**许多芯片设计为了适应不断变化的模型结构，需要在设计的通用性和效率之间寻求平衡，带来较大设计难度。未来，可信的智能芯片设计可能使用神经符号化方法构建可组合的模型，将芯片转变为由对应算法、具有可解释硬件模块的集合，从而在保证通用性和可解释性的前提下实现所需的性能。

（二）算法技术

1. 计算机视觉

计算机视觉技术囊括很多能够理解图像（包括图片和视频）的算法[23][27][28][30]，得益于深度学习技术的不断进步，计算机视觉在近些年飞速发展，在感知领域的研究已经相对成熟，内容合成与图像识别等某些人工智能任务已经能够通过图灵测试，并在金融、安防、制造等场景中落地，拥有一批相对成熟的产品应用。

在为生产生活带来便利的同时，**计算机视觉算法的应用也在隐私泄露、识别失效、偏见歧视等方面引发新关注。**例如，在商业零售领域对用户在不知情的情况下进行人脸识别和营销活动等产生了个人信息和隐私保护问题，造成了恶劣的社会影响。面具仿冒、对抗样本攻击可能造成识别失效，此外，也可能涉及到针对不同人

种、老年人的偏见歧视等伦理问题，美国一些零售店和警务工作中使用的面部识别技术会错误识别黑人，引发了抗议和监管。

资料来源：中国信息通信研究院整理

图 10 计算机视觉领域典型可信应用

为了增强计算机视觉算法的可信能力，产业界进行了多种尝试，不断提升计算机视觉算法可信水平，为重要产品提供核心能力支撑。**一是通过联合解译和认知推理深入理解场景或事件，增强可解释性**，人造物体和场景设计中暗含了潜在的、未以像素表示的实体和关系（近似于人类的常识），通过推理这些可见像素以外的不可见因素，使用有限的数据来实现各种任务的泛化[17][18][19][26]，形成“以小数据驱动大任务”的新型范式。**二是采用数学可证明的形式，融合不同形态的噪声进行改造**，以掩码等方法使其满足不可逆、可撤销、不可关联等特性，提升模型鲁棒性，避免样本不均衡，实现安全、可信且准确率高的识别。三是对于已有的生物识别系统，可以**应用安全多方计算和同态加密等技术手段**，在生物特征的密文状态下进行计算，并将最终结果恢复成明文，有效保护原始生物特征的安全。在实际应用中，同态加密与安全多方计算经常结

合使用，在金融领域反洗钱和跨实体欺诈分析、抗击新冠疫情敏感健康数据等场景下得到应用。**四是形成行业合力，推进与不同风险场景、主体结合的分级分类标准建设**，例如，上海在全国率先立项人脸识别地方标准《公共场所人脸识别分级分类应用规范》，积极探索使用主体和实施主体对公共场所人脸识别系统的分级分类应用原则，并提出相应的评估方法；中国信通院发起成立“可信人脸应用守护计划”，联合多方力量，通过标准制定、测试评估和行业自律等手段，共同规范人脸应用健康发展。

2. 智能语音

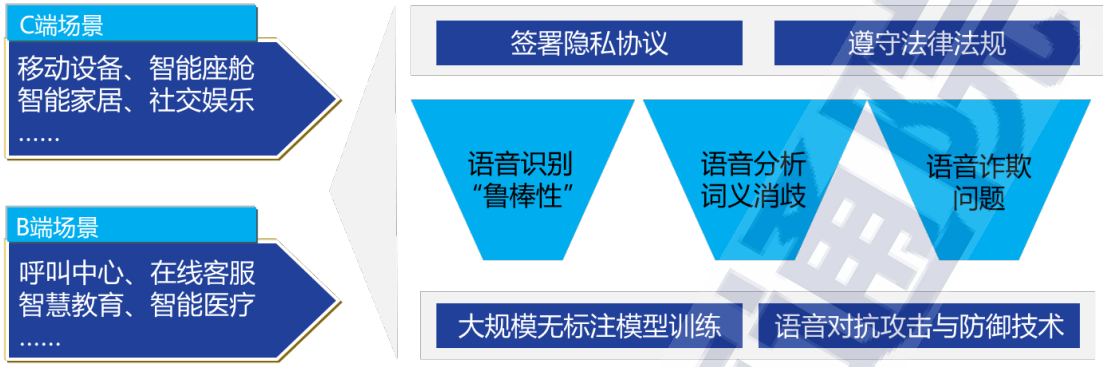
智能语音技术是人工智能主要算法技术之一，从最初只能识别孤立的数字以及有限的词汇，逐步发展到通过声音模式和特征设置参数实现基于大量词汇的连续语音识别，再到基于概率统计建模、基于深度学习的识别。目前，智能语音算法已经逐步成熟，在移动设备、汽车、家居等C端场景以及呼叫中心、在线客服等B端场景深入赋能，并逐渐渗透到安防、旅游、法律等行业中。

语音识别稳定性、语音诈欺等问题成为焦点。语音识别的鲁棒性问题显著，业内普遍宣称的高准确率，更多是在安静室内近场识别中实现的，在真实使用场景中则要考虑远场、方言、噪音、断句等问题，准确率会大打折扣。语言往往一词多义，语音分析目前主要是浅层处理，词义消歧依然是瓶颈，想要让机器像人一样运用知识储备结合上下文进行理解和交互，还需要更多的探索。此外，随

着各种语音合成工具的普及，语音合成滥用等语音欺诈问题也随之而来，成为产业发展关注的焦点问题。

为解决以上问题，企业先后开展探索，积累了一批成功经验。**一是向用户充分告知相关风险，通过签署协议等方式对数据的采集和使用进行限制，**允许用户对何时和如何上传使用语音信息做出选择，例如设置手动关闭语音采集装置的开关、允许不自动升级、只有在用户允许的情况下进行语音的采集和识别。**二是在数据存储、处理和删除等方面遵循各类法律法规的规定，推进技术向善**，严格按照个人信息保护法、数据安全法、网络安全法、个人信息安全规范等国内规章制度以及欧盟 GDPR、美国 CCPA 等条例的规定开展业务。例如，《网络音视频信息服务管理规定》对基于新技术新应用制作、发布、传播音视频信息明确了安全评估、标识、信息管理、辟谣等方面的要求，避免利用智能语音技术侵害他人合法权益。**三是尝试基于大规模无标注训练**，语音领域的无监督学习成功案例较少，有监督训练成本较高，因此，半监督学习获得广泛关注，即在海量无标签数据上训练大规模通用预训练模型，并对少量有标签数据进行精细调整，从而更好地强化训练效果，提升可信能力。这种利用少量带标注数据来训练大量无标注数据的超大规模自监督学习技术已展现出很强的通用学习能力。**四是语音对抗攻击与防御技术获得更多关注**，语音领域的对抗攻击将从当前的白盒攻击，进□步进化成黑盒攻击，攻击内容将从当前流行的 untarget 攻击进化成 target 攻击。基于这些尝试，智能语音正在从以往基于语

音交互的智能辅助工具形态进化为基于虚拟人多模态交互的智能助手形态。



资料来源：中国信息通信研究院整理

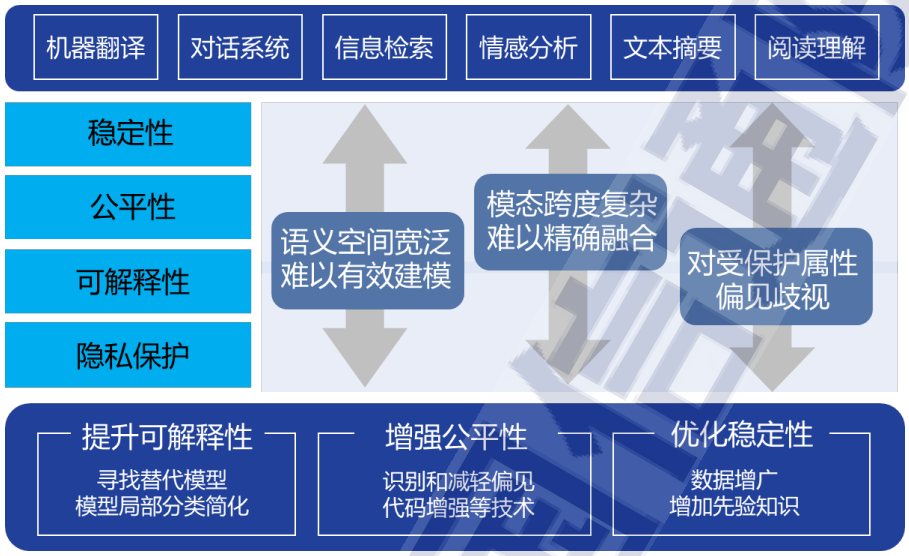
图 11 智能语音领域的典型可信应用

3. 自然语言处理

自然语言处理是人工智能从感知迈向认知的关键技术，近年来，自然语言处理加快与知识图谱等的融合，推动机器翻译、对话系统、阅读理解等技术在特定任务上超越人类水平，从而持续提升在搜索引擎、对话交互、个性推荐等场景的性能，并随着全球各地协同发展和语言文化交流融合而不断培育出新的需求。

自然语言处理在发展过程中还存在不少技术挑战，其中之一便是很难获取到大量高质量的标注数据。自然语言处理领域认知类任务较多，数据标注的时间成本和人力成本相较与语音识别、图像处理等感知类任务更大。**从可信人工智能的视角来看，自然语言处理发展仍存在建模、融合、消除歧视等挑战。**一是自然语言语义空间宽泛，依赖特定语料学习，通用性和迁移性存在瓶颈，难以有效建模；二是自然语言涉及多种模态，跨模态关系抽取、语义理解等技术还缺乏深层结构分析，难以精准融合；三是大型预训练模型可能

导致对性别、种族和年龄等受保护属性的偏见，例如一项研究表明，通过某特定报纸和一家面向老年人的报纸训练的自然语言处理结果脱离了年轻人和女性的交流方式。



资料来源：中国信息通信研究院整理

图 12 自然语言处理领域的典型可信应用

目前，自然语言处理领域中已经开展了不少可信相关工作。**可解释性方面，采用特征重要性、替代模型、样例驱动、溯源、陈述归纳等机制提供解释**。例如，构建基于 probing task 测试模型的语义理解能力，寻找简单的替代模型或者将模型的局部分类面简化；理论分析 RNN 的泛化性[18]；运用一阶导数显著性衡量每个输入单元对最终决策的贡献量；注意力机制模拟人类理解语言时会集中注意到一些关键词的行为，在一系列任务上显著提升模型性能[19]。此外，还可以模仿人类解决问题的过程进行可解释的结构设计，由于该架构包含模拟人类认知的组件，学习到的模型（部分）可解释。**公平性方面，扩大和丰富样本数据来源，并对样本进行偏移，识别和减**

轻偏见，扩大自然语言样本来源范围和种类，除了常见的社交媒体、报纸等数据来源，将样本来源进一步扩大到维基百科、城市字典(Urban Dictionary)，甚至圣经和古兰经的解释及公开杂志[20]。在预训练模型使用的各个阶段尽可能消除偏见[21]，一些算法可以在保持有用信息的同时，修改实际矢量以删除定型信息，实现对模型的偏移，从而改善公平性。**可靠性方面，小样本学习性能和鲁棒性提升，同时出现了很多能够可靠评估模型的方法，性能指标逐渐由单一转向多元。**通过数据增广、增加先验知识等实现小样本学习性能和鲁棒性的提升，有效避免故障模式的发生。以欺诈短信分类识别场景为例，通过对公开的短信数据进行改造，对样本量少的类别进行过采样操作，以及综合使用 Macro-F1 等多分类的评价函数、FastText 算法等技术，能够有效提升模型的稳定性[22]。随着近年来自然语言处理模型的快速改进，基准度量能力也在不断提升，动态对抗性评估、社区驱动型评估、跨多种错误类型的交互式细粒度评估、超越单一性能指标评估模型的多维评估等新基准逐渐出现。

（三）应用场景

1. 智慧金融

人工智能驱动金融服务应用程序已经成为金融创新的一大趋势，在金融产品设计、市场营销、风险管理、客户服务和其它支持性活动等金融行业主要业务链条均有落地，生物特征识别、知识图谱、智能语音等技术已经衍生出智能营销、智能客服等典型场景。

人工智能赋能金融业的深度逐渐加深，金融系统复杂性日益提升，对金融市场的公平性、透明度和稳定性提出了新的挑战。依托可信人工智能适当和透明的设计能够合理规避风险，增进消费者保护和信任。一方面，通过解释人工智能算法如何开发与运作，能够促进建立对人工智能应用的信任，整体提升金融市场的可信度；另一方面，减少人工智能结果的偏见，避免产生歧视性结果及市场趋同和羊群行为，从而确保市场的稳定性和可复现性。

资料来源：中国信息通信研究院整理

图 13 智慧金融领域的典型可信需求与实践

可信人工智能在金融领域的实践主要集中于业务运营、风险管控、销售营销等环节。**在业务运营环节，匿踪私密查询等技术能够保护用户隐私，防止用户信息被误用滥用。**金融机构在联合运营商、政务机构、保险机构、支付机构等获取相关客户信息解决日常业务的过程中，使用匿踪私密查询技术对查询方的客户身份ID进行混淆加密，使数据源方无法确切知道客户信息，有效避免客户身份ID信息泄漏。通过不可更改的固件安全启动、硬件加密算法保证的

加密信任链以及离线签名的加解密算法，保证系统运行环境的无人为篡改和可信安全性。**在风险管控环节，通过联合计算与统计解决风险信息不对称问题，丰富目标主体风险评价信息维度，提升金融服务的普惠性和公平性。**通过安全求交等技术[24][25]，横纵向打通不同领域数据，从而实现金融机构风控信息的全面补充，形成跨行业线上线下真实经营数据信息的评估链条，交换机器学习参数，提升对目标主体信贷风险洞察能力，实现本行优质高潜客户挖掘，有效缓解中小银行客户相对资质稍差的问题。**支付服务环节，以注入式攻击防御方案增强支付系统安全稳定性。**针对刷脸支付等新型支付潮流和潜在的注入式攻击威胁，金融机构主要采用整套防控方案，既包含硬件环境检测+传感器参数+图像视觉特征融合的多模态防控方案+TEE+暗水印，又融合隐式多帧、微表情等算法创新，可以防御99.9%的注入式攻击。

2. 智慧医疗

人工智能作为一种“通用”技术，几乎渗透医疗系统所有领域，从临床决策到生物医学研究和卫生系统管理等，将临床医生从非临床事务中解放出来，在影像诊断、药物发现、肿瘤诊治等行医过程中甚至具有更高的效率。

如何在人工智能可信框架内规范智慧医疗，高效发挥人工智能技术作用至关重要。可信治理框架强调透明度、信任的重要性，可信人工智能在医疗领域的推广能够在保护数据隐私的同时打通数据群岛、实现数据价值、提升医疗算法的可解释性，从而提高社会整

体医疗水平与效率。目前，医疗行业内存在数据共享难、分析难的困境，并伴随着医疗数据异构、类型复杂等问题。

资料来源：中国信息通信研究院整理

图 14 智慧医疗领域的典型可信需求与实践

可信人工智能在医疗领域的运用主要体现在医疗辅助诊断、药物发现与罕见病治疗方面。**医疗辅助诊断方面，可信人工智能能够实现高质量医疗服务的标准化，缓解地区医疗水平发展不平衡的问题。**通过数字化、标准化的专家经验和知识图谱，可将高质量医疗服务复制并输出，增加医疗资源的总体供给，快速提升基层医院的医疗水平，使得患者无论是在发达地区或是偏远地区，均可就近就医，享受到基本同质的医疗服务，促进医疗卫生资源均衡化发展。**药物研发和罕见病研究方面，以联邦学习协同各机构合作，提高精度和成功率，保护病人隐私。**单个机构罕见病数据量偏少且高度有偏，共享存在高壁垒、高成本、高机密性等困难，联邦学习框架下，机构之间仅通过共享模型权重即可协同训练，彼此增强模型效果，还可通过蒸馏学习解决参与聚合的模型参数量过大的问题，维

护系统稳定性。此外，**将随机对照试验作为检验医疗人工智能创新的金标准，保证泛化性和可靠性。**随机对照试验（RCT）通常被认为是医学临床试验的黄金标准，2021 年 Nature Review Cancer 的一篇研究显示，3578 项与深度学习相关的癌症诊断技术研究中，符合三期临床 RCT 标准的仅有 3 项。目前，结肠镜计算机辅助诊断、冠脉 CTA 等领域已经开展了符合医学标准的科学验证，从而保证医疗人工智能的准确度和可靠性。

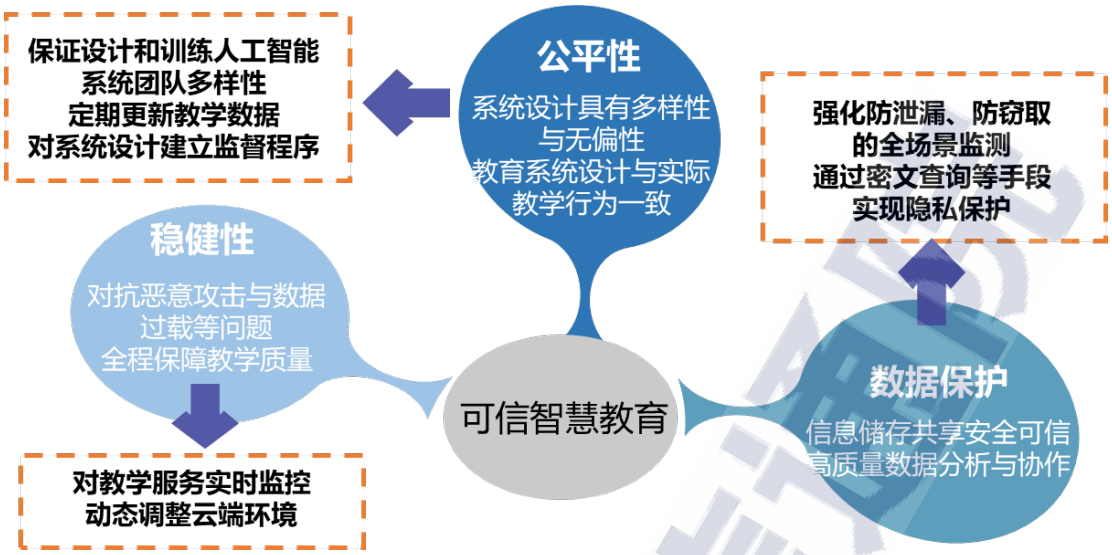
3. 智慧教育

智慧教育在发展过程中已催生出大量应用和新业态，覆盖教育各个环节，依托知识图谱、情感计算、自然语言处理等技术，通过计算机辅助教学、学习管理系统、自适应学习等方式，全面促进教育机会多样性，内容丰富性，方式灵活性与途径便捷性。

教育的社会性与敏感性意味着人工智能带来的不确定性和风险更加显著。可信人工智能在公平，稳健、保护隐私的前提下充分发挥人工智能在教育领域的潜力。一方面，降低系统复杂性，给出选择特定学习轨迹的过程与理由，避免信息不对称及歧视性结果，确保包容、公平的优质教育。另一方面，在保护用户个人数据的前提下促进因材施教和个性化学习，避免个人信息滥用引发道德风险。

可信人工智能在教育领域的实践主要体现在确保公平性、增强系统稳健性、完善数据安全保障三方面。**可信智慧教育的公平性体现在系统设计的多样性与无偏性。**一方面，保证设计和训练人工智能系统团队多样性，如在专业领域、教育背景方面，并定期更新教

学数据避免系统固化；另一方面，对系统设计建立监督程序，明确分析系统的目的、操作范围、限制和要求等，评估可能出现不公平偏见的情况，并保证人工智能教育系统态度、诊断和教学行为的一致性。**可信智慧教育能够确保系统的稳健型，对抗恶意攻击与数据过载等问题，全程保障教学质量。**一方面，建立稳健教学服务引擎，对教学服务实时监控，并设立预警通知机制，保障监测到警告信息后及时处理，以维持底层服务的稳定性；另一方面，通过对各个关键业务点的智能模拟访问，保障教育系统实时可用，并依据运行状态数据分析，动态调整云端环境的配置和数量，提升业务高峰期的运行稳定性。**可信智慧教育将教育大数据的信息储存和共享变得安全、可信。**一方面，强化防泄漏、防窃取的全场景监测、预警和应急处置等能力建设，并建立起数据流动的异常行为画像机制，对可能存在的风险部署人工智能检测模型，将数据的时效性和检测行为融为一体；另一方面，建立可信执行环境，通过密文查询手段防止工作人员查询敏感数据，保护学生隐私。同时，提升区域教育大数据含金量，实现隐私保护下的高质量数据协作，利用上链数据实现事中数据流通可控，事后可溯源监管、审计。



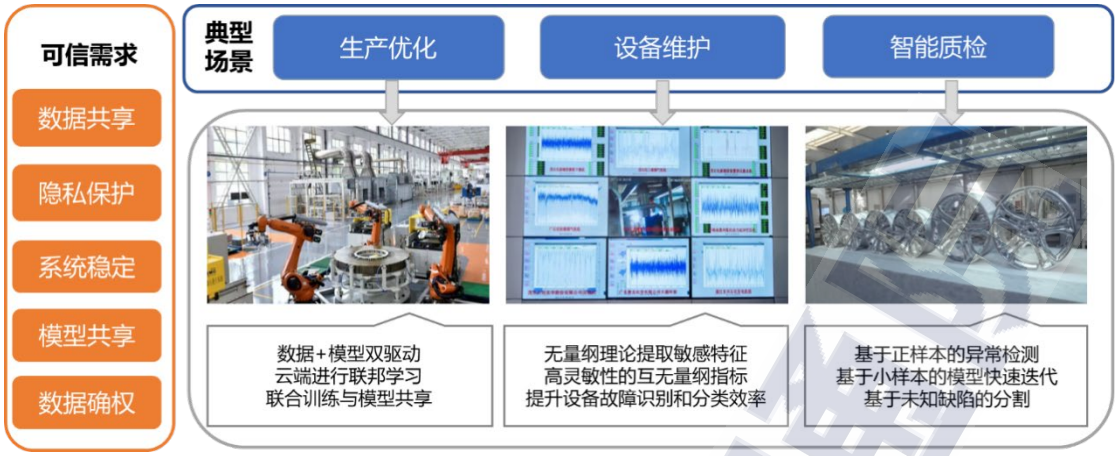
资料来源：中国信息通信研究院整理

图 15 智慧教育领域的典型可信需求与实践

4. 智能制造

在机器视觉、语音技术、机器学习等技术助力下，人工智能在工业领域的应用推动了生产优化、智能质检、生产设备预测性维护以及供应链管理等制造体系的全局决策优化，推动产业提质增效。

目前，人工智能在工业领域的实践存在工业数据流通不畅、信息泄露，数据价值融合、释放困难，数据涉及主体众多、无法确权等问题。可信人工智能试图寻找一种适合工业场景的新型数据共享流通解决方案，在数据价值释放及数据安全需求中取得平衡，保障工业数据合规流动与有序利用，有助于构建数据驱动、人机协同、跨界融合，共创分享的智能工业经济形态。



资料来源：中国信息通信研究院整理

图 16 智能制造领域的典型可信需求与实践

可信人工智能在工业领域的实践主要体现于生产优化、智能质检与生产设备预测性维护等方面。**可信人工智能能够避免信息孤岛带来的效率低下、工艺水平参差、重复性工作等问题，促进生产优化。**通过建立以数据+模型双驱动的数据共享方案，将样本数据传输至云端进行联邦学习实现多条产线间的联合训练与模型共享，并通过数据清洗与储存、数据计算处理等步骤沉淀量化行业知识，构建高品质、高效率的生产新模式。**通过无量纲和互无量纲指标实现生产设备的预测性维护，降低随机故障几率。**原始振动与特征信号通常隐藏了大量设备状态信息，能够反映设备系统状态与变化规律，通过对该信号的提取与分析，能够有效进行设备故障诊断。一些企业利用无量纲理论提取敏感特征，从而避免特征冗余，提升设备故障识别和分类效率；在一些对精细度要求较高的领域，可以利用实际模拟仿真，推导出具有更高灵敏性的互无量纲指标，适用于环境和设备结构复杂的情况。**使用基于正样本的异常检测、基于小样本的模型快速迭代和基于未知缺陷的分割等方法，优化智能质检。**人

工智能质检在汽车制造、电子制造、新能源等多个工业细分领域均有落地，目前主要采用基于正样本的异常检测、基于小样本的模型快速迭代和基于未知缺陷的分割等方法，解决较为常见的换型效率与模型精度要求高、难以使用一个通用模型覆盖所有应用场景、样本数量不足以满足训练需求、数据质量参差不齐等问题，优化问题工件预测模型，进行全量自动化质检。

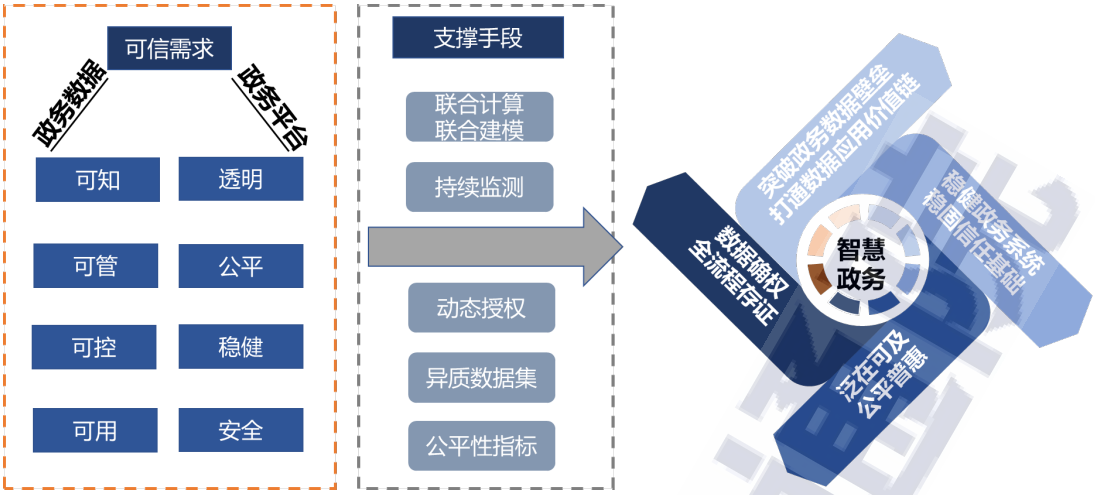
5.智慧政务

人工智能融合知识图谱、语义分析、文字识别、语音识别等技术，已经应用到政府办公、信息管理和公共服务等多个场景中，助力政务决策、业务流程优化，提升利企便民服务体验，提高城市政务服务能力与水平。

智慧政务在助力政府信息化建设和实现政务应用无障碍的过程中，数据共享、隐私安全与协同效率等受到挑战。可信人工智能帮助促进政府信息在各部门间及时交换和广泛共享，为跨部门协同与领导决策提供准确可靠的数据支撑，全面保证数据的可知、可管、可控、可用，并为公众建立一个更为透明、公平、功能强大的公众服务平台，提高政府的公共服务管理水平。目前，智慧政务主要面临横、纵向信息交换存在“数据壁垒”，政务协同缺乏信任基础等问题。

可信人工智能在政务领域的实践主要体现在实现跨域数据共享，加强政务系统可靠性与实现公平普惠。**可信人工智能能够突破政务跨域数据壁垒，打通政务跨域数据应用价值链。**建立非人为控

制的信任系统，通过联合计算、联合建模等实现政务数据协同计算与数据资源合规市场化、安全应用化、价值最大化，为跨部门跨层级数据互联互通提供安全、可信环境，并实现政务信息的全流程存证与数据确权，面向数据权属、采集、存储、存证、计算，提供全生命周期安全保障。提高异构应用访问的适应性，实现外部公共数据分布式访问，进一步提升公共数据融合和流通服务质量及效率，形成社会效益与经济效益双发展。**可信人工智能能够构建稳健政务系统，稳固信任基础。**政府层面，通过人工智能对抗技术，帮助政府解决AI应用中的安全问题。定期检测AI模型安全漏洞，防御攻击AI系统的行为，提升政务系统安全性，保证政府业务的全流程安全访问，并防止政务敏感信息截屏、转发、复制等。用户层面，持续监测终端设备和用户的安全风险，根据授权主体、客体环境和行为风险进行动态授权。通过检测鉴别AI伪造内容，降低因滥用AI导致的社会隐私和伦理风险。**可信人工智能通过确保系统决策的公平性，帮助构建泛在可及、公平普惠的公共服务信息体系。**一方面，构建完整的异质数据集，避免因文化、政策或历史因素所造成的社会偏见，如针对少数和弱势群体，在设计理念与实际应用中考虑特殊需求，推出相应的页面与符合群体特征的应用机制。另一方面，定期检查数据集以确保数据的高质量，使用公平性指标来减轻或消除偏见和潜在歧视。



资料来源：中国信息通信研究院整理

图 17 智慧政务领域的典型可信需求与实践

（四）产品设备

1. 医疗设备与器械

医疗设备和器械在临床检查治疗中占有重要地位，是保证医院正常医疗研究教学工作的必要条件。发达国家和地区起步较早，产品水平总体领先，通用电气、飞利浦、西门子等企业垄断了医疗器械行业的主要份额，尽管我国企业也紧抓机遇快速发展，但中高端诊疗设备仍主要依赖进口。

医疗设备和器械的可信能力是监管机构重点关注的能力之一，国家药监局器审中心《医疗器械网络安全注册审查指导原则（2022 年修订版）》要求医疗器械对于网络安全威胁应具备必要的识别、保护能力和适当的探测、响应、恢复能力，保密性、完整性等方面的风险在全生命周期应处于可接受水平。**未经授权的泄密已经成为医疗行业数据安全的首要风险**，多类型、多型号的IoMT设备分布在多科室，且设备厂商的远程运维方式多样，这导致了风险暴露面积

的增加，原有安全防护手段难以应对。**医用级智能可穿戴设备可能将代表性不足的群体排除在医学研究之外**，美国国立卫生研究院的一项研究表明，由于设备价格太高，大多数使用智能手表和其他可以追踪健康状况的可穿戴设备的人都是受过良好教育的富有白人，低收入群体和少数种族群体被排除在使用可穿戴数据的研究之外。

资料来源：中国信息通信研究院整理

图 18 医疗设备与器械领域的典型可信需求与实践

通过政府、行业和企业的共同努力，医疗设备与器械领域的可信实践正在加快落地。实践中，**医疗设备和器械行业需要遵循法律法规和相关标准的要求**，其中既有《网络安全法》《数据安全法》等法律法规，也有相关标准予以规范。此外，监督管理部门根据申请，按照相应的法定程序，对拟上市医疗器械进行系统评价，其中安全性是重要的指标之一。从产品角度看，由于可解释性差、医疗数据无法完全代表临床决策要素等原因，企业结合应用场景需求和产品特点，运用多种不同的技术手段提升产品可信能力。**智能医疗设备领域**，强化自动精准校正和全局轨迹规划，全面保障设备运行，提升系统稳定性；采用多维神经网络进行分布渐进式学习优化，兼顾噪声、对比度、分辨率，有效保证成像和影像识别结果的

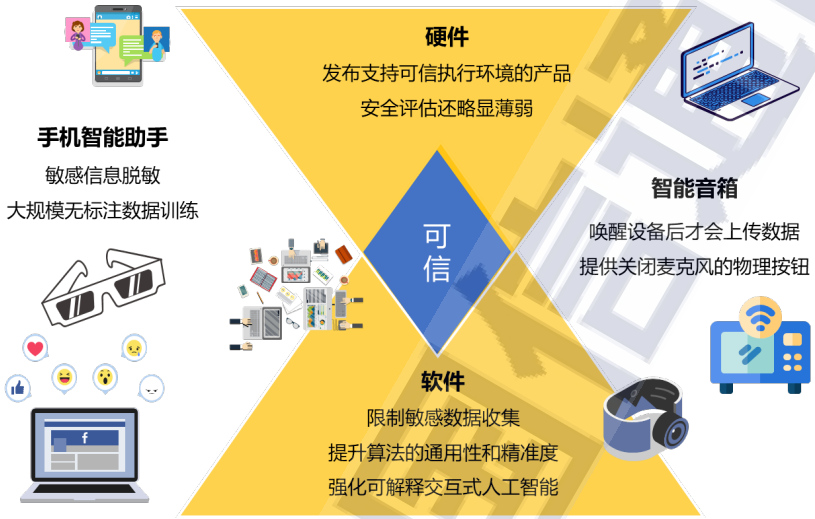
可靠性；聚焦患者关怀，为妇幼胎儿、肥胖患者、运动医学检查精准诊断与个性化治疗提供针对性支撑产品。**胃镜等器械领域**，将知识图谱、计算机视觉等技术融入产品设计，使用诊疗规范指导产品研发，从而增强系统可解释性；精准分类、快速分离，在不增加检查时长的前提下，有效降低漏检率，最大限度地减少不必要的活检，显著改善患者就诊体验；将人工智能与 5G 结合，支持远程检查，增强高水平医疗服务的公平普惠。**医用可穿戴设备方面**，产品设计时考虑到肥胖、重症、特殊疾病的病人需求，并推出不同价格层次的产品，让更多患者享受到优质服务；以专业医生临床诊断需求为标杆，对标医院场景下的数据准确度，提供医用级数据监测，利用循证人工智能算法帮助患者解决危急情况预警等痛点，指导用户的行为、用药和生活方式。

2. 智能终端

智能终端覆盖了智能手机、PAD、智能音箱、智能车载终端等多种类别，几乎是数字化、智能化场景中都不可缺少的构件。受益于人均收入增加、城镇化水平提升和老龄化进展，智能终端相关产品需求将得到持续不断的拉动，帮助提升社会资源的普惠性、均衡性分布，增进民众体验感和满意度。

作为数字经济的关键入口和主要创新平台，**智能终端在提供更多便利的同时也面临着严峻考验，特别是在稳定性和隐私保护方面**。为了实现更多功能的集成，必然要求智能终端增加交互接口，这也意味着可能被攻击的入口数量增加。据媒体报道，来自英国和

意大利的研究团队远程黑入某智能音箱，让智能音箱给自己下达恶意指令，平均成功率达88%。同时，为了提供个性化服务，智能终端需要大量收集用户隐私及交互数据，用户数据和隐私环境被侵害的可能大大增加。



资料来源：中国信息通信研究院整理

图19 智能终端领域的典型可信需求与实践

沙箱技术、身份鉴别、访问控制等传统安全手段已经不再适用余智能终端新的可信需求。近几年来，**针对硬件的可信执行环境（TEE）逐渐流行**，基于指纹等生物特征支付成为主流支付手段后，智能终端厂商纷纷将TEE技术作为保障终端支付安全的基础平台，发布支持TEE的产品，苹果、高通等企业培育和发展出从底层硬件到上层软件的TEE整体方案。然而，由于TEE组件数量多、组件交互复杂，涉及的底层技术众多，目前国内对这类可信智能终端的安全评估还略显薄弱，可能成为未来重点发展的方向。软件方面，基于深度学习模型产生的各类智能化产品通用性较差，终端厂商主要**限制敏感数据收集，提升算法的通用性和精准度，强化可解**

释交互式人工智能来提升可能能力，一是通过严格限制非必要的数据采集、敏感信息脱敏、不强制升级、用户数据收集协议等方式强化可信能力；二是参考可信人工智能的相关准则及框架进行系统涉及开发，微软、谷歌、IBM等企业已将自身提出的可信理念贯彻于产品研发设计之中；三是强化算法模型的可解释性，在复杂度、透明度等维度上寻求平衡，一般选用自解释模型或引入注意力机制、深化统计模型、基于物理模型等构建具有内置可解释性的事前解释模型，也可运用激活最大化、概念激活矢量测试、知识蒸馏等事后解释模型或算法。三是将可解释性纳入前期用户研究考察维度，针对智能终端使用者所希望了解的解释内容以及相关要素的潜在影响等进行广泛、科学的调研，为产品的实际研发提供参考，提升智能终端产品交互的可解释能力，使其更加贴合人类认知。

3. 智能驾驶

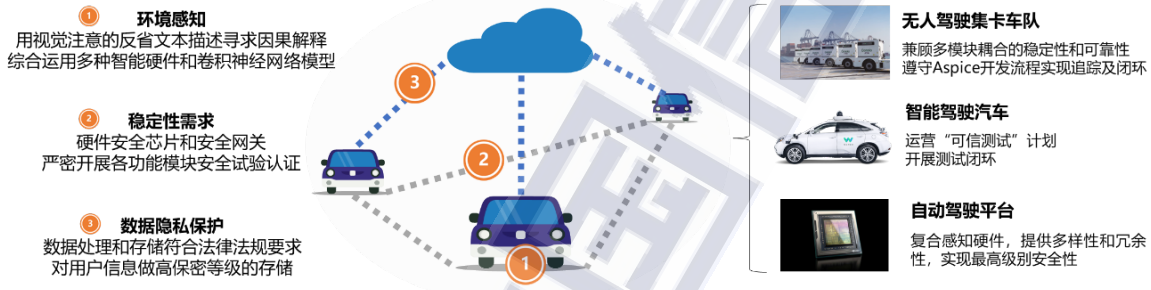
人工智能技术的发展加速了汽车智能化、网联化变革步伐，汽车成为具备感知和决策能力的智能载具。围绕芯片、系统、算法、人车交互四大核心，在智能算力的支持下，车企通过自研和接入第三方智能驾驶系统获得智能驾驶能力，头部车企已经能够实现“芯片+操作系统+算法”垂直整合，打造智能化、集成化智能驾驶系统，逐步打通全场景链路。

嵌入式系统、高精度导航、智能传感器等技术的发展为智能驾驶奠定了良好基础，使得传统上完全依赖人为控制的机动车辆具备了智能化的数据采集、汇总、分析、决策能力。智能驾驶车辆配备

的高精度定位设备，通过同步定位与建图模式算法即可进行地图测绘作业，配合车载摄像头拍摄周边的精确环境与地貌，在精准定位的同时也存在重大安全隐患；人类驾驶员在使用过程中对智能驾驶系统的高度认可，或误认为其已拥有与人类相近的驾驶能力，导致相关交通事故增多。因此，**对智能驾驶产品的要求不能仅仅局限于做出安全、实时的决策，而且需要解释这些决策是如何做出的，从而建立对人工智能的信任。**

针对智能驾驶产品环境感知、数据处理和存储、系统稳定可靠等不同模块产生了不同的可信方法。**融合多种传感手段和神经网络算法提升环境感知能力。**为弥补目前各类传感器在低光或恶劣天气工况下的缺陷，智能驾驶企业大多选择同时使用多种传感器，通过融合多种传感手段，结合 Transformer 神经网络与卷积神经网络(CNN)技术，帮助感知系统更深刻地理解环境语义，并解决 AI 大模型量产部署的难题；同时，用视觉注意的反省（introspective）文本描述寻求因果（post-hoc）解释，通过卷积神经网络模型 CNN 实时获取交通参与者精准的位置、类别和速度朝向等信息。**运用数据分级分类和区块链等技术做好数据处理和存储。**对各类数据进行分类分级，人脸、车牌等敏感数据实现轮廓保留前提下的脱敏处理；宝马、通用汽车、雷诺、福特等传统汽车制造商正鼎力支持利用区块链技术提供可信存证服务，并成立了移动开放区块链倡议（MOBI），探索区块链在汽车和移动出行领域的潜力；数据存储方面需要按照法律法规要求，对用户信息做高保密等级的存储，个人

信息或者重要数据应当依法在境内存储，确需向境外提供的，应当通过数据出境安全评估。**通过保证硬件安全和系统测试提升稳定可靠性能**。依靠硬件安全芯片和安全网关，构筑安全、可信的人工智能平台环境，实现网段隔离、访问控制、识别异常入侵，规避黑客攻击风险；从可信需求规划、设计、实施、集成、验证、确认、配置等方面，严密开展基本功能测试、软件测试、运营流测试、意外测试等功能模块的安全试验认证。



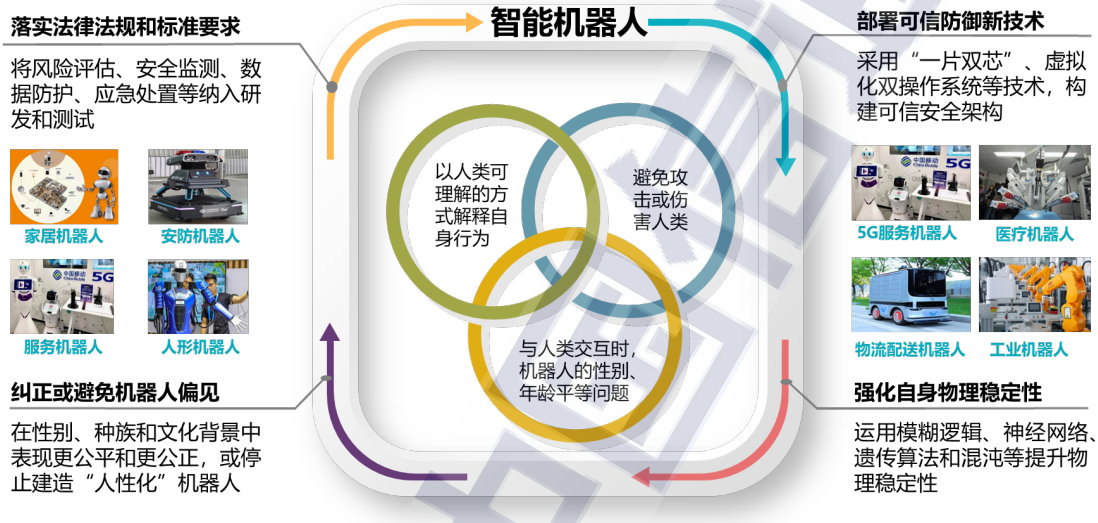
资料来源：中国信息通信研究院整理

图 20 智能驾驶领域的典型可信需求与实践

4. 智能机器人

智能机器人是产业智能化转型升级的重要切入口。因为不受严苛环境限制、不被情绪影响，能按照设定的算法和程序代替人类完成危险、繁琐的工作，20世纪中期，智能机器人产业逐步发展成型，满足了大批量生产的迫切需求，显著提升社会经济效率。近年来，随着自动化、新一代信息技术的迅速发展，智能机器人正向着轻型化、柔性化和人机协作发展，并不断地将人的认知能力与机器人的工作效率相结合，满足更多应用场景的需要。

在为人类带来方便的同时，智能的机器人的发展和应用也带了新的问题，例如，手术机器人通过计算机视觉技术识别病灶，一旦视觉系统受到污染，算法做出错误判断，将出现医疗事故；工业机器人系统一旦被黑客恶意操作，可能转而攻击人类；协作机器人人机交互算法未能及时识别和同步协同，影响使用效果。



资料来源：中国信息通信研究院整理

图 21 智能机器人领域的典型可信需求与实践

为实现智能机器人领域的可信落地，企业主要采取了以下几类做法。**一是落实法律法规和标准要求，在研发和测试环节加强可信能力建设，**结合《网络安全法》、等保 2.0 等法律法规，将风险评估、安全监测、数据防护、应急处置等纳入产品研发和测试环节，将智能机器人与人类交互的可解释性、自身的稳定性等作为智能机器人产品的重要亮点。**二是部署可信防御新技术，加强接口管理和物理隔离，增强抗攻击能力，**结合动态安全防护、威胁态势感知等，采用“一片双芯”、虚拟化双操作系统等技术把安全环境与互

联网环境进行严格物理隔离，构建智能机器人产品可信安全架构；基于区块链技术搭建安全高速独立隐身的机器人 VBN 网络，提供节点和源站间高频率链路质量探测和智能切换，打造多层安全保障。**三是灵活运用机器视觉、步态控制等算法优化动作控制，强化自身物理稳定性，**将模糊逻辑、基于概率论的推理、神经网络、遗传算法、多平面分割和混沌等具有更高鲁棒性的软计算技术应用到智能机器人中，通过柔顺控制、精准步态规划、视觉伺服、模型预测控制器等技术，让智能机器人稳定灵活地移动。**四是通过增加身份选项或消除身份信息，纠正或避免潜在的机器人偏见**，为用户提供多种机器人身份选项（比如允许用户选择机器人的“性别”“姓名”“形象”），或是避免建造具有人名和身份的“人性化”机器人，使机器人在性别、种族和文化背景中表现更公平、更公正，从而来识别和纠正潜在偏见。

5. 虚拟现实设备

虚拟现实（Virtual Reality，VR）设备强调全方位的沉浸式体验，弥补了智能手机等终端只能在二维图像层面接受文字、图片、音视频等信息的缺陷，将信息传播拓展到三维层面，进一步提升了媒介传递信息的效率和复杂程度，改变了参与者与媒介信息的交互模式。未来，以VR为代表的未来视频正在成为数字孪生、元宇宙等新业态的重点发展路径，既能作为日常生活的虚拟助手，又能在一定程度上成为生产力平台或文娱平台的演进形态。

虚拟现实设备在增强公平性方面具有独特优势，通过提供“身临其境”“感同身受”的虚拟体验，能够减少偏见和歧视，满足残障等人士的需求，增强人之人之间的交互。**随着虚拟现实设备的加速推广，隐私保护、歧视性、人身安全舒适性和安全性等方面的可信需求也逐步提升。**一是AR/VR技术需要以用户提供的基本信息和交互信息为起点，可能涉及所处位置、运动状态、生物特征，可能带来新的隐私保护和歧视性问题；二是针对虚拟现实设备的人身安全舒适性和安全性研究还有待深入，近期一起“VR性侵案”登上热搜，在虚拟世界发生的猥亵行为是否构成犯罪引发民众激烈讨论。

资料来源：中国信息通信研究院整理

图 22 虚拟现实设备领域的典型可信需求与实践

通过多环节的措施可以有效提升虚拟现实设备的可信水平。一方面，**赋予用户更大权限控制**，提供离线等多种可选的访问方式，允许用户自主选择何时以及如何向设备提供可以推断身份信息和共享敏感数据，以此限制敏感信息收集；设置“个人边界”等多项安全功能，使得其他虚拟人物无法触碰，以此杜绝虚拟现实骚扰行

为。另一方面，**将可信列入产品开发和测试必须考虑的范围，丰富身份选择，优化算法增强“现实感”，**邀请不同性别、职业、年龄等多类型的用户参与测试，并为用户提供多样化的身份选择，允许用户选择身份呈现的具体内容；不断向全面沉浸发展，融合运用近眼显示、感知交互、渲染处理等新一代技术，为用户提供身临其境的体验。**三是改进VR设备细节设计，提升用户使用舒适度，**虚拟现实设备大多直接由用户佩戴使用，其舒适度体验主要与重量尺寸、发热散热程度、面部贴合度和透气性等因素有关。近年来，随着设计水平和原料器件的改善提升，VR设备使用体验进一步优化，以VR显示器件为例，经历了从CRT到TFT-LCD/AMOLED的变革，在分辨率提升、效应速度加快的同时，屏幕体积不断缩小，重量持续降低。**四是制定引入安全保障和行为准则，构建良好的虚拟社区氛围，**通过邀请URL引入新用户，确保链接的分发仅限于受信任个人，限定可以创建或邀请新用户的范围；创建并发布行为准则，指定受信任的用户充当管理者，帮助新用户熟悉虚拟社区，严格限制和处罚恶意行为，积极引导形成友善、公平的虚拟社区氛围，减轻虚拟体验中的偏见和歧视。

四、可信人工智能前景展望

（一）未来发展趋势

形成产业共识，由各自表述向统一理念迈进。与人工智能产业发展历程相似，可信人工智能也经历了早期的摸索阶段，对于可信人工智能的内涵，虽然各个国家、机构、企业的表述并不相同，但

是侧重点较为统一，并已针对透明度、隐私保护、公平等方面形成初步共识，已经迈入了形成统一理念的新发展阶段。在此基础上，学术界和产业界对可信人工智能理念的共识将进一步凝聚和增强，随着可信人工智能在更多领域的持续落地，理念与实践间将形成良性互动，以可信理念指导产业实践落地，用产业实践不断丰富可信内涵，从而进一步巩固统一的产业共识。

突出理念落地，由抽象概念向具体实践发展。产业是可信理念的落脚点，目前，可信人工智能正处在由抽象的学术概念向产业实际加速落地的关键时期，产业主体正在由以理论界为主向产业界落地为主转变，产出形态由论文为主向产品服务为主转变，专利数量持续增长，并已经形成了一批可见可用的产业成果，诞生了一批代表性企业，覆盖人工智能产业链主要环节。未来，可信人工智能理念将更深层次地与人工智能产业链各环节结合，与元宇宙、人工智能生成内容等新技术、新业态互相促进，在人工智能场景和产品中得到更广泛的体现，为加快推进城市数字化转型、提升社会经济效率、改善人民生活水平发挥更大更多的赋能作用，确保人工智能造福人类。

注重动态平衡，由单一维度向综合框架转变。可信人工智能各项能力并不是完全独立的，比如可解释性和稳定性、稳定性和隐私保护间存在平衡的关系，可解释性和隐私保护又有协同的关系。这就要求可信人工智能的实现与衡量不能以单一要素为决定因素，而是应该结合实际应用需求，寻求最优的平衡效果。因此，可信人工

智能的一体化研究将是未来的重要趋势，不再仅仅围绕单一维度开展研究，而将会逐步过渡到关注各项可信能力之间的相互作用及其影响，使之具有较强的综合性和整体性，实现各项可信能力之间的动态平衡，以综合的可信框架促进产业协调健康发展。

优化技术布局，由前期研究向更深层次探索。可解释性方面，如何通俗易懂地向用户解释人工智能技术原理、算法自动决策机制、潜在风险及防范措施成为首要任务，并希望以此消除社会公众对新一代人工智能技术的疑惑和担忧。公平性方面，还需要为现有的公平性定义提供更强的条件，进一步拓宽公平性覆盖面，涵盖现有问题中没有考虑到的歧视，同时合理松弛以提高预测模型准确度。隐私保护方面，进一步推进分级分类，特别是针对人脸、指纹等敏感数据，如何在便捷地实现尽可能高的识别准确率的同时提升隐私保护性能。安全性（鲁棒性）方面，当前可验证鲁棒性的理论和实践存在较大落差，如何把一些外部知识交给机器学习模型来帮助其提高防御能力成为一个有前景的研究方向。

融合多元主体，由产学研向社会大众拓展。随着社会认知的深入，可信人工智能参与主体将进一步丰富，产业链多主体、多要素协调互动。在《网络安全法》《数据安全法》《个人信息保护法》等法律法规框架下，深圳、上海、北京等地加快推动人工智能立法，建立公共场所人工智能分级分类应用规范，形成具有地方特色的可信人工智能治理体系。行业协会、联盟和研究机构发挥积极作用，联合行业企业围绕先行领域制定和发布行业标准，在安全性、

可靠性等领域已经取得了一些成果供落地实施进行参考。“人脸识别第一案”等引发社会关注，人民群众对可信人工智能的了解和需求增加，参与程度大幅提升，从需求层面倒逼可信人工智能发展。

（二）产业发展建议

加强政策法规协同，协调制度、技术、人员整体推进。围绕政策法规体系完善、指南规范指引路径、从业人员可信理念培育，形成合理协同推进可信人工智能产业化落地。当前，可信人工智能政策法规更偏向理念指引，顶层设计还不完善，政策法规间没有形成完整、协调的体系关系，在模型权属、风险责任主体、结果公平等方面还需要予以明确。制定形成人工智能企业规范化发展指南，为企业提升可信人工智能技术、将 AI 向善贯穿于人工智能企业业务全流程提供方法论指引，帮助企业更好地实践可信。技术是可信落地的关键手段，从业者是可信落地的实际执行者，要注重从业人员可信理念培育，不断增强人工智能从业者可信认识，形成自觉践行可信人工智能理念的良好氛围，以制度、技术和人员协同加快可信人工智能产业化落地进程。

前瞻布局技术研究，以技术创新带动可信持续发展。前瞻布局可信开源学习框架，探索通用人工智能可信研究，强化代码可视化，持续优化现有技术，实现技术驱动的可持续创新。可信的通用人工智能的研究需要前瞻性发展，针对分布式计算、联邦学习、隐私计算等细分技术领域，不断建立完善可信开源学习框架，探索通用人工智能甚至是超级智能的可信研究。进一步强化代码可视化，

加快推广集成了OCR、NLP等人工智能能力的低代码平台，为用户提供一体化的智能服务能力，提升开发效率。围绕可解释的人机交互、提供更强的公平性定义、提供公平且可靠的算法、隐私数据分级分类、可验证鲁棒性等方面，持续优化技术，实现可信能力之间的均衡协调。

健全标准评估体系，系统性推进更多领域可信落地。构建创新治理闭环，以标准提升可信人工智能规范化程度，以评估、测试、认证体系建设加快应用和产品落地。一是构建完善的产品标准及评估体系。什么样的产品和应用能称得上是“可信的人工智能”还需要确立标准，既帮助大众快速接受、选择相关产品，提升对可信产品的认可度，又能为可信产品和服务供应商提供能力评估，帮助其更好地发挥长处补齐短板，从而加快推动可信人工智能的落地。二是分领域分阶段推进测试认证机制。发展可信人工智能评估评测及认证等监管技术，形成从理论原则、法律法规、合规增强技术、评估评测及认证的治理创新闭环；在医疗健康、自动驾驶、公共服务等领域分阶段推进研究能够适应场景特性、敏捷响应、动态的认证机制，系统性解决可信人工智能落地难的问题。

强化可信流程管理，将可信理念融入系统原生设计。运用可信理念重塑流程管理新能力、新形态，将人工智能可信融入系统原生设计，确保产品的可信品质。在《可信 AI 操作指引》等的基础上，面向业务数字化转型需求和生产研发实际，将可信理念融入流程发现、流程监控、流程优化、流程管理等各个环节，塑造形成全新的

流程管理能力。同时，将人工智能可信融入系统原生设计至关重要，在人工智能应用系统建设过程中，可将人工智能可信作为基础能力及要求纳入其设计范畴，以弹性适应未来人工智能安全的问题及挑战，以更加丰富的技术手段适配人工智能系统的不同发展阶段、不同应用要求、不同风险强度，为人工智能的可信应用保驾护航。

推动产业交流合作，共同打造可信产业生态朋友圈。加强国内外沟通合作，共建人工智能产业国际可信机制，做好宣传引导，增进社会公众的可信人工智能意识。进一步发挥行业组织和高水平专业平台作用，为可信人工智能企业和从业者提供技术交流、成果发布、资源对接的渠道，聚集行业各方力量共同打造可信人工智能生态圈。积极开展国际交流和技术合作，从人工智能系统测试和管控、危机沟通渠道等方面着手，主动参与国际人工智能治理，共建人工智能产业国际可信机制。加强对可信人工智能的宣传，通过对可信人工智能理念、典型案例的宣传和引导，增进社会公众对可信人工智能的认识了解，提高民众对人工智能潜能、挑战和局限性的认识，为可信人工智能产业发展营造良好的社会环境。

参考文献

[1] Bai T, Luo J, Zhao J, et al. Recent advances in adversarial training for adversarial robustness[J]. arXiv preprint arXiv:2102.01356, 2021.

[2] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[C]//International conference on machine learning. PMLR, 2018: 274-283.

[3] Kim M, Song Y, Wang S, et al. Secure logistic regression based on homomorphic encryption: Design and evaluation[J]. JMIR medical informatics, 2018, 6(2): e8805.

[4] Privacy Preserving Multi-party Machine Learning with Homomorphic Encryption

[5] Fengxiang He, Bohan Wang, and Dacheng Tao. "Piecewise linear activations substantially shape the loss surfaces of neural networks." International Conference on Learning Representation (ICLR), 2020.

[6] Chao Zhang, Dacheng Tao. “Risk bounds of learning processes for Lévy processes.” Journal of Machine Learning Research, 2013.

[7] Shaopeng Fu, Fengxiang He, Dacheng Tao. "Knowledge Removal in Sampling-based Bayesian Inference." International Conference on Learning Representation (ICLR), 2022.

[8] Zhang Q, Gu B, Deng C, et al. Secure bilevel asynchronous vertical federated learning with backward updating[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(12): 10896-10904.

[9] Zhang Q, Gu B, Deng C, et al. AsySQN: Faster Vertical Federated Learning Algorithms with Better Computation Resource Utilization[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 3917-3927.

[10] Zeke Xie, Fengxiang He, Shaopeng Fu, Issei Sato, Dacheng Tao, and Masashi Sugiyama. "Artificial neural variability for deep learning: On overfitting, noise memorization, and catastrophic forgetting." Neural Computation, 2021.

[11] Shaopeng Fu, Fengxiang He, Yang Liu, Li Shen, Dacheng Tao. "Robust

Unlearnable Examples: Protecting Data Privacy Against Adversarial Learning." International Conference on Learning Representation (ICLR), 2022.

[12] Fengxiang He, Shaopeng Fu, Bohan Wang, and Dacheng Tao. "Robustness, privacy, and generalization of adversarial training." 2020.

[13] Fengxiang He, Bohan Wang, and Dacheng Tao. "Tighter generalization bounds for iterative differentially private learning algorithms." Conference on Uncertainty in Artificial Intelligence (UAI), 2021.

[14] Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, Dacheng Tao. "DisPFL: Towards Communication-Efficient Personalized Federated learning via Decentralized Sparse Training." International Conference on Machine Learning (ICML), 2022.

[15] Liu Y, Kang Y, Zhang X, et al. A communication efficient collaborative learning framework for distributed features[J]. arXiv preprint arXiv:1912.11187, 2019.

[16] He F, Liu T, Tao D. Control batch size and learning rate to generalize well: Theoretical and empirical evidence[J]. Advances in Neural Information Processing Systems, 2019, 32.

[17] Fengxiang He, Tongliang Liu, and Dacheng Tao. "Why ResNet works? Residuals generalize." IEEE Transactions on Neural Networks and Learning Systems (TNNLS). 2020.

[18] Zhuozhuo Tu, Fengxiang He, and Dacheng Tao. "Understanding generalization in recurrent neural networks." International Conference on Learning Representation (ICLR), 2020.

[19] Galassi A, Lippi M, Torroni P. Attention in natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4291-4308.

[20] Jacobs A Z, Blodgett S L, Barocas S, et al. The meaning and measurement of bias: lessons from natural language processing[C]//Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020: 706-706.

[21] Kaneko M, Bollegala D. Debiasing pre-trained contextualised embeddings[J]. arXiv preprint arXiv:2101.09523, 2021.

[22] Bragg J, Cohan A, Lo K, et al. Flex: Unifying evaluation for few-shot nlp[J]. Advances in Neural Information Processing Systems, 2021, 34: 15787-15800.

[23] Fengxiang He, and Dacheng Tao. Foundations of Deep Learning. Springer, forthcoming.

[24] Huang Y, Evans D, Katz J. Private set intersection: Are garbled circuits better than custom protocols?[C]//NDSS. 2012.

[25] Angelou N, Benaissa A, Cebere B, et al. Asymmetric private set intersection with applications to contact tracing and private vertical federated machine learning[J]. arXiv preprint arXiv:2011.09350, 2020.

[26] Jingwei Zhang, Tongliang Liu, Dacheng Tao. “An Optimal Transport Analysis on Generalization in Deep Learning.” IEEE Transactions on Neural Networks and Learning Systems, 2021.

[27] Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, Dacheng Tao. "Topology-aware Generalization of Decentralized SGD." International Conference on Machine Learning (ICML), 2022.

[28] Tongliang Liu, Gábor Lugosi, Gergely Neu, Dacheng Tao. “Algorithmic stability and hypothesis complexity.” International Conference on Machine Learning (ICML), 2017.

[29] Zhang Q, Gu B, Dang Z, et al. Desirable Companion for Vertical Federated Learning: New Zeroth-Order Gradient Based Algorithm[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 2598-2607.

[30] Chao Zhang, Xianjie Gao, Min-Hsiu Hsieh, Hanyuan Hang, Dacheng Tao. “Matrix Infinitely Divisible Series: Tail Inequalities and Their Applications.” IEEE Transactions on Information Theory, 2020.

编制说明

本报告由中国信息通信研究院华东分院、中国信息通信研究院云计算与大数据研究所及京东探索研究院共同牵头。在编写过程中，得到以下单位的大力支持，在此特别表示感谢（按拼音首字母排序）：

- 高校:

  清华大学、上海交通大学、中国科学技术大学

- 科研机构:

  上海市计量测试技术研究院、上海人工智能实验室、上海人工智能研究院有限公司、上海市人工智能行业协会、之江实验室科技控股有限公司

- 企业:

  北京瑞莱智慧科技有限公司、北京市天元律师事务所、北京寓科未来智能科技有限公司、第四范式（北京）技术有限公司、杭州锘崴科技有限公司、华为技术有限公司、交通银行股份有限公司、蚂蚁科技集团股份有限公司、OPPO 广东移动通信有限公司、上海富数科技有限公司、上海商汤智能科技有限公司、上海燧原科技有限公司、上海西井信息科技有限公司、深圳市洞见智慧科技有限公司、数坤（北京）网络科技股份有限公司、腾讯云计算（北京）有限责任公司、新华三技术有限公司、中国电信集团有限公司、中企网络通信技术有限公司

中国信息通信研究院 华东分院

地址：上海市徐汇区云锦路 600 号航汇大厦 7 楼

邮编：200232

电话：021-64171028

传真：021-64171028

网址：www.caict.ac.cn