CAICT 中国信通院　京东探索研究院 JD EXPLORE ACADEMY

No.202106

可信人工智能白皮书

中国信息通信研究院
京东探索研究院
2021年7月

版权声明

前　言

当前，新一代人工智能技术迅猛发展，并向社会各个领域加速渗透，给人类生产生活带来了深刻变化。人工智能在带来巨大机遇的同时，也蕴含着风险和挑战。习近平总书记在 2018 年 10 月主持中央政治局第九次集体学习时强调，“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”。增强人工智能使用信心，推动人工智能产业健康发展已经成为重要关切。

发展可信人工智能正在成为全球共识。2019 年 6 月，二十国集团（G20）提出“G20 人工智能原则”，强调要以人为本、发展可信人工智能，这一原则也得到了国际社会的普遍认同。欧盟和美国也都把增强用户信任、发展可信人工智能放在其人工智能伦理和治理的核心位置。未来，将抽象的人工智能原则转化为具体实践，落实到技术、产品和应用中去，是回应社会关切、解决突出矛盾、防范安全风险的必然选择，是关系到人工智能长远发展的重要议题，也是产业界急需加快推进的紧迫工作。

无论是回顾可信人工智能的背景和历程，还是展望新一代人工智能的未来，本白皮书认为人工智能的稳定性、可解释性、公平性等都是各方关注的核心问题。立足当下，本白皮书从如何落实全球人工智能治理共识的角度出发，聚焦于可信人工智能技术、产业和行业实践等层面，分析了实现可控可靠、透明可释、隐私保护、明

确责任及多元包容的可信人工智能路径，并对可信人工智能的未来发展提出了建议。

由于人工智能仍处于飞速发展阶段，我们对可信人工智能的认识还有待进一步深化，白皮书中存在的不足之处，欢迎大家批评指正。

目 录

图 目 录

表 目 录

一、可信人工智能发展背景

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，正在对经济发展、社会进步、国际政治经济格局等诸方面产生重大而深远的影响。2020 年人工智能产业保持平稳增长，根据 IDC 测算，全球人工智能产业规模为 1565 亿美元，同比增长 12%；根据中国信息通信研究院测算，我国产业规模达到约 434 亿美元（3031 亿人民币），同比增长 15%。人工智能在带来巨大机遇的同时，也蕴含着风险和挑战。习近平总书记高度重视人工智能治理工作，强调要“确保人工智能安全、可靠、可控”，倡议推动落实二十国集团人工智能原则，引领全球人工智能健康发展。

（一）人工智能技术风险引发信任危机

当前，人工智能应用的广度和深度不断拓展，正在成为信息基础设施的重要组成。但在此过程中，人工智能也不断暴露出一些风险隐患，主要体现在以下几个方面：

算法安全导致的应用风险。以深度学习为核心的人工智能技术存在脆弱和易受攻击的缺陷，使得人工智能系统的可靠性难以得到足够的信任。如优步（Uber）自动驾驶汽车未能及时识别路上行人而致其死亡；据美国《财富》杂志报道，一家人工智能公司利用 3D 面具和合成照片实施欺骗攻击，成功破解多国的人脸识别系统[1]。

黑箱模型导致算法不透明。深度学习具备高度复杂性和不确定性，从而容易引发不确定性风险。由于人们无法直观地理解决策背后的原

[1] https://new.qq.com/omn/20191230/20191230A0FX0R00.html

因，人工智能与传统行业的进一步融合受到阻碍。如美国德州某学校使用人工智能系统判断老师教学水平，由于系统不能解释争议性决策的判断依据，遭到该校教师的强烈抗议，最终导致系统下线。

数据歧视导致智能决策偏见。人工智能算法产生的结果会受到训练数据的影响，因此，如果训练数据中存在偏见歧视，算法会受到歧视数据的影响，并进一步固化数据中存在的偏见歧视，导致依托人工智能算法生成的智能决策形成偏见。如美国芝加哥法院使用的犯罪风险评估系统（COMPAS）被证明对黑人存在歧视[2]。

系统决策复杂导致责任事故主体难以界定。人工智能的系统的自动化决策受众多因素影响，使得责任主体难以界定。对于自动驾驶、机器人等应用安全事故频发，法学专家表示，从现行法律上看人工智能本身还难以成为新的侵权责任主体，但人工智能的具体行为受程序控制，发生侵权时，到底是由所有者还是软件研发者担责，仍需进一步探讨[3]。

数据滥用导致隐私泄露风险。生物识别信息的频繁使用使得个人隐私数据泄露的可能性增大，数据一旦丢失会造成极大的安全风险。如 ZAO 通过用户协议条款违规收集人脸数据[4]，加重了人们对隐私数据滥用可能造成刷脸支付和身份认证相关安全风险的担忧。

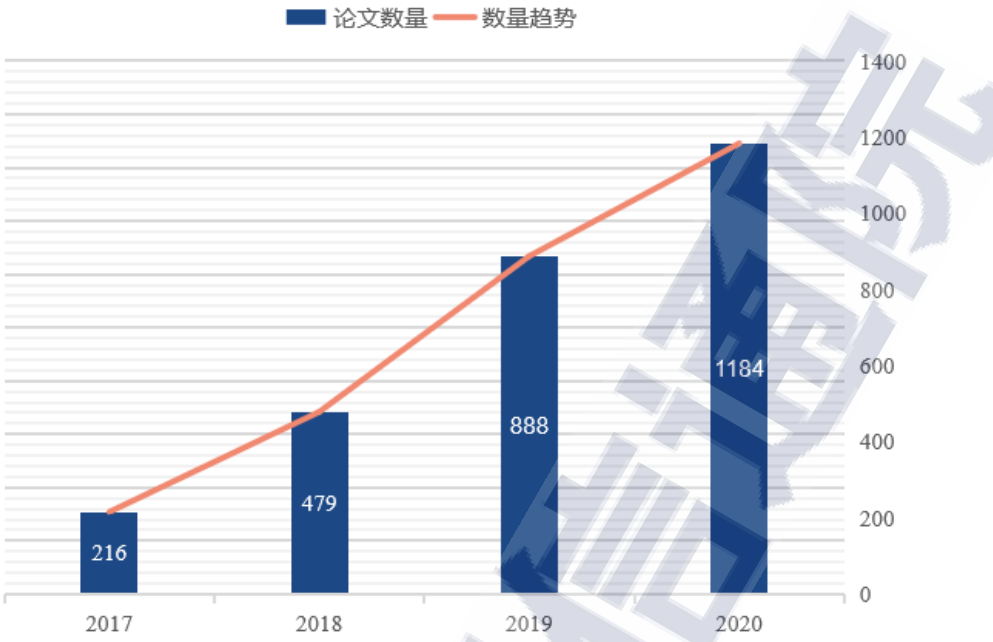
（二）全球各界高度重视可信人工智能

面对人工智能引发的全球信任焦虑，发展可信人工智能已经成为

[2] https://www.sohu.com/a/299700146_358040
[3] http://media.people.com.cn/n1/2018/0502/c40606-29959959.html
[4] http://finance.china.com.cn/industry/company/20190909/5075700.shtml

全球共识。2019 年 6 月，二十国集团（G20）提出“G20 人工智能原则”，在其五项政府建议中明确提出的“促进公共和私人对人工智能研发的投资力度，以促进可信赖的人工智能（Trustworthy Artificial Intelligence）的创新；需创建一个策略环境，为部署值得信赖的人工智能系统开辟道路。”已经成为国际社会普遍认同的人工智能发展原则。

学术界首先推开了可信人工智能的大门。中国科学家何积丰院士于 2017 年 11 月香山科学会议第 S36 次学术研讨会首次在国内提出了可信人工智能的概念，即人工智能技术本身具备可信的品质。**从学术研究角度，可信人工智能研究范畴包含了安全性、可解释、公平性、隐私保护等多方面内容。**2020 年可信人工智能研究论文数量相比 2017 年增长近 5 倍；美国国防高级研究计划局发布学术报告《可解释人工智能》并开展相关资助活动，致力于推动可信人工智能发展；顶级会议 AAAI 连续 2 年组织可解释人工智能（Explainable AI）专题研讨，并一直保持火热的研究态势。同时，围绕着机器学习公平性、可问责和透明性的研究已经形成“FAccT ML”（Fairness, Accountability and Transparency in Machine Learning）社区，在此基础上，ACM 从 18 年开始连续 4 年发起学术会议 ACM FAccT（ACM Conference on Fairness, Accountability, and Transparency）。



来源：Web of Science 官网

图 1 可信人工智能相关论文数量图[5]

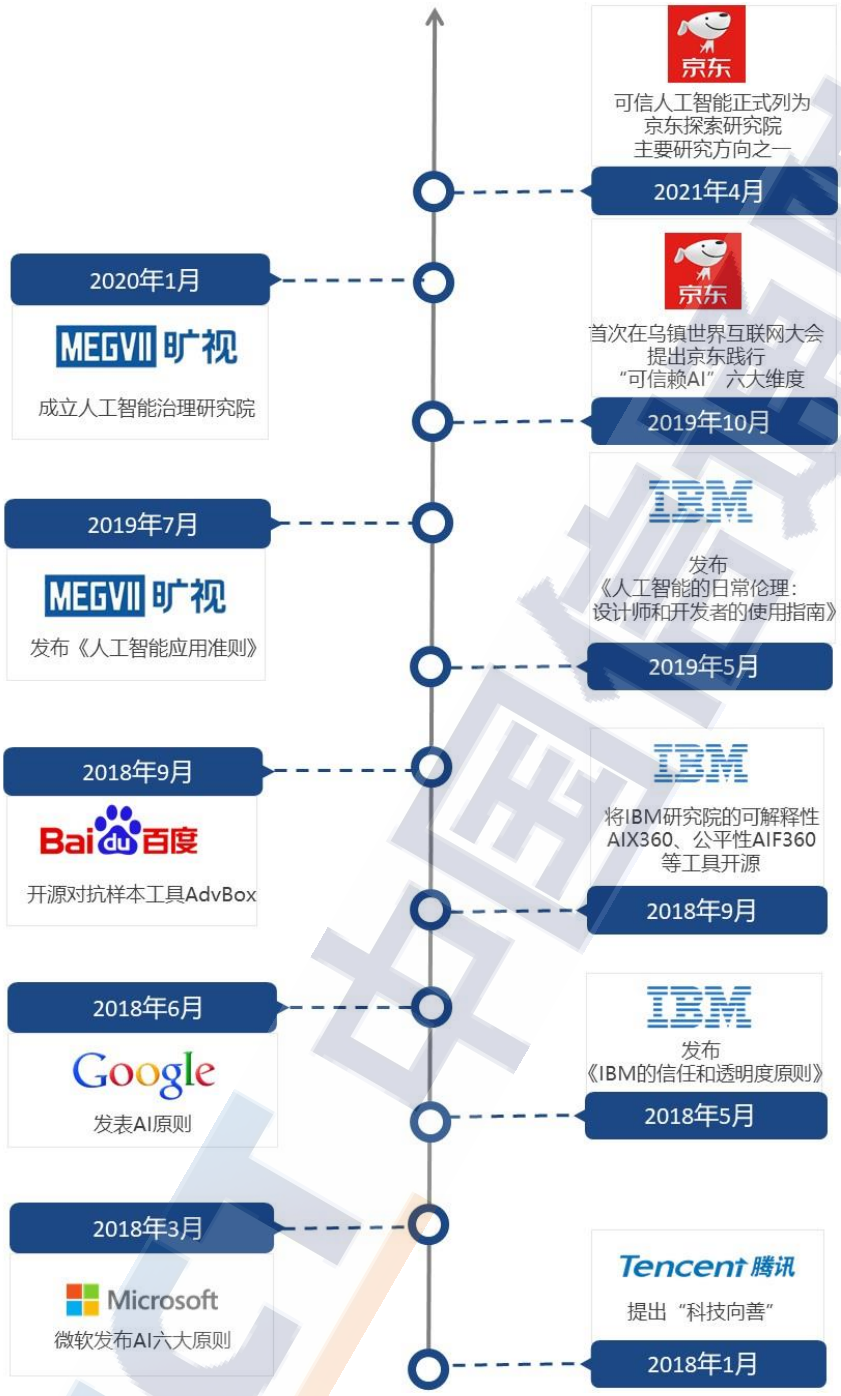
政府把增强用户信任、发展可信人工智能，放在其人工智能伦理和治理的核心位置。2020 年欧盟的《人工智能白皮书》[1]提出了人工智能“可信生态系统”，旨在落实欧洲人工智能监管框架，提出对高风险人工智能系统的强制性监管要求。同年 12 月，美国白宫公布了一项名为《促进政府使用可信人工智能》的行政命令[6]，该命令为联邦机构使用人工智能制定指导方针，旨在促进公众接受并信任政府在决策中使用人工智能技术。

标准化组织布局可信人工智能标准。ISO/IEC JTC1 SC42 专门设置了 WG3 可信人工智能工作组，目前已发布《信息技术 人工智能 人工智能的可信度概述》，正在推进《信息技术 人工智能 评估神经

[5] 中国信息通信研究院根据 Web of Science 检索整理。
[6] https://www.thepaper.cn/newsDetail_forward_10263830

网络的鲁棒性》系列研究工作。国内成立全国信息技术标准化技术委员会人工智能分技术委员会（SAC/TC 28/SC 42），同步推进相关研究。2020 年 11 月，全国信息安全标准化技术委员会 TC260 工作组发布了《网络安全标准实践指南—人工智能伦理道德规范指引》意见征求稿，针对可能产生的人工智能伦理道德问题，提出了安全开展人工智能相关活动的规范指引。

企业积极探索实践可信人工智能。IBM Research AI 于 2018 年开发多个人工智能可信工具，以评估测试人工智能产品在研发过程中的公平性、鲁棒性、可解释性、可问责性、价值一致性。这些工具已捐献给 Linux Foundation 并成为了开源项目。微软、谷歌、京东、腾讯、旷视等国内外企业也在积极开展相关实践工作，图 2 梳理了部分企业在可信人工智能方面的探索情况。



来源：资料整理

图 2 企业开展可信人工智能实践情况[7]

结合各方的表述，本白皮书认为“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

[7] 根据公开资料整理。

可信赖程度，**可信人工智能则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。**未来，随着人工智能技术、产业的不断发展，可信人工智能的内涵还将不断丰富。

（三）可信人工智能需要系统方法指引

当前对可信人工智能的要求及评价方法实操性不断加强。各国都意识到，伦理等“软性”约束如果缺乏相应落地机制，就容易出现道德漂白（ethics washing）的情况[8]，因此需要操作性更强的手段。2021年2月，德国发布了人工智能云服务一致性评价目录AI Cloud Service Compliance Criteria Catalogue (AIC4)[9]，从实操层面定义了评价云环境下人工智能的可信程度。4月，欧盟委员会公布了“制定人工智能统一规则（人工智能法）”并修订了相关立法提案，提出了一种平衡和相称的人工智能横向监管方法，围绕民生、人民基本权益划分了人工智能的四级风险框架，并规定了相应的处罚方式，意图通过法律手段提高市场的信任度，推动人工智能技术的推广和落地，推进人工智能可信。 5月，美国国家标准与技术研究院提出了评估人工智能系统中用户信任度的方法，并发布《人工智能和用户信任》（NISTIR 8332）[10]，从实操层面定义了评价人类使用人工智能系统时的信任体验。6月，美国国防部致力于通过教育和培训建立可信赖的人工智能能力，通过系统工程和风险管理方法在整个采购生命周期实施监管。

[8] https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/

[9] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.html

[10] https://www.nist.gov/news-events/news/2021/05/nist-proposes-method-evaluating-user-trust-artificial-intelligence-systems

人工智能立法进程不断加快，但具体细则仍需进一步明确；同时产业界探索可信人工智能也逐步迈入深水区。总体上看，**针对可信人工智能的实践仍处于相对分散的状态，缺少一套体系化的方法论，**以实现全面贯彻相关治理要求，体系化落实相关操作的实践性指引。基于此，本白皮书在全面梳理人工智能伦理约束、规范立法及优秀实践的基础上，**提出“可信人工智能框架”，**作为落实人工智能治理要求的一整套方法论，从产业维度出发，围绕企业和行业的可信实践进行了深入剖析，致力于在人工智能治理和产业实践之间搭建起连接的桥梁。

来源：中国信息通信研究院

图 3 可信人工智能核心内容

企业可信实践层面，框架以企业人工智能系统生命周期为参照，结合五项可信特征要求，针对周期各个环节提出了实操性要求，并对企业可信文化及可信管理机制建设提出了细节性建议。**行业可信实践层面，**框架从标准、评估及保障三个维度进行了详细阐述。

二、可信人工智能框架

可信人工智能从学术界提出，到各界积极研究，再到产业界开始

落地实践，其内涵也在逐步的丰富和演进。本白皮书认为，可信人工智能已经不再仅仅局限于对人工智能技术、产品和服务本身状态的界定，而是逐步扩展至一套体系化的方法论，涉及到如何构造“可信”人工智能的方方面面。图 4 给出了可信人工智能的总体框架。

来源：中国信息通信研究院

图 4 可信人工智能总体框架

可信人工智能是落实人工智能治理的重要实践，所遵循的可信特征与人工智能伦理和相关法律法规等要求一脉相承，均将以人为本作为其要求。从治理方式上来看，相较于伦理从宏观层面做出指引、法律以结果为导向做出约束，可信人工智能深入到企业内部管理、研发、运营等环节，以及行业相关工作，**将相关抽象要求转化为实践所需的具体能力要求**，从而提升社会对人工智能的信任程度。

可信特征层面。通过对全球范围内已经发布的 84 份政策文件按照词频进行梳理，可以看到当前人工智能治理原则已经收敛在**透明性、安全性、公平性、可问责、隐私保护**等五个方面[2]。尽管不同的组织

由于其文化背景、业务性质及管理制度等存在差异，对于这些共同原则的理解及实施方法有不同倾向，但从产业维度来看，**以上五项共识的核心理念均是围绕如何构建多方可信的人工智能而细化提出的。**这五项共识对于如何增强供给侧和需求侧双方使用人工智能的信任，协助监管机构培育可信的健康产业生态提供了指引。本白皮书参考全球五项共识（图 5）、中国人工智能产业发展联盟（AIIA）倡议以及发布的《人工智能行业自律公约》[3]和《可信 AI 操作指引》[4]，总结提出**可靠可控、透明可释、数据保护、明确责任、多元包容**等五项可信特征要素，用以指引实践可信人工智能时所需具备的操作能力。

伦理原则 Ethical principle	文档数量 Number of documents	关键词 Included codes
透明度 Transparency	73/84	Transparency, explainability, Excitability, understandability, interpretability, communication, disclosure, showing
正义与公平 Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
非恶意行为 Non-maleficence	60/84	Non-maleficence, security, safety,harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
责任 Responsibility	60/84	Responsibility, accountability,liability, acting with integrity
隐私权 Privacy	47/84	Privacy, personal or private information
仁慈 Beneficence	41/84	Benefits, beneficence, well-being,peace, social good, common good
自由与自治 Freedom and autonomy	34/84	Freedom, autonomy, consent, autonomy choice, self-determination, liberty,empowerment
信任 Trust	28/84	Trust
可持续性 Sustainability	14/84	Sustainability, environment(nature), energy, resources (energy)
尊严 Dignity	13/84	Dignity
团结 Solidarity	6/84	Solidarity, social security, cohesion

来源：资料整理

图 5 全球 84 份人工智能伦理文件中的主要关键词[11]

可信支撑技术层面，围绕着可靠可控、透明可释、数据保护、明确责任和多元包容等可信特征的要求，以理论研究和技术创新为核心

[11] 引自《人工智能原则的全球格局（The global landscape of AI ethics guidelines）》，中国信息通信研究院整理。

抓手，从而弥补当前技术的不足。例如研究可解释的新一代人工智能算法、具备隐私保护能力的隐私计算技术等，这需要学术界和工业界积极的探索。

企业可信实践层面，人工智能在社会上的广泛应用依赖于企业等主体将技术商品化，因此企业在可信人工智能方面的实践是可信方法论中的关键一环。应该意识到没有完美的技术，关键在于如何正确使用技术：一方面需要培育可信文化和建设可信管理制度等；另一方面需要在人工智能系统研发和使用的整个生命周期中贯彻可信特征的要求，从产品源头确保“可信”品质。

行业可信实践层面，可信人工智能需要整个行业的参与和实践。主要包括可信人工智能标准体系的建设，可信人工智能评估测试，以及人工智能可信保障等，通过构建保险等社会化方式分担人工智能技术和系统应用带来的风险。

三、可信人工智能支撑技术

随着社会各界对人工智能信任问题的不断关注，安全可信的人工智能技术已成为研究领域的热点[5,6]。研究的焦点主要是提升人工智能系统稳定性、可解释性、隐私保护、公平性等，这些技术构成了可信人工智能的基础支撑能力。

（一）人工智能系统稳定性技术

人工智能系统面临着特有的干扰，这些干扰来自于针对数据和系统的多种攻击方式，包括中毒攻击、对抗攻击、后门攻击等。这些攻击技术既可互相独立也可以同时存在。例如，中毒攻击通过按照特殊

的规则进行恶意评论等方式，向训练数据集投入干扰数据，继而影响推荐系统的准确度[17]；对抗攻击通过在道路交通标志牌上贴上特殊设计的图案，可以误导自动驾驶系统使其错误识别路牌上的信息，进而造成交通事故[18]；后门攻击具有隐蔽性，可能会被用于对 AI 供应链发动攻击。相比于传统的软件系统，此类干扰对人工智能系统的稳定性提出了更高的要求。

人工智能的稳定问题引起了持续而广泛的研究。针对人工智能模型的对抗攻击与中毒攻击早在 2012 及 2013 年就已出现。其中，对抗攻击的目的在于通过构造针对性样本来诱使人工智能系统决策出错；而中毒攻击的目的在于通过向人工智能模型的训练数据集注入中毒样本来劣化训练得到的模型的性能。在此之后，对抗攻击相继发展出了 FGSM（Fast Gradient Sign Method）、Carlini-Wagner 及 PGD（Projected Gradient Descent）等攻击方法；而中毒攻击的发展同样十分迅速，在其基础上更出现了后门攻击。后门攻击通过后门样本向人工智能系统植入后门，从而达到定向操纵人工智能系统的目的。该攻击与中毒攻击存在一定相似性，且常通过中毒攻击的方式来向系统植入后门。为抵御这些攻击，一些工作提出各类异常数据检测方法来检出并清除对抗样本、中毒样本、后门样本等恶意数据，从而减轻恶意攻击带来的干扰；通过在对抗样本上进行对抗训练来抵抗对抗攻击；利用模型剪枝、后门检测等技术抵抗后门攻击。

人工智能的稳定性仍然面临着较大的挑战。一方面，各种干扰手段层出不穷、持续演进，而新的攻击方法容易让旧的防御方法失效；

另一方面，干扰的形式正在逐步从数字世界向物理世界蔓延，例如通过打印对抗样本等手段能够直接对自动驾驶和人脸识别系统造成物理层面的干扰。未来在人工智能稳定性技术方面的研究将持续增多。

（二）人工智能可解释性增强技术

目前，以深度学习算法为核心的人工智能系统的运作就像是一个黑箱，人们只能看到数据的导入和输出，而不清楚内部的工作原理和判断依据。一方面，人们对训练得到的人工智能模型为何能具有极高的性能尚不清楚；另一方面，人工智能系统在做出决策时具体依赖哪些因素人们也不清楚。

针对人工智能算法可解释性的研究仍处在初期阶段，部分算法的理论框架有待完善[7,8,9,14,15]。例如，优化算法的有效性在决策树、支持向量机等一些简单的人工智能模型上已被很好地证明。然而，对于随机梯度下降算法为何能高效优化深度神经网络，学术界已经开展了大量的研究，但目前对于该问题的讨论仍未有定论。又如，针对人工智能模型如何利用数据特征做出预测，学术界已通过实验取得了一定的成果，但还缺乏理论支撑。为了使人工智能模型具有更好的可解释性，研究学者提出可以通过建立适当的可视化机制尝试评估和解释模型的中间状态；通过影响函数来分析训练数据对于最终收敛的人工智能模型的影响；通过 Grad-CAM（Gradient-weighted Class Activation Mapping）方法分析人工智能模型利用哪些数据特征做出预测；通过 LIME（Local Interpretable Model-agnostic Explanations）方法使用简单的可解释模型对复杂的黑盒模型进行局部近似来研究黑盒模型的可

解释性；还有部分研究则提出可以通过建立完善的模型训练管理机制提升人工智能系统实现过程的可复现性。

在人工智能的产业落地过程中，应最大限度地使人工智能系统的行为对人类更透明、更容易理解、更可信。一味地相信人工智能系统所做出的决策，而不对其决策过程进行解释会极大限制人工智能系统在国防、法律、医疗、教育等关键领域的普及，甚至引发严重的社会问题。增强人工智能系统的可解释性迫在眉睫。

（三）人工智能隐私保护技术

人工智能系统需要依赖大量数据，然而数据的流转过程以及人工智能模型本身都有可能泄漏敏感隐私数据。例如，在数据流转的任意阶段，恶意攻击者可以对匿名数据集发起攻击，从而窃取数据；在数据发布阶段，恶意攻击者可以使用身份重识别对匿名数据集发起攻击，从而窃取隐私信息；恶意攻击者也可以直接针对人工智能模型发起攻击，从而窃取隐私信息。例如，模型反转攻击可以根据受攻击模型的输出推断并重建其训练数据，从而窃取隐私信息；成员推断攻击可以推断给定数据样本是否来自受攻击模型的训练数据集，从而造成隐私泄露。

学界针对上述隐私泄露问题提出了多种针对性的保护方法，最常见的为基于差分隐私和基于联邦学习的隐私保护方法。差分隐私最早由美国学者 Cynthia Dwork[10]于 2006 年提出，是人工智能系统隐私保护能力的一个主要量化指标。其核心思想是一个具有优秀隐私保护能力的人工智能算法应当对输入数据中的微小扰动不敏感。基于该思想，

可以通过对数据进行下采样、顺序置换、添加噪声等方式，来防御攻击者进行隐私窃取。2016 年，谷歌公司的一项工作首次将差分隐私应用于深度学习中，其通过在模型训练过程中向梯度加入高斯噪声来增强深度模型的隐私保护能力。该工作展现了差分隐私法在大规模人工智能模型中的应用潜力。目前，一些头部科技公司已将差分隐私法应用于部分真实的业务中。联邦学习[19]在 2015 年提出，其能在不收集用户数据的条件下进行人工智能模型的训练，以期保护隐私信息。具体来说，联邦学习将模型部署到用户设备；各用户设备使用自己的私有数据，计算模型参数的梯度，并将其上传中央服务器；中央服务器对收集到的梯度进行融合，传回各用户设备；各用户设备利用融合后的梯度更新模型参数。需要指出的是，一些初步研究表明，联邦学习方法仍存在一定的隐私泄露风险。有实验显示，联邦学习可能泄露一定量的本地用户数据[11]，同时有理论指出，联邦学习可能会在一定程度上弱化人工智能系统的隐私保护能力[12]。因此，还需要针对联邦学习进一步优化，提升其用户隐私保护的能力。一个可行的方向是将联邦学习和差分隐私相结合，以构建隐私保护能力更强的人工智能系统。

在当前时代下，越来越多的隐私信息承载于数据之中，人们对隐私数据保护的关注更胜以往，部分国家也开始从立法层面制定隐私数据的使用规范。针对隐私保护进行研究能使得人工智能系统符合法律的基本规范和要求，完善可信人工智能的建设。

（四）人工智能公平性技术

随着人工智能系统的广泛应用，其表现出了不公平决策行为以及对部分群体的歧视。学术界认为，导致这些决策偏见的主要原因如下：受数据采集条件限制，不同群体在数据中所占权重不均衡；在不平衡数据集上训练得到的人工智能模型，可能会为了在整体数据上的平均性能，而牺牲在少量数据上的性能，造成模型决策不公平。

为了保障人工智能系统的决策公平性，相关研究者主要通过构建完整异构数据集，将数据固有歧视和偏见最小化；对数据集进行周期性检查，保证数据高质量性。此外，还有通过公平决策量化指标的算法来减轻或消除决策偏差及潜在的歧视。现有的公平性指标可以分为个体公平性与群体公平性两大类[13,16,20]。其中，个体公平性衡量智能决策对于不同个体的偏见程度，而群体公平性则衡量智能决策对于不同群体的偏见程度。另一方面，基于公平性指标的算法则大致能分为预处理方法、处理中方法及后处理方法共三大类。预处理方法通过删除敏感信息或重采样等方式对数据进行清洗，从而降低数据中存在的偏差。处理中方法通过在人工智能模型训练过程中加入与公平性量化有关的正则项，提高训练得到的模型的公平性，例如，有工作采用 Rényi 相关性作为正则项，并利用最小-最大优化算法来减少模型预测与敏感属性之间的任意潜在相关性。后处理方法通过对模型输出进行调整，进一步提高训练得到的模型的公平性，例如，有工作基于多重精确度（Multiaccuracy）的概念提出多精度提升法（Multiaccuracy Boost），以减轻黑盒人工智能系统的决策偏差。

人工智能在敏感领域的应用越来越多，包括招聘、刑事司法、医疗等，其公平性也受到了广泛的担忧。公平性技术能够从技术角度对数据进行均衡，从而进一步引导模型给出公平的结果，这对于提高人工智能系统决策公平性具有重要意义。

当前越来越多的研究关注到人工智能在稳定性、可解释性、隐私保护、公平性等问题上的挑战，随着研究的不断深入，势必将会涌现出更稳定、更透明、更公平的人工智能理论及技术，而这些技术是未来实现可信人工智能的基石与重要保障。

四、可信人工智能实践路径

本白皮书参考中国人工智能产业发展联盟发布的《可信 AI 操作指引》相关内容，结合调研访谈人工智能企业研发实际情况，从企业和行业层面总结提出了可信人工智能的实践路径。

（一）企业层面

企业是人工智能技术、产品或服务的研发和使用的核心主体，也是可信人工智能落地实践中最重要的主体。可信人工智能在企业的实践是一项整体的、发展的、非传统的系统工程，需要从企业文化、管理制度等方面入手，同时在人工智能系统研发中全面落实相关技术要求。

1. 将可信人工智能融入企业文化

企业文化是一个企业整体价值观、共同愿景、使命及思维方式的具体体现，企业要发展可信人工智能，就要把可信理念融入企业文化。

（1）企业管理层要认可“可信”的方向

作为企业运营的核心，企业管理层要在发展可信人工智能层面达成一致，全面树立以人为本的价值观，认同透明可释、多元包容、可靠可控、明确责任和隐私保护的特征要素，将可信人工智能融入到企业经营管理的各个方面，以促进企业整体可信度的提高。

（2）员工要加强“可信”的学习和实践

企业可以制定有关可信人工智能的学习培训计划，通过邀请外部专家宣讲、发放可信人工智能书籍或介绍材料等方式，在员工中普及“可信”理念，推广使用可信相关技术或工具，鼓励员工在工作中不断创新和实践可信人工智能。

（3）企业要营造“可信”的文化氛围

企业可在办公场地、网站、宣传资料、新闻稿件中，体现可信人工智能的元素，展现企业自身探索可信人工智能的实践案例，鼓励员工探讨可信人工智能话题，激励在可信人工智能实践中做出贡献的团队或个人。

2. 完善可信人工智能的管理制度

管理制度是实施管理行为的依据，是社会再生产过程顺利进行的保证。企业要实现可信人工智能，就要在管理制度中有所体现。

（1）建立可信人工智能团队

在企业内部建立专门的团队（或虚拟组织），负责可信人工智能的管理工作。建议由企业主要负责人担任领导职务，便于直接指挥、协调其他部门参与可信人工智能相关工作；可根据具体业务情况，细

分成若干子组；人员构成方面建议由有法律、研发背景的人员专职或兼职组成。明确落实部门及有关人员的责任义务。

（2）建立并落实可信人工智能人员管理制度

由可信人工智能管理部门牵头，人力资源、研发、法务等部门配合，共同制定可信人工智能相关人员的管理制度，主要对企业内部涉及人工智能需求分析、产品设计、研发、测试及可信管理的相关人员，明确人员管理、教育培训、考核等要求。要切实落实人员管理制度，对相关人员定期进行教育培训及考评，逐步提升人员专业水平。

（3）建立并落实可信人工智能系统研发与使用的管理制度

建立人工智能系统研发阶段的管理制度，明确责任部门和人员、工作内容、工作方法、工作流程和工作要求，由可信人工智能管理部门牵头督促落实；要明确可信人工智能系统使用阶段的管理制度，制定应急预案及救济措施，确保系统在使用阶段能满足可信的要求，或在发生问题后能及时有效解决，最大限度降低伤害和减小损失。

（4）配备实现可信人工智能的必要资源

企业内部要做好统筹，为实现可信人工智能配备必要的资源，包括但不限于必要的人员、资金、场地、设施等。

（5）建立制度的迭代和和更新机制

企业要随着人工智能治理态势的变化及相关政策法规的出台，由可信人工智能管理部门牵头，根据实际情况，不断优化和完善管理制度，确保能够及时适应，达到最优效果。

3. 将可信人工智能要求嵌入到研发应用全流程

（1）规划设计阶段

企业在人工智能系统生命周期的开始就需要充分考虑落实可信人工智能的特征要素，将可信的理念根植于需求分析和系统详细设计等规划设计的关键环节中，从而使后续的研发测试和运营能够始终符合可信人工智能的核心要求。

结合当前软件产品设计的常见流程，企业可以通过专门设立的可信团队，从两个方面帮助产品团队制定人工智能系统可信设计方案：

一是提出人工智能系统的可信设计要求。在完成产品需求分析之后，应充分调研人工智能智能系统面临的潜在风险，有针对性地提出应对手段，如针对系统安全性、失效保护机制、可解释性、数据风险、系统责任机制、用户权利义务、系统公平性等方面，提出相应的可信设计要求清单。

二是评审人工智能系统的可信设计方案。可信团队中各个专业领域方向专家，需要结合其自身的专业知识、工作经验、典型案例等，验证人工智能系统可信设计方案的可行性，发现潜在问题，提供启发性引导和更多的可信设计延展思路，为后续可信设计方案的修改完善提供意见，确保将可信人工智能的核心特征与系统设计进行融合，减少信任漏洞，预防潜在风险事故的发生。

（2）研发测试阶段

可靠可控方面，应着力提升人工智能系统自身的防御能力并确保人类的监督和接管权力。人工智能系统自身的防御能力可以从数据和

模型两个层面进行提升，数据层面的防御方法包括恶意数据预清洗、利用数据增广等技术提高模型鲁棒性。而在模型层面，除了对生产环境中的模型进行加密、限制生产环境中模型的恶意查询交互次数等传统安全防御方法外，另一个主要方法为对抗训练。人工智能模型极易被特殊构造的攻击样本干扰，对抗训练算法可以使用对抗样本来训练人工智能模型，从而提升模型对于对抗样本的鲁棒性，使得模型更不容易被对抗样本干扰。另外，在研发过程中需要针对人工智能系统设置后备计划，确保在上线部署后面对突发情况时，人工智能系统能够自动调节恢复、被专业人员快速接管，或通过“一键关停”的方式被人为终止服务。

透明可释方面，应重点提升人工智能系统的可复现性。当前算法的可解释性研究与人工智能技术在应用领域的高速发展相比仍较为落后，因此企业在研发和测试阶段应主要从提升系统的可复现性入手，不仅可以增强系统透明性，同时也能一定程度上降低后期系统审计和责任追溯的难度。相关的主要措施包括：建立完善的数据集管理机制，结合现有数据管理策略和工具，详细记录系统各版本训练过程中训练集、测试集的来源和构成情况，以及训练过程所采用的数据预处理操作；建立完善的模型训练管理机制，详细记录训练模型时所用的硬件平台、系统配置、软件框架、模型版本、模型初始化、超参数、优化算法、分布式运行策略、网络速率、指标、测试结果、以及所采用的其他技巧和工程技术手段等。

数据保护方面，应通过开展数据治理以避免训练数据的非法收集、

滥用和泄漏等问题，同时探索使用隐私保护算法训练人工智能系统。从算法层面提高人工智能系统的隐私保护能力，使用差分隐私或联邦学习等技术，例如苹果公司的用户数据收集，美国人口普查等。微软和哈佛大学合作创办的人工智能项目 OpenDP 开发了很多开源的差分隐私工具包，以对模型和数据提供更多保护。

明确责任方面，应全面审计人工智能系统的实现流程，提升系统可追溯能力，确保系统及服务的源头可信。审计的主要环节包括数据准备、模型训练、模型评估三个环节。数据准备过程的审计有助于确认训练数据的收集是否合法合规、是否涉及侵犯隐私等情况，数据的处理是否遵照了标准的标注和预处理手段，数据的存储是否采用了加密、访问限制等安全性措施。模型训练环节是为人工智能系统赋予“智能”的关键，对硬件平台、软件框架、算法选择、调参过程等训练的关键环节进行全面审计，能够帮助对人工智能系统进行追溯。模型的评估很大程度上能够反映人工智能系统在实际应用中的性能表现和泛化能力，标准严谨的评估过程往往能够发现错误，衡量模型质量，并判断其是否能满足设计要求，帮助回溯系统实现过程中存在的问题从而进行不断改进，因此需要详细审计模型在验证集和测试集上的指标表现和变化。

多元包容方面，应着重关注训练数据集的公平多样性，避免数据偏差造成的信任缺失。人工智能系统的表现依赖于训练数据的质量，数据集可能包含隐含的种族、性别或意识形态偏见等问题（表 1），从而导致人工智能系统可能会做出不准确或带有偏见和歧视的决策。

企业应注重提升训练数据的多样性和公平性从而符合多元包容的要求，一方面留意数据中可能会出现的固有的歧视和偏见，以采取主动措施来削弱偏见带来的影响；另一方面，对数据集进行周期性检查，保证数据高质量性。此外，**测试环节采用基于公平决策能力的量化指标对人工智能系统进行测试**。目前，具体操作可包括：

● 通过可靠、合法的来源收集数据，保证数据来源的可信程度。

● 通过统计学的方式或相关工具集，检查数据集中样本、特征、标签的准确性和完整性，并根据检查结果及时进行相应的调整。

● 根据真实部署环境的变化及时更新数据集，保证数据集的时效性和相关性。

● 构建易用的数据集格式和接口，简化数据集的读取和调用流程，防止误操作。

● 在对人工智能模型的公平决策能力进行定量分析时，需要根据具体应用场景和特定需求选取合适的量化指标，兼顾考虑个体公平性和群体公平性指标。

表 1 数据集中常见的固有偏见

序号	数据质量	描述
1	报告偏见	人工记录数据集收集情况属性，无法准确反映真实客观情况
2	自动化偏见	自动化的软件工具生成的结果本身存在偏见
3	选择偏见	数据集中选择的样本未能反映样本的真实分布情况
4	群体归因偏见	人们倾向于将个体的真实情况泛化到所属的整个群体
5	隐形偏见	通常根据不一定普遍适用的模型和个人经验做出假设

来源：资料整理

（3）运营使用阶段

在人工智能实际运营和使用阶段，需要做好人工智能系统的解释说明工作，持续监测人工智能系统的各项可信风险，积极优化人工智能系统。

一是对用户披露人工智能系统的技术意图。在算法的可解释性尚未成熟的情况下，对于人工智能系统技术意图的理解可以从建立适当的人机交流机制、披露系统决策的功能逻辑和使用要求、明示系统的潜在错误决策风险等方面入手。具体来说，企业应在上线部署人工智能系统时，建立适当的人机交流机制，如设置一个功能模块，通过一种通俗易懂的表达方式，如文字、图形标识、语音提示等，明确地告知用户当前是否正在与人工智能系统进行交互。在实际应用过程中，用户也应被明确告知有关人工智能系统的基本功能、性能表现、使用要求、面向对象、以及系统在决策流程中扮演的角色等基本信息。

二是持续开展人工智能风险监测等。建立用户反馈渠道，及时收集用户的真实意见，并对整个系统进行优化和迭代。监测人工智能系统在实际使用过程中的各种风险，不断完善监督、赔偿等机制，对造成实际损害的人工智能系统及时开展责任追溯工作和赔偿工作。

（二）行业层面

可信人工智能的实现不仅仅是企业单方面的实践和努力就能够完成的，更需要多方参与和协同，最终形成一个相互影响、相互支持、相互依赖的良性生态。这个生态主要包括标准体系、评估验证、合作交流等具体内容。

一是构建可信人工智能的标准体系。政策法律只能规定原则和底线，需要标准从可执行、可落地的层面来进行具体的指导和约束。目前，部分国家已在制定或出台人工智能治理的原则或法律，在此基础上，可结合人工智能技术、产品或场景制定具体的标准规范。如：2021年4月，我国《信息安全技术人脸识别数据安全要求》国家标准面向社会公开征求意见，该标准主要为解决人脸数据滥采、泄露或丢失，以及过度存储、使用等问题，对于《个人信息保护法》草案中人脸识别相关的规定也作了更多阐述和细化。

二是开展第三方的评估验证。第三方的评估验证是检验目标对象是否达到相关要求的有效手段，由于人工智能技术的复杂性，更加需要专业的第三方机构给予支持。围绕可信人工智能的特征，要重点考虑系统安全性、鲁棒性、可复现性、数据保护、可追溯性、公平性等方面的表现。中国人工智能产业发展联盟2020年也发布了首批商用人工智能系统可信评估结果，涉及11家企业的16个人工智能系统，为用户选型提供了重要参考；欧盟最新发布的人工智能立法提案中，也提出了将由权威第三方机构开展可信评估等举措。

三是摸索市场化的保险机制。人工智能技术应用与其他信息系统一样，无论达到多高的保障等级，出现问题的风险始终存在。这就需要创新工作方法，以其他方式转移风险损失。保险是风险补偿的重要措施，可最大程度转移风险，弥补用户损失。建议人工智能企业及保险机构可探索人工智能产品应用的保险机制，针对风险事故进行量化评估，提供风险补偿，帮助完善可信人工智能生态。

五、可信人工智能发展建议

打造可信任的人工智能系统已经成为各界关注的焦点和努力方向。通过实践可信人工智能方法论，有助于提升人工智能的可信水平，让其更好的被社会大众接受。**可信人工智能并非一成不变，而是伴随着人工智能技术、伦理、法律的发展，将会不断演进以适应新的发展需要。**这也将对所涉及到的各类主体提出新的要求。

（一）政府层面加快推动我国人工智能监管及立法进程

构造体系化的人工智能法律监管框架。一是完善现行法律法规以适应发展需要，在《网络安全法》、《数据安全法》，以及未来将发布的《个人信息保护法》等基础上，梳理人工智能系统监管过程面临的适用问题，不断完善法律法规。二是推进新立法工作主动应对新风险，深入研究人工智能引发的新问题和新态势，及时梳理形成立法建议。三是创新手段推进法律的落地执行，探索采用试点、沙箱等监管方式，研发智能化监管工具，不断提高监管的效率和灵活性。此外，要坚持统筹推进人工智能领域国内法治和涉外法治，积极参与多双边区域合作机制，推动国际间人工智能治理规则制定，寻求共识、弥合分歧。

（二）技术研究层面需全面做好体系化前瞻性布局

可信人工智能一体化研究将是未来重要趋势。当前针对可信人工智能的研究，多是从安全、隐私、公平等单一维度展开。已有研究工作表明，安全性、公平性、可解释性等不同要求之间存在相互协同或相互制约的关系，若仅考虑某一个方面的要求则可能会造成其他要求

的冲突。因此需要针对可信人工智能构建一体化研究框架，以保持不同特征要素之间的最优动态平衡。

面向可信通用人工智能（AGI）的研究需要提前布局。目前无论是人工智能治理还是可信人工智能的工作，大多是面向弱人工智能技术及应用来进行的，通用人工智能甚至是超级智能尚未引起足够关注，而这些一旦出现将是关乎人类命运的重大事件，需要具有前瞻性的布局，如通过发展超级深度学习、量子机器学习等前沿技术探寻通用人工智能的发展路径。同时，我们也需要在探索强人工智能时开展可信相关的研究工作。

（三）企业实践层面需匹配业务发展实现敏捷可信

企业拓展人工智能技术应用过程中应注重可信人工智能敏捷迭代。随着人工智能技术与不同行业的广泛融合，其应用深度与日俱增，企业所面临的可信特质要求将不断扩充，这就对企业应具备的可信实践能力提出了更高的要求。一方面应研发可信人工智能检测和监测工具，以匹配业务发展需要，并针对行业应用的独特性进行升级和迭代。另一方面应积极与监管部门对接，主动配合参与数字沙盒、安全港、试点应用、标准合规等监管措施，构建内部和外部相协调的敏捷可信机制。

（四）行业组织层面需搭建交流合作平台打造可信生态

鼓励行业组织围绕可信人工智能领域搭建专门交流平台，号召产业各方共同打造可信人工智能生态。可信人工智能是一项复杂的系统

化工程，需要多方共同参与，应充分发挥行业组织优势，广泛吸纳各方优秀实践经验，编制可信人工智能操作指引；围绕人工智能研发管理、技术保障、产品应用等方面，建立完善可信人工智能标准体系；加快研发人工智能测评和监测能力，运用评估测试、跟踪监测等多种手段，持续推动可信人工智能在产业界落地。

参考文献

[1] EUROPEAN COMMISSION. WHITE PAPER On Artificial Intelligence-A European approach to excellence and trust[R/OL]. (2020-02-19) https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[2] Jobin A., et al. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(2).

[3] 中国人工智能产业发展联盟. 人工智能行业自律公约[R/OL].(2019-0808)http://aiiaorg.cn/uploadfile/2019/0808/20190808053719487.pdf

[4] 中国人工智能产业发展联盟. 可信 AI 操作指引[R/OL].(2020-0923)http://aiiaorg.cn/uploadfile/2020/0923/20200923064427421.pdf

[5] 张钹等. 迈向第三代人工智能[J]. 中国科学:信息科学, 2020, v.50(09):7-28.

[6] 何积丰. 安全可信人工智能[J]. 信息安全与通信保密, 2019(10):4-8.

[7] Liu T., et al. Algorithm-dependent generalization bounds for multi-task learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 39, pages 227-241, 2016.

[8] He F., et al. Control batch size and learning rate to generalize well:

Theoretical and empirical evidence[C]. In Advances in Neural Information Processing Systems, pages 1141-1150, 2019.

[9] Tu Z., et al. Theoretical analysis of adversarial learning: A minimax approach[C]. In Advances in Neural Information Processing Systems, pages 12280–12290, 2019.

[10] Dwork C., et al. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, volume 9, pages 211–407, 2014.

[11] Zhu, L., et al. Deep leakage from gradients[C]. In Advances in Neural Information Processing Systems, 2019.

[12] He F., et al. Tighter generalization bounds for iterative differentially private learning algorithms[J]. arXiv preprint arXiv:2007.09371, 2020.

[13] Dwork C., et al. Fairness through awareness[C]. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pages 214–226, 2012.

[14] Ribeiro M. T., et al. "Why should i trust you?" Explaining the predictions of any classifier[C]. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135-1144, 2016.

[15] Ribeiro M. T., et al. Anchors: High-precision model-agnostic explanations[C]. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[16] Calders, T., et al. Building classifiers with independency constraints[C]. IEEE International Conference on Data Mining Workshops. pages 13-18, 2009.

[17] Fang, M., et al. Poisoning attacks to graph-based recommender systems[C]. In Proceedings of the 34th Annual Computer Security Applications Conference, pages 381-392, 2018.

[18] Eykholt, K., et al. Robust physical-world attacks on deep learning visual classification[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1625-1634, 2018.

[19] McMahan, B., et al. Communication-efficient learning of deep networks from decentralized data[C]. In Artificial Intelligence and Statistics, pages 1273-1282, 2017.

[20] Hardt, M., et al. Equality of opportunity in supervised learning[C]. In Advances in Neural Information Processing Systems, pages 3323-3331, 2016.

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62309514

传真：010-62304980

网址：www.caict.ac.cn