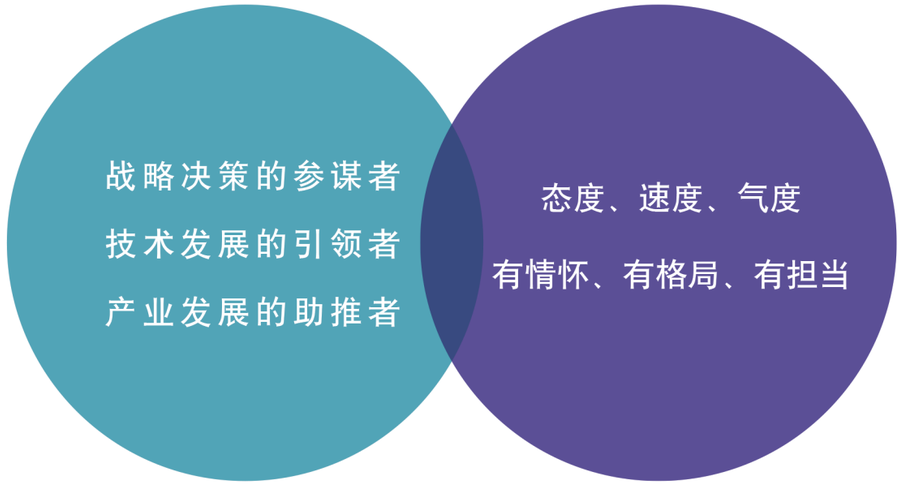
中国联通研究院根植于联通集团（中国联通直属二级机构），作为中国联通科技创新专业子公司，自2022年起，挂牌成立中国网络安全研究院、下一代互联网宽带业务应用国家工程研究中心及首个国家级网络安全产业知识产权运营中心，形成“两院两中心”发展格局，开创了研究院高质量发展的新局面。中国联通研究院作为服务于国家战略、行业发展、企业生产的战略决策参谋者、技术发展的引领者、产业发展的助推者，坚持以高水平科技自立自强为使命担当，聚焦网络强国、数字中国主责，拓展联网通信、算网数智业务，形成“态度、速度、气度、有情怀、有格局、有担当”的企业文化，以下一代互联网、光网络、5G-A/6G、网络安全、数智网优、低空智能网联和新型智库研究七个领域为主攻方向，坚持“四个聚焦”，开展关键核心技术攻关、科创力量建设、专业技术人才队伍建设、创新成果转化等工作，着力实现价值创造提升、战略科技力量提升、专精特新能力成色提升，争做通信行业科技创新主力军，努力建设成为“国家信赖、行业领先、集团倚重、员工自豪”现代化一流研究院。



中国联合网络通信有限公司研究院

地址：北京市亦庄经济技术开发区北环东路1号
电话：010-87926100
邮编：100176

BRICS Business Council
金砖国家工商理事会
中方数字经济工作组

2024

中国人工智能伦理风险白皮书

中国联通研究院 | 中国联通国际公司 · 2024年9月

中国联通
China unicom

版权声明

本报告版权属于中国联合网络通信有限公司研究院，并受法律保护。转载、摘编或利用其他方式使用本报告文字或者观点的，应注明“来源：中国联通研究院”。违反上述声明者，本院将追究其相关法律责任。

目录

前言 1
一、人工智能方兴未艾，风险防范未雨绸缪 4
（一）数字经济发展孕育人工智能全面成长 4
（二）人工智能高歌猛进，融入百业大显身手 6
（三）中国 AI 快速应用，发展举世瞩目 9
（四） 伦理规范有利于人工智能长治久安 13
二、他山之石可以攻玉，先进经验学习借鉴 20
（一）国际 AI 伦理研究现状 20
（二）国际 AI 伦理研究贡献及未来走向 21
（三）全球主要国家 AI 政策 23
（四）全球智者的观点 26
（五）全球头部企业的选择和态势 29
三、居安思危如履薄冰，杜绝风险伦理安全 34
（一）伴随社会和技术进步，AI 伦理安全理念与时俱进 34
（二）加强监管，防范失控性风险 35
（三）完善法规，杜绝侵权性风险 39
（四）确保透明公正，规避歧视性风险 42
（五）加强培训教育，缓解人类失业风险 45
（六）促进政策公平和广泛合作，抹平数字鸿沟风险 48
（七）构建指导原则，减少人类道德价值观受损风险 52
（八）展望未来伦理演进，探索 45°平衡律 56
四 、金砖国家扩展加盟，伦理安全展望新篇 59
（一）金砖国家应正视 AI 伦理风险 59
（二）金砖国家需合作应对共同挑战 60
参考文献 63
附录 66

前 言

2023 年，我国经济社会发展遭遇国内外多重超预期因素冲击，风险挑战前所未有。中国的数字经济克服了重重困难，取得了长足的发展，人工智能更是如耀眼的明珠，在各行各业中广泛应用，凸显了新质生产力对经济的推动、带动了人类文明的发展。人工智能的突破性成就带动了人工智能应用的快速蔓延，数字经济的快速增长有了新的引擎，通信产业等数字基础设施有了长足的进步。同时，数字经济也给人工智能技术进步带来新的需求和宽广的市场。

人工智能已经成为我们生活中不可或缺的一部分。从智能手机、自动驾驶汽车到医疗诊断和金融交易，AI 的应用已经深入到社会的各个角落。然而，人工智能技术的应用在给人们带来生活和工作便利的同时，也带来了一系列伦理风险与挑战。AI 伦理研究，就是在这样的背景下应运而生，旨在探索和定义 AI 的道德边界。

人工智能伦理研究是一个跨学科的领域，它探讨 AI 带来的伦理问题及风险，研究解决 AI 伦理问题，以促进 AI 向善并引领其健康发展。这种研究领域涉及到哲学、计算机科学、法律、经济等众多学科的交汇碰撞。

中国新一代人工智能发展规划推进办公室于 2019 年专门成立了新一代人工智能治理专业委员会，全面负责开展人工智能治理方面的政策体系、法律法规和伦理规范研究和工作推进。这体现了我国对人工智能伦理及其治理的密切关注程度和积极推进决心，同时也突显了这一问题的重要性。

前瞻研究和制定人工智能伦理规范有助于更好地发挥其“头雁”效应，对产业健康发展、人民幸福生活具有重要意义。在智能时代，人与机器的关系将会被重新定义。因此，国务院关于印发新一代人工智能发展规划的

通知中专门强调要探索建立无人驾驶、在线医疗、金融科技、智能配送等监管框架，完善相关法律法规和伦理审查规则。

国家新一代人工智能治理专业委员会 2021 年 09 月 25 日发布了《新一代人工智能伦理规范》，旨在将伦理道德融入人工智能全生命周期，为从事人工智能相关活动的自然人、法人和其他相关机构等提供伦理指引。

本文首先描述了 2023 年人工智能的数字化环境，这既是人工智能发展推动的结果，也是人工智能发展的广阔需求，人工智能的伦理问题，也在其具体的应用中发生了警示和碰撞。接下来本文介绍了常见的人工智能应用，呈现了 2023 年中国人工智能欣欣向荣的社会画卷。本文还介绍了人工智能伦理的内涵、研究范围、国外人工智能的发展情况、各国政策、智者和头部企业等在人工智能伦理方面的观点，以资参考、借鉴。

本文随后具体描述了伴随人工智能的应用产生的潜在人工智能风险，在应用性能效益和风险的二维平面上，我们审慎地认为，兼顾二者的 45 度角发展方向可能会是未来人类的选择，并提出了预防和应对的建议以供读者参考。

最后，考虑到金砖国家人工智能方兴未艾，提出了几点简单的建议，我们相信，在这飞快变化的时代，金砖国家正不断壮大发展，在世界政治经济的地位日益突出，在人工智能这个科技高地上的共识和合作，将有效规避人工智能的伦理风险，极大地促进金砖国家的经济增长和社会进步。

编写组成员

牵头编制单位：
中国联合网络通信有限公司研究院、中国联通国际有限公司

专家指导委员会：
唐永博 中国联合网络通信有限公司副总裁
孟树森 中国联通国际有限公司董事长兼总经理
李红五 中国联合网络通信有限公司研究院院长
陈辉 金砖国家未来网络研究院（中国·深圳）院长
魏进武 中国联合网络通信有限公司研究院副院长
张立仁 金砖国家未来网络研究院（中国·深圳）副院长

编写成员：
康旗、刘亚琪、金春花、吴钢、吕金宝、傅瑜瑜、余东海

参与编制单位：
金砖国家未来网络研究院中国分院

一、人工智能方兴未艾，风险防范未雨绸缪

2023 年人工智能（下文简称 AI）的快速发展，是人类生产力的一次跳跃。一方面，AI 给数字基础设施的建设带来直接的刺激和需求，数字经济的发展空间也得到大大提升，另一方面，数字经济的强大需求拉动了 AI 的研发，数字经济的广泛市场也给 AI 带来一望无垠的试验田，AI 的伦理问题研究就是在这样的背景下展开的。

(一) 数字经济发展孕育人工智能全面成长

2023 年是数字中国战略深入实施的关键一年，数字中国建设的基础更加稳固，赋能经济社会发展的效果显著。

1、数字基础设施扩容提速

网络基础设施进入提速升级新阶段。5G 网络覆盖面更广、技术更先进，并加速向重点场所深度延伸。截至 2023 年底，5G 基站数达 337.7 万个，同比增长 46.1%；平均每万人拥有 5G 基站 24 个，较上年末提高 7.6 个百分点；5G 移动电话用户数达 8.05 亿，在移动电话用户中占比 46.6%；5G 虚拟专网数量超 3 万个。基础电信企业 IP 骨干网、城域网、接入网 IPv6 改造全面完成，全国网络基础设施已全面支持 IPv6。截至 2023 年底，IPv6 活跃用户数达到 7.78 亿，移动网络 IPv6 流量占比达到 60.88%，固定网络 IPv6 流量占比达到 19.57%。

算力基础设施达到世界领先水平。截至 2023 年底，中国在用数据中心标准机架超过 810 万架，算力总规模达到 230EFLOPS，居全球第二位。8 个国家算力枢纽节点进入落地应用阶段。算力供给结构逐步优化，

包括数据中心、智算中心、超算中心等多种类型。截至 2023 年底，智能算力规模达到 70EFLOPS，增速超过 70%。全国累计建成国家级超算中心 14 个，全国在用超大型和大型数据中心达 633 个、智算中心达 60 个（AI 卡 500 张以上），智能算力占比超 30%。

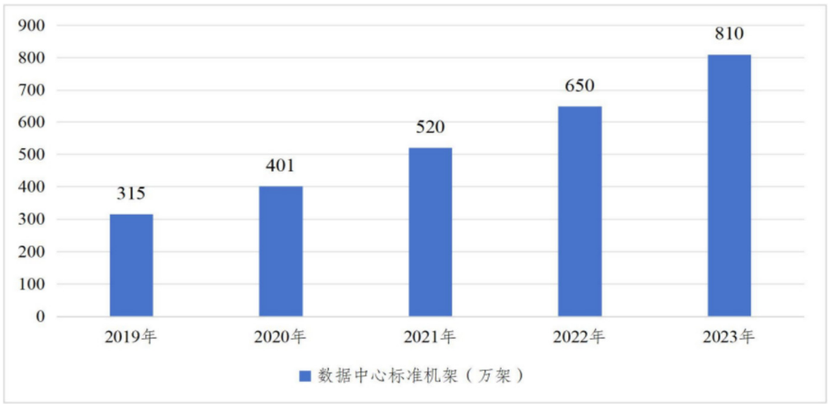


图 1 2019 年—2023 年数据中心发展情况

数据来源：工业和信息化部

2、数字技术创新活力涌现

基础数字技术能力持续增强。数字技术领域保持较高的创新热度。**先进计算、人工智能等关键核心技术不断取得突破。**高性能计算持续处于全球第一梯队。智能芯片、通用大模型等创新成果加速涌现。人形机器人进入提速发展阶段，专利申请数量增长较快。人工智能核心企业数量超过了 4500 家。**量子计算机、脑机接口等前沿技术的研发进度不断加快。**

算力网络实现阶段性跃升。积极落实国家东数西算战略，持续完善全国性算力网络布局，截至 2023 年底，三家基础电信企业为公众提供

的数据中心机架数达 97 万架，比上年末净增 15.2 万架，净增量是上年的近两倍，可对外提供的公共基础算力规模超 26EFlops(每秒万亿亿次浮点运算)。

构建人工智能（AI）基础设施体系。内部应用与外部赋能相结合探索 AI 应用新模式，推动“云网智”融合发展，全面向“AI+”战略转变。在网络规划 - 建设 - 维护 - 优化等网络运营各环节引入 AI 和大数据技术，提升网络智能化水平，减低网络故障率、节约基站能耗、优化信号覆盖。三家基础电信企业积极构建“1 个通用 +N 个专用”的通专结合大模型体系，通用大模型达千亿级参数级别，行业大模型可面向政务、应急、文旅等多个领域。

(二) 人工智能高歌猛进，融入百业大显身手

2023 年，人工智能领域发生了许多重大事件，深刻地影响了我们的生活和社会，2023 年被誉为“AI 时代开启元年”，人工智能技术的创新与应用突飞猛进。其中，生成式人工智能的出现是最具代表性的进展，例如 OpenAI 推出的 ChatGPT 订阅版、NVIDIA 的 BioNeMo 平台，以及 AWS 的‘AI Ready’培训计划。这些技术改变了人工智能技术与应用的发展轨迹，加速了人与 AI 的互动进程，是人工智能发展史上的新里程碑。特别是 ChatGPT 的火爆反应，让全球见证了人工智能技术的突破性发展和广泛应用。

以下是人工智能（AI）在全球范围内的应用取得的进展：

1、 AI 与各行各业的融合：AI 技术在金融、医疗、教育、交通等领

域的应用日益广泛，为各行业带来了更高效、智能的解决方案。

2、大模型的发展：随着深度学习技术的不断进步，大型神经网络模型在各领域取得了显著成果，提高了 AI 的准确性和泛化能力。

3、跨领域融合与创新：AI 技术与其他领域的融合进一步加深，如生物学、物理学、心理学等，推动了各行各业的创新发展。

4、强人工智能的发展：AI 逐渐向强人工智能发展，使得 AI 能够理解、思考和感知，甚至在某些方面超越人类的能力。

5、AI 与人类协作：AI 技术正逐渐与人类形成协同作战的局面，如在医疗、法律、教育等领域，AI 辅助专业人士进行工作，提高工作效率和质量。

总之，2023 年以来，AI 在全球范围内取得了重要进展，为人类带来了诸多便利和价值。

AI 能力一览表

序号	应用领域	说明
1	ChatGPT	主要应用在对话系统中，为用户提供问题解答和信息查询等服务。据统计，ChatGPT 的访问量高达 146 亿次，占总流量的 60% 以上。
2	GAN	主要用于图像生成，可以创建出逼真的人脸图像。
3	语音识别	将人的语音转化为文字，广泛应用于智能助手、语音输入等领域。
4	自然语言处理	使计算机能理解和生成人类语言，被广泛应用在搜索引擎优化、机器翻译等场景中。

5	推荐系统	通过分析用户的行为和偏好，为用户推荐他们可能感兴趣的商品或内容。
6	计算机视觉	让计算机能够“看”并理解图片或视频中的内容。
7	强化学习	让机器通过与环境的交互来学习如何做出最优决策。
8	语音合成	将文字转化为语音，常用于阅读器和智能助手等设备中。
9	机器翻译	实现不同语言之间的自动翻译。
10	人脸识别	通过分析人脸特征来识别身份。
11	自动驾驶	利用 AI 技术实现汽车的自动驾驶。
12	医疗诊断	利用 AI 对病症进行预测和诊断。
13	金融风控	使用 AI 进行风险评估和欺诈检测。
14	智能家居	利用 AI 技术实现家居设备的自动化控制。
15	游戏 AI	在游戏中实现 NPC（非玩家角色）的智能化。
16	教育 AI	应用于个性化教学和智能辅导等教育领域。
17	情感分析	通过分析文本中的情感倾向，了解人们对某一主题的态度和观点。
18	图像分割	将图像划分为多个区域，每个区域代表一个物体或背景。
19	目标跟踪	在视频中自动追踪目标的位置和运动轨迹。
20	AI 辅助设计	利用 AI 技术帮助设计师进行创作和设计。

表 1 AI 能力一览表

（三）中国 AI 快速应用，发展举世瞩目

2023 年以来，中国在人工智能领域取得了显著的进展和成就，不仅在技术研究上实现了重大突破，而且在产业发展和国际合作方面也展现出了强大的实力和影响力。

1、政策支持与战略规划

2023 年 02 月 26 日中国政府发布了《数字中国建设整体布局规划》，强调了人工智能在数字化转型中的重要性。

规划的主要目标：到 2025 年，基本形成横向打通、纵向贯通、协调有力的一体化推进格局，数字中国建设取得重要进展。到 2035 年，数字化发展水平进入世界前列，数字中国建设取得重大成就。

规划的主要内容：数字基础设施高效联通，数据资源规模和质量加快提升， 数字经济质量效益大幅增强， 政务数字化智能化，数字文化建设跃上新台阶， 数字社会精准化普惠化便捷化，数字生态文明建设取得积极进展，数字技术创新实现重大突破，数字安全保障能力全面提升。

《新一代人工智能发展规划》在 2017 年由国务院印发。由国家发展改革委、科技部、工业和信息化部等部门共同负责实施。该规划明确了我国新一代人工智能发展的战略目标、重点任务和保障措施，为推动 AI 技术的广泛应用和产业化提供了政策指导。《新一代人工智能发展规划》提出了“三步走”的立法规划，并在 2023 年 6 月 6 日发布的《国务院 2023 年度立法工作计划》中明确将《人工智能法》列入立法规划。2023 年 8 月，世界上第一个有关生成式人工智能的成文法《生成式人

工智能服务管理暂行办法》生效实施。

《关于加强科技伦理治理的意见》在 2022 年由中共中央办公厅、国务院办公厅印发，2023 年主要负责部门科技部进行了实施部署，同时涉及国家发展改革委、教育部、工业和信息化部等，他们共同参与科技伦理治理工作，针对 AI 技术在发展过程中可能面临的伦理风险，该意见提出了加强科技伦理治理的要求和措施，确保 AI 技术的健康发展。在战略规划方面，政府部门、信息化骨干企业、互联网大厂、运营商等纷纷出台了一系列规划方案，共同推动 AI 产业的发展。

政府部门：通过设立专项基金、建设创新平台、推动产学研合作等方式，加大对 AI 产业的扶持力度。同时，加强 AI 技术的监管和引导，确保其在合法合规的范围内发展。

信息化骨干企业：结合自身业务优势和技术积累，制定了一系列 AI 发展规划。例如，华为发布了多款具有自主知识产权的 AI 芯片和算法，推动了 AI 技术在多个领域的应用落地。

互联网大厂：凭借其庞大的用户基础和丰富的数据资源，在 AI 领域较早布局，投入较大。例如，阿里巴巴、腾讯等公司纷纷推出了自己的 AI 开放平台，为中小企业提供 AI 技术支持和服务。阿里巴巴宣布将投资 1000 亿元人民币建设全球最大的人工智能数据中心，

运营商：作为信息基础设施的建设者和运营者，运营商在 AI 领域也发挥着重要作用。他们通过优化网络架构、提升数据传输效率等方式，为 AI 技术的发展提供了良好的网络环境。

2、技术研究与创新

中国在 2023 年以来在 AI 领域取得了显著成就，包括大量生成式 AI 专利申请、大模型快速发展、AI 技术在各个行业广泛应用。在自然语言处理（NLP）、语音识别、图像处理、自动驾驶技术和医疗诊断等领域，AI 技术的准确率和应用水平不断提高，达到了商业化应用的水平。

据国家工业信息安全发展研究中心发布的相关报告指出，2014 年至 2023 年，中国发明人申请的生成式人工智能专利数量超过 3.8 万件，居世界领先行列。中国在全球生成式 AI 专利申请中占据主导地位，这一成就反映了中国在 AI 技术研发方面的强大实力和创新能力。

世界知识产权组织（WIPO）发布的最新报告显示，全球生成式人工智能相关的发明申请量达 54000 件，超过 25% 是在 2023 年出现的。其中，中国发明人申请的生成式人工智能专利数量最多，涵盖了从自动驾驶、出版到文档管理等广泛领域。全球范围内生成式 AI 专利申请量的激增，表明 AI 技术正处于快速发展期，且中国在这一全球趋势中扮演着重要角色。

2023 年，中国在人工智能领域取得了显著的技术突破，不仅在理论研究上取得了重大进展，而且在商业应用、社会服务等领域的融合也日益深入。包括商汤科技发布多模态大模型，华为展示智能产品与服务，中兴通讯推出星云编程大模型，中国气象局与清华大学联合研发的「风清」大模型。

这些技术突破不仅提升了中国在全球 AI 领域的竞争力，也为各行

各业提供了智能化升级的新路径，推动了社会经济的快速发展。

3、产业发展与应用

2023 年，中国国产大模型频频亮相、加速迭代。据腾讯科技报道显示，中国有至少 130 家公司研究大模型产品，其中 100 亿级参数规模以上的大模型超过 10 个，10 亿级参数规模以上的大模型已近 80 个，大模型数量位居世界第一梯队。中国大模型的发展速度惊人，数量众多，这反映了中国在 AI 技术领域的深厚积累和快速创新能力。

生成式 AI 已遍及生命科学、制造、交通、安全和电信等行业。具体而言，图像和视频数据类在生成式人工智能专利中占主导地位，其次是文本和语音 / 音乐类，分子、基因和蛋白质数据类的生成式人工智能专利增长迅速。生成式 AI 在不同行业中的应用广泛，显示出其强大的潜力和广泛的应用前景。特别是在生命科学和制造领域的应用，预示着 AI 技术将对这些行业产生深远影响。

艾瑞咨询的研究报告显示，中国人工智能产业规模达到 2137 亿元，预计到 2028 年将超过 8000 亿元，大模型带来的产业加成比例在 2028 年或达到 32.9%。AI 技术在办公、设计、传媒、法律、游戏、教育、汽车等多个领域加速应用和扩散。

4、国际合作与治理

在国际合作与治理方面，中国发挥了积极作用。中国提出的加强人工智能能力建设国际合作决议在联合国获得通过，显示了国际社会对加强 AI 领域国际合作的共识。此外，中国还与其他国家签署了《布莱切

利宣言》，这是一份针对人工智能的国际性声明，旨在关注 AI 技术的安全性。

中国在第三届“一带一路”国际合作高峰论坛期间提出《全球人工智能治理倡议》，围绕人工智能发展、安全、治理三方面系统阐述了人工智能治理的中国方案。中国在全球AI治理中的积极参与和提出的倡议，展示了中国作为负责任大国的形象，同时也反映了中国在 AI 技术发展方面的自信和实力。

当前，人工智能技术对法律、伦理和人道主义层面的冲击及其对国际政治、经济、军事、社会等领域带来的复杂影响，已引发国际社会的关注和讨论。全球 AI 治理面临的挑战包括技术、法律和伦理等多方面的复杂性，需要国际社会共同努力，制定有效的治理框架来应对这些挑战。

综上所述，2023 年，中国在 AI 领域的发展不仅体现在技术创新和产业应用上，还在于其在全球 AI 治理中的积极参与和贡献，展现了中国在全球科技舞台上的重要地位和市场潜力。

(四) 伦理规范有利于人工智能长治久安

1、人工智能发展带来一系列伦理风险与挑战

随着以 ChatGPT、Sora 等大模型为代表的新一代生成式人工智能技术取得突破性发展，颠覆性技术的广泛应用正牵引人类社会发生深刻变化，当前正朝向通用人工智能迈进。人工智能技术的应用在给人们带来工作和生活便利的同时，也不断暴露出一些风险隐患，如隐私泄露、

偏见歧视、算法滥用、安全问题等等，引发了社会各界广泛关注。根据经济合作与发展组织（OECD）发起组织的人工智能事件监测数据，仅2023 年 12 月单月，人工智能伦理事件就达 495 件，伦理问题也反向制约着人工智能的健康发展。频发的伦理风险问题引发了人们对未来人工智能技术是否能够继续坚守造福人类初心的质疑，甚至对人类自身造成威胁的担忧。

事件 1：比利时男子被 Al“怂恿 " 自杀身亡

2023 年 3 月 ，根据比利时《自由报》, 一名比利时男子皮埃尔 (化名) 自杀身亡。其妻子克莱尔 (化名) 称，皮埃尔是被一个名为“艾丽莎”的智能聊天机器人诱导走向死亡的。

事件 2：福州 AI 诈骗事件

2023 年 5 月 , 包头市公安局电信网络犯罪侦查局发布一起使用智能 AI 换脸和拟声技术，进行电信诈骗的案件，福州市某科技公司法人代表郭先生 10 分钟内被骗 430 万元。

事件 3：俄罗斯象棋比赛机器人夹断儿童手指

2022 年 7 月，据英国《卫报》报道，在莫斯科国际象棋公开赛上，象棋机器人在对弈过程中，突然抓住并折断对面七岁小男孩的手指。据悉，该男孩是莫斯科 9 岁以下最强的 30 位棋手之一。

事件 4：chatgpt 出卖个人隐私，被起诉赔偿 30 亿美元

2023 年 6 月 28 日，位于北加州的 Clarkson 律所代表 16 位匿名人士以及数以亿计的 ChatGPT 用户、ChatGPT API 用户、ChatGPT plus

用户、微软用户在加州北部地区巡回法院向 OpenAl 和微软提起集体诉讼，指控其在利益的驱使下，未经用户知情同意或用户不知情的情况下，使用从数亿互联网用户 (包括各个年龄段的儿童) 抓取的私人信息 (包括个人身份信息) 来创建其人工智能产品，严重侵犯了用户的财产权、隐私权等权益，给社会带来了潜在的灾难性风险，要求赔偿 30 亿美金。

事件 5：Facebook Al 种族歧视，误将黑人标记为灵长类动物

有 Facebook 用户在观看一段以黑人为主角的视频时收到推荐提示 : 询问他们是否愿意“继续观看有关灵长类动物的视频”。该视频是由《每日邮报》在 2020 年 6 月 27 日发布的 , 其中包含黑人与白人平民和警察发生争执的片段 , 并且视频内容与灵长类动物无关。

事件 6：AI 换脸侵权案

2023 年 6 月 9 日，杭州市萧山区人民检察院对一起利用“AI 换脸”技术侵犯公民个人信息的案件，依法向杭州互联网法院提起民事公益诉讼。2020 年左右，虞某某从互联网上接触到“AI 换脸”软件及技术，并于 2021 年开始，以牟利为目的，在未取得被编辑人同意的情况下，利用上述“AI 换脸”软件，将从互联网等渠道收集到的他人人脸信息与部分淫秽视频中的人脸信息进行替换合成，制作生成虚假的换脸淫秽视频，在网络社交软件上进行传播并获利。

在未来的发展中，人工智能伦理问题将继续存在并不断演变。我们需要加强跨学科的研究和合作，推动人工智能技术的可持续发展，同时也需要加强 AI 伦理研究，建立健全保障人工智能健康发展的法律法规、

制度体系、伦理道德，以确保人工智能技术的发展和应用符合人类的价值观和利益。

2、AI 伦理研究的重要性

目前，以深度学习为核心的新一代人工智能技术取得了极大的成功，大模型的发展令人工智能在下游任务的性能逐步趋近类人智能，并体现出极强的应用赋能潜力。人工智能新技术正在不断刷新着人们的认知极限，颠覆性地重塑着人类生产、生活方式，为人类社会实现人机协同、跨界融合、共创分享带来重大机遇，但同时也带来了隐私泄漏、偏见歧视、责权归属、技术滥用等伦理问题和挑战。因此，加强 AI 伦理研究，建立健全保障人工智能健康发展的法律法规、制度体系、伦理道德是至关重要的。

AI 伦理研究的重要性主要体现在以下几个方面：

防止隐私泄露。AI 系统通常需要访问大量个人数据来进行学习和优化。在这个过程中，个人隐私可能会受到侵犯。如何在收集和使用数据的同时，保护用户的隐私，是 AI 伦理研究需要解决的重要问题。

消除偏见与歧视。在训练 AI 模型时，使用的数据可能会反映出社会偏见和歧视。如果不加以纠正，不公平或偏见的决策会加剧社会不平等。例如，AI 系统可能会因为性别、种族等因素，对不同的人做出不同的决策。

防止技术滥用。AI 技术的发展和应用，可能会被恶意使用，例如用于制造深度伪造视频、网络攻击等。AI 伦理研究需要防止这种情况的发生。

提高透明度与可解释性。AI 系统通常被视为“黑匣子”，难以理解其决策过程。这使得难以追踪和解释 AI 系统的错误或偏见。为了增强透明度和可解释性，AI 应用需要开发可解释的算法，并提供用户和监管机构访问决策过程的途径。

规范数据质量与多样性。确保训练数据的质量和多样性是解决伦理问题的第一步。数据应该经过严格筛选和清理，以减少潜在的偏见。此外，应该采用多样化的数据来源，以避免单一视角的影响。

合理监管加强法治。政府和监管机构应该制定严格的法规，要求 AI 开发者和运营者遵守伦理标准。这些法规应该包括数据隐私保护、公平性要求、透明度标准等，以确保 AI 系统的道德使用。

AI 伦理是确保人工智能技术能够为社会带来益处而不是伤害的关键要素。我们必须认真对待这些问题，采取积极的措施来解决潜在的风险和挑战。只有在伦理问题得到妥善处理的前提下，AI 技术才能够持续地为人类社会的进步和发展作出贡献。

3、 AI 伦理研究的主要内容

人工智能伦理是一个研究人工智能伦理问题的新兴的跨学科领域。人工智能伦理学主要研究与人工智能相关的伦理理论、指导方针、政策、原则、规则和法规。具体来说，人工智能伦理领域的研究内容包括但不限于以下几个方面：

1） 人工智能伦理风险的阐释和评估。分析并预测 AI 技术可能带来的社会、经济以及道德风险。

2）伦理准则和治理原则的研究。形成包括教育改革、伦理规范、技术支撑、法律规制、国际合作在内的多维度伦理治理框架。

3）制定专门的 AI 伦理原则和规则。比如尊重个人自主权，保护可能由于疾病、智力障碍、年龄限制等原因所导致自主权减弱的个人。

4）跨国比较研究。通过比较不同国家对 AI 的监管政策和实践，了解各国在 AI 伦理方面的优先关注点和发展趋势。

5）人工智能与就业。人工智能的发展可能会导致大量工作岗位被取代，从而引发社会就业问题。因此需要讨论和研究人工智能对就业的影响和应对之策。

6）人工智能军事化部署。随着人工智能技术在军事领域的应用，涉及到军事化也带来了一系列道德问题。例如无人机的使用、自动化武器的研发等，这些都引发了人工智能伦理的深刻思考和讨论。

总的来说，AI 伦理研究的主要目标是确保 AI 技术的发展和应用能够符合社会公众利益和道德价值，避免可能带来的负面影响。本报告主要侧重于 AI 伦理风险的研究，其余的部分也有一定程度的涉及。

4、AI 伦理研究面临的挑战

尽管 AI 伦理研究的重要性已经被广泛认识，但是在实际的研究过程中，还面临着许多挑战：

1） 伦理观念多样性。让人工智能大模型的能力和行为跟人类的价值、真实意图和伦理原则相一致，才能确保人工智能与人类协作中的安全、互信和可靠。但自然人的价值观因国家、民族、宗教、文化、性别、

道德、个人观念等不同而各异，这使得在全球范围内形成统一的 AI 伦理原则和规范变得困难。如何在尊重各国文化传统的基础上，寻求共识并制定普遍适用的 AI 伦理原则是一个重要挑战。

2）技术发展迅速带来的不确定性。随着人工智能技术的不断发展和进步，新的应用场景和技术手段不断涌现，给伦理研究带来了新的挑战。例如，深度学习、神经网络等技术的发展使得 AI 系统越来越复杂，难以理解和解释，这给伦理评估和监管带来了困难。

3）法律法规滞后。现有的法律法规往往难以适应快速发展的 AI 技术，导致很多伦理问题无法得到有效解决。如何制定和完善相关法律法规，以适应 AI 技术的发展和应用，是一个重要的挑战。

4）数据隐私和安全保护产生的障碍。AI 技术需要大量的数据来训练模型，但数据的收集和使用往往涉及到个人隐私和安全问题。一些不法分子可能会利用 AI 技术进行数据泄露、身份盗窃等犯罪活动。例如，某大型科技公司曾因未充分告知用户其数据被用于 AI 模型训练而面临隐私诉讼。

5）算法歧视。AI 算法在处理数据时可能会存在偏见，导致不公平的结果。如何在 AI 系统的设计和开发过程中充分考虑公平性问题，避免歧视现象的发生，是一个重要挑战。

6）责任和问责制如何界定。当 AI 系统出现错误或误判时，如何确定责任归属是一个复杂的问题。由于 AI 系统决策过程的复杂性和缺乏透明度，责任归属成为一个争议焦点。

二、他山之石可以攻玉，先进经验学习借鉴

如何应对 AI 伦理的风险和挑战，我们可以借鉴一下国际主流的做法，尤其是处于领先地位的美国。

2024 年 6 月 28 日星期五，由 OpenAI 首席执行官 Sam Altman 和 Operation HOPE 首席执行官 John Hope Bryant 创立的人工智能伦理委员会在亚特兰大举行了首次会议。该委员会由 Altman 和 Bryant 于去年春天在克拉克亚特兰大大学共同发起的倾听之旅发展而来，旨在确保传统上代表性不足的社区能够在整个人工智能的发展中拥有发言权，帮助构建围绕该技术的人性和伦理考虑，以及广泛参与人工智能的经济机会。

作为引领科技革命和产业变革的战略性技术，人工智能成为国际竞争新焦点，自 2016 年以来，中国、美国、欧盟等国家或地区陆续发布人工智能发展和治理相关政策法规。2022 年，以 ChatGPT 为代表的生成式人工智能引发新一轮人工智能革命，其可能引发的生成虚假有害信息、侵犯知识产权等风险再次引发全球对人工智能治理的关注。

AI 技术的迅猛发展，使得其在医疗、教育、交通等多个领域展现出巨大的应用潜力。然而，随之而来的隐私保护、数据安全、就业市场变动、道德伦理等问题也日益凸显。在此背景下，国际社会纷纷投入大量资源，深入研究 AI 伦理问题，以期在技术发展的同时，确保人类社会的和谐与稳定。

(一) 国际 AI 伦理研究现状

1、理论框架的构建

近年来，国际上众多学者致力于构建 AI 伦理的理论框架，以期为实践提供指导。这些理论框架主要围绕 AI 技术的责任归属、道德判断以及价值观等问题展开。例如，部分学者提出了以“公平性、透明性、可解释性、数据隐私”为核心的 AI 伦理原则，强调在 AI 系统的设计、开发和部署过程中，应充分考虑这些原则。

2、实践应用的探索

在实践应用方面，各国政府、企业和研究机构纷纷开展 AI 伦理相关的项目与实验。例如，欧盟推出的《通用数据保护条例》(GDPR)，明确规定了在 AI 处理个人数据时，应遵循的隐私保护原则；谷歌、微软等科技巨头则在其内部建立了 AI 伦理委员会，负责审查 AI 技术的道德风险。

3、跨学科研究的推动

AI 伦理问题的复杂性决定了其需要跨学科的解决方案。目前，国际上许多研究团队正积极促进伦理学、社会学、法学、计算机科学等多学科的交叉融合，共同研究 AI 伦理问题。这种跨学科的研究模式不仅有助于深入理解 AI 技术的本质及其社会影响，还为寻求有效的解决策略提供了更多可能性。

(二) 国际 AI 伦理研究贡献及未来走向

1、 国际 AI 伦理研究贡献

1）提升公众对 AI 伦理问题的认识

通过广泛的研究与讨论，国际社会成功提升了公众对 AI 伦理问题的认识。如今，越来越多的人开始关注 AI 技术背后的道德与法律问题，积极参与到相关讨论中来。这种社会共识的形成为 AI 技术的健康发展提供了有力的社会支持。

2）为政策制定提供理论支持

国际上的 AI 伦理研究成果为各国政府的政策制定提供了有力的理论支持。许多国家在制定 AI 产业政策时，都充分考虑了伦理因素，以确保技术进步的同时，维护社会稳定和公共利益。例如，中国的《新一代人工智能发展规划》中明确提出要加强人工智能伦理研究，确保人工智能安全、可靠、可控。

3）推动 AI 技术的可持续发展

AI 伦理研究的深入进行有助于推动 AI 技术的可持续发展。通过不断反思与调整技术发展的方向，国际社会有望避免潜在的道德风险，确保 AI 技术在为人类创造福祉的同时，不会带来负面影响。例如，在自动驾驶汽车的研发过程中，工程师们就积极引入伦理原则，以确保在紧急情况下能够做出符合道德要求的决策。

2、国际 AI 伦理研究未来走向

尽管国际上的 AI 伦理研究已取得显著成果，仍面临诸多挑战。未来，研究继续关注以下方面：一是加强跨学科的交流与合作，共同推动 AI 伦理研究的深入发展；二是关注新兴技术领域中的伦理问题，如人工智能在医疗、教育、金融等领域的应用；三是加强国际合作，共同制定

全球性的 AI 伦理原则和标准。

国际上在 AI 伦理领域的研究已取得丰富成果和显著成就。这些成果不仅为 AI 技术的健康发展提供了有力保障，还为人类社会的和谐与稳定做出了积极贡献。展望未来，我们期待国际社会能够继续深化 AI 伦理研究，共同推动人工智能技术的可持续发展。

(三) 全球主要国家 AI 政策

1、美国：以发展为导向开展 AI 风险管理，力争保持全球领头羊地位

1）重发展轻监管

相比于监管，美国更关注 AI 产业的创新发展。早在 2016 年，美国密集发布《为人工智能的未来做好准备》、《国家人工智能研究与发展战略计划》等，将 AI 提至国家战略的高度，此后持续表现出鼓励发展的态度，从资金投入、资源开放、人才培养等各方面给予政策支持。而直到 2020 年，美国才开始陆续发布治理为主的政策文件，如《人工智能监管原则草案》、《在联邦政府中促进使用可信赖的人工智能》等，但相关政策并不强调严格监管，避免因强监管阻碍 AI 创新发展。

2）AI 监管尚未全面立法

美国目前尚未完成 AI 监管全面立法，相比于出台具有极强约束力的法律文件，更侧重于提出 AI 监管原则和框架，如 2020 年发布的《人工智能应用监管指南》提出 10 项 AI 监管原则，2022 年发布的《人工智能权利法案蓝图》提出负责任地使用 AI 路线图，2023 年发布的《人

工智能风险管理框架》提供了可参考的 AI 风险管理框架。2023 年 9 月，美国第一个严格、全面的立法蓝图——《两党人工智能立法框架》被提出，尚未被国会通过。

2、欧盟：以安全为导向开展 AI 监管立法，抢占全球伦理规则主导权

1）强调制定用于 AI 监管的法案

欧盟作为偏重国家间统一规范的组织，期望制定明确的法案来对 AI 进行监管。2016 年以来，欧盟围绕 AI 先后出台多部强调监管、数据规范、边界的政策文件。2021 年，欧盟委员会提出全球首部监管 AI 的法律草案——《人工智能法案》，旨在为 AI 引入一个共同的监管和法律框架，此后该法律草案经多轮修改，近期又针对横空出世的生成式 AI 不断完善内容，预期 2024 年前后通过。

2）采用风险分级监管治理框架与全过程风险控制体系

《人工智能法案》采用分级分类的风险规制路径，将 AI 系统可能引发的伤害人类安全、基本权利的风险划分为四个等级，分别是极低风险、有限风险、高风险和不可接受的风险，不同级别对应不同的监管措施，如完全禁止不可接受风险、重点限制高风险等。此外，《人工智能法案》针对高风险 AI 系统进入市场前后的完整流程进行了风险规制，覆盖范围包括设计开发、研发设计、产品评估、注册、CE 标志、产品监测，如在设计开发阶段建立和维护风险管理系统、入市后建立和记录售后监测系统等。

世界主要国家 AI 政策		
1	美国	美国政府高度重视 AI 技术的发展，将其视为国家战略。2019 年，美国发布了《美国人工智能研究和开发战略》，旨在确保美国在 AI 领域的领导地位。此外，美国政府还通过各种资金支持和政策引导，推动 AI 技术的创新和应用。
2	中国	中国政府将 AI 技术视为国家战略，制定了一系列政策和规划，以推动 AI 产业的发展。例如，2017 年，国务院发布了《新一代人工智能发展规划》，明确提出了到 2030 年，中国要成为世界人工智能创新中心的目标。
3	欧盟	欧盟将 AI 技术视为关键基础设施和竞争力的重要组成部分。2018 年，欧盟发布了《欧洲人工智能战略》，旨在通过政策、投资和国际合作等手段，推动 AI 技术的发展和应用。
4	日本	日本政府高度重视 AI 技术的发展，将其视为实现经济增长和社会福祉的关键因素。2017 年，日本政府发布了《人工智能技术战略》，提出了加强 AI 研究、人才培养和国际合作等方面的政策措施。
5	德国	德国政府将 AI 技术视为国家战略，制定了一系列政策和规划，以推动 AI 产业的发展。例如，2018 年，德国政府发布了《德国联邦政府人工智能战略》，明确了德国在 AI 领域的发展目标和政策措施。
6	英国	英国政府高度重视 AI 技术的发展，将其视为提高国家竞争力和实现可持续发展的关键因素。2018 年，英国政府发布了《英国人工智能战略》，提出了加强 AI 研究、人才培养和产业发展等方面的政策措施。
7	韩国	韩国政府将 AI 技术视为国家战略，制定了一系列政策和规划，以推动 AI 产业的发展。例如，2019 年，韩国政府发布了《韩国人工智能发展战略》，明确了韩国在 AI 领域的发展目标和政策措施。
8	印度	印度政府高度重视 AI 技术的发展，将其视为提高国家竞争力和实现经济增长的关键因素。2018 年，印度政府发布了《印度国家人工智能战略》，提出了加强 AI 研究、人才培养和产业发展等方面的政策措施。
9	加拿大	加拿大政府将 AI 技术视为国家战略，制定了一系列政策和规划，以推动 AI 产业的发展。例如，2017 年，加拿大政府发布了《加拿大人工智能战略》，明确了加拿大在 AI 领域的发展目标和政策措施。

10	澳大利亚	澳大利亚政府高度重视 AI 技术的发展，将其视为提高国家竞争力和实现经济增长的关键因素。2018 年，澳大利亚政府发布了《澳大利亚国家人工智能战略》，提出了加强 AI 研究、人才培养和产业发展等方面的政策措施。

表 2 世界主要国家 AI 政策

这些政策和措施有助于保护个人隐私、促进公平公正、确保 AI 系统的透明度和可解释性，以及明确责任分配。

不足之处是伦理规范与法律、标准等衔接不畅，可能导致伦理规范缺乏可操作性。此外，如何在促进创新和保护隐私之间找到平衡点，仍是各国面临的挑战

综上所述，AI 伦理治理是一个全球性的挑战，需要各国政府、企业和研究机构共同努力，通过制定和实施有效的政策和措施，确保 AI 技术的健康、安全和可持续发展。

(四) 全球智者的观点

自 2023 年以来，随着人工智能技术的飞速进步，全球范围内的智者、学者、政策制定者和企业家纷纷就 AI 伦理问题展开深入探讨。他们的观点虽不尽相同，但共同构成了一个多维度的 AI 伦理讨论框架。

首先，众多智者强调数据隐私的重要性。他们认为，在 AI 系统的训练和运营过程中，个人数据的收集和使用必须遵循严格的隐私保护法规。这包括确保数据的匿名化、透明化处理，以及给予用户明确的数据使用授权。此外，一些专家还提出了“数据主权”的概念，即个人应有权决定其数据如何被使用，甚至有权撤回这些数据。

其次，AI 的公平性和不歧视问题也受到了广泛关注。专家们指出，

AI 系统可能会因为训练数据的偏见而产生歧视性决策，如性别、种族或年龄歧视。因此，设计和开发 AI 系统时，必须确保算法公正、无偏见，并能够适应多样化的应用场景。

再者，AI 的透明度和可解释性成为热议焦点。许多专家认为，AI 系统的决策过程应尽可能透明，以便人们理解其背后的逻辑和原因。这不仅可以增强公众对 AI 的信任，还有助于在出现问题时及时定位和纠正。此外，AI 的安全性和责任归属也是讨论的重点。随着 AI 技术在各个领域的广泛应用，如何确保 AI 系统的安全运行，防止恶意攻击和数据泄露，成为了一个亟待解决的问题。同时，当 AI 系统导致损害时，如何界定责任归属，是归咎于开发者、用户还是 AI 本身，这也是一个复杂而棘手的伦理和法律问题。

最后，关于 AI 与人类的关系，智者们的观点也呈现出多样性。一些人认为，AI 的发展将导致大规模的劳动力市场变革，人类需要适应这种变化，并与 AI 建立合作关系。另一些人则警告说，过度依赖 AI 可能导致人类的自主性降低，因此应保持警惕，并寻求在技术进步中保持人类的主体地位。

从整体上看，2023 年以来全球智者关于 AI 伦理的观点体现了对技术进步与社会责任之间平衡的深刻思考。这些观点不仅涵盖了数据隐私、公平性、透明度等传统伦理问题，还触及了安全性、责任归属以及人机关系等前沿议题。

这些讨论不仅丰富了 AI 伦理的理论体系，也为政策制定和实践应

用提供了重要参考。然而，也应看到，这些观点之间仍存在分歧和争议，这反映了 AI 伦理问题的复杂性和多维度性。未来，随着 AI 技术的进一步发展和应用，这些伦理问题将继续面临挑战和变革，需要全球范围内的智者持续关注和深入探讨。

综上所述，2023 年以来的 AI 伦理讨论展现了人类对科技进步的审慎态度和对社会责任的担当。这些讨论不仅为 AI 的未来发展指明了方向，也为我们如何更好地与智能机器共存提供了思考基础。

全球智者的观点集萃		
1	埃隆・马斯克 (Elon Musk)	特斯拉和 SpaceX 的创始人，对 AI 技术持谨慎态度。他曾多次警告称，人工智能可能对人类构成威胁，甚至可能导致人类灭亡。
2	马克・扎克伯格 (Mark Zuckerberg)	Facebook 创始人兼 CEO，对 AI 技术持积极态度。他认为 AI 将带来许多创新和便利，但同时也需要关注其潜在的隐私和安全问题。
3	杰弗里・辛顿 (Geoffrey Hinton)	著名计算机科学家，被称为“深度学习之父”。他对 AI 技术充满信心，认为 AI 将彻底改变人类生活。
4	吴恩达 (Andrew Ng)	斯坦福大学教授，谷歌大脑创始人。他对 AI 技术持乐观态度，认为 AI 将带来巨大的经济和社会价值。
5	尼克・博斯特罗姆 (Nick Bostrom)	牛津大学哲学家，研究人工智能伦理问题。他认为 AI 技术的发展将对人类社会产生深远影响，需要制定相应的伦理规范。
6	尤瓦尔・赫拉利 (Yuval Noah Harari)	以色列历史学家，《人类简史》作者。他对 AI 技术持谨慎态度，认为 AI 可能导致大量失业和社会不平等问题。
7	彼得・泰尔 (Peter Thiel)	PayPal 创始人，风险投资家。他对 AI 技术持积极态度，认为 AI 将带来巨大的商业机会。
8	雷・库兹韦尔 (Ray Kurzweil)	著名发明家和未来学家，对 AI 技术充满信心。他预测，到 2045 年，人工智能将达到与人类智能相当的水平。
9	杨振宁	中国著名物理学家，诺贝尔物理学奖获得者。他对 AI 技术持谨慎态度，认为 AI 可能对人类构成威胁，需要加强监管和伦理规范。
10	李开复	创新工场创始人，曾是谷歌大中华区总裁。他对 AI 技术持乐观态度，认为 AI 将带来巨大的经济和社会价值，但同时也需要关注其潜在的隐私和安全问题。

表 3 全球智者的观点集萃

(五) 全球头部企业的选择和态势

近年来，随着人工智能（AI）技术的迅猛发展，全球著名企业纷纷将 AI 伦理纳入其产品研发、投资、运营和市场策略的核心考量。这不仅是对社会期望的回应，也是企业自身长期发展的必要条件，从市场表现来看，这些企业都显现出积极的态势。

1、企业策略的重点概括：

1）产品研发方向

强调数据隐私与安全：在产品设计初期，企业便着重考虑数据收集、存储和处理的安全性。例如，苹果公司在其设备中内置了强大的加密功能，以保护用户数据不被未经授权的访问。

促进算法公平性与无偏见：企业致力于开发和部署减少算法偏见的算法，确保 AI 决策不会因种族、性别或其他特征而对某些群体产生不利影响。谷歌的“公平性指标”项目就是这一努力的体现。

增强透明度和可解释性：为了提高用户信任，企业正在努力提高 AI 系统的透明度。微软推出了“解释性 AI”工具，帮助用户理解机器学习模型的决策过程。

2）投资策略

加大伦理技术研发投入：企业不仅在传统 AI 技术上投资，还特别重视那些能够解决伦理问题的创新技术。亚马逊的 Alexa 基金就投资了多个致力于开发更加人性化、符合人类伦理的 AI 技术的初创公司。

建立伦理风险评估体系：在投资决策过程中，企业开始引入伦理风险评估，确保所投资的项目不仅符合经济效益，也符合社会伦理标准。

3）运营管理

制定内部伦理准则：许多企业制定了详细的内部 AI 伦理准则，为员工提供明确的指导。IBM 的 AI 伦理准则就是一个典型的例子，它要求所有员工在进行 AI 相关工作时遵守一系列伦理原则。

设立伦理委员会：为了确保 AI 项目的伦理合规性，一些企业成立了专门的伦理委员会，负责审查和监督 AI 项目的实施。例如，微软设立了一个人工智能伦理委员会，负责审查公司的 AI 项目是否符合伦理标准。

4）市场策略

塑造负责任的品牌形象：企业通过公开强调其 AI 产品的伦理特性，努力塑造负责任的品牌形象。这不仅能吸引那些关心技术伦理的消费者，还能提升企业的社会声誉。

推动行业合作与标准制定：企业积极参与行业合作，共同推动 AI 伦理标准的制定。例如，谷歌、微软和亚马逊等企业联合成立了“人工智能伦理联盟”，旨在推动行业内的伦理讨论和标准制定。

2、企业策略实际效果

在 AI 伦理的指导下，全球著名企业纷纷调整其产品研发、投资、运营和市场策略，以期在追求技术创新的同时，兼顾社会责任和道德标准。

从实际效果来看，这些企业在考虑 AI 伦理的基础上制定的策略已经取得了一定的成效。一方面，这些策略有助于提升企业的品牌形象和社会声誉，吸引更多的人才和客户。另一方面，这些努力也有助于推动整个行业的健康发展，减少技术滥用和伦理风险的可能性。

1）提升企业形象与品牌价值

企业将 AI 伦理纳入其核心策略，显著提升了其品牌形象。在公众对 AI 技术的道德和安全问题日益关注的背景下，企业的这些举措展示了其对社会责任的承担，增强了消费者的信任和忠诚度。例如，谷歌通过其“AI for Good”项目，展示了其在利用 AI 解决全球性问题上的努力和承诺，有效提升了其品牌价值。

2）促进技术创新与可持续发展

在 AI 伦理的指引下，企业更加注重技术的可持续性和创新性。通过投资研发更加透明、可解释和公平的 AI 算法，企业不仅解决了技术上的难题，也为其长期发展奠定了坚实的基础。此外，这些技术创新还有助于企业开拓新的市场和应用领域，如医疗健康、教育等，从而实现业务的多元化和可持续发展。

3）增强内外部信任与合规性

随着 AI 技术的广泛应用，数据隐私和安全问题日益凸显。企业通

过制定严格的内部伦理准则和设立伦理委员会，加强了对 AI 技术的监管和审查，确保了其业务运营符合相关法律法规的要求。这不仅增强了投资者、合作伙伴和社会公众对企业及其产品的信任，也有助于降低因违规操作而引发的法律风险。

4）推动行业合作与标准制定

企业在 AI 伦理方面的积极行动不仅限于自身，还推动了整个行业的合作与标准制定。通过参与或主导国际 AI 伦理倡议和组织，企业与其他利益相关者共同探讨了 AI 技术的道德、法律和社会影响，为推动全球范围内的 AI 伦理规范的形成发挥了积极作用。

3、企业策略的社会影响与面临挑战

企业的 AI 伦理策略在社会层面产生了广泛的影响。一方面，它提高了公众对 AI 技术的认知和理解，促进了科技与社会的良性互动。另一方面，它也激发了公众对 AI 伦理问题的关注和参与，推动了社会对 AI 技术道德和法律边界的深入探讨。面对不断变化的伦理挑战和市场需求，企业仍需持续改进和创新其策略，以实现更加可持续和负责任的发展。

尽管企业在 AI 伦理策略上取得了一定的成效，但仍面临诸多挑战。技术的快速发展使得新的伦理问题不断涌现，要求企业持续跟进和适应。此外，不同文化和社会背景下对伦理的理解存在差异，这要求企业在国际化战略中更加审慎。因此，企业需要建立一个动态的伦理管理机制，以应对不断变化的伦理挑战。

全球 10 大头部企业市场态势		
1	谷歌	特斯拉和 SpaceX 的创始人，对 AI 技术持谨慎态度。他曾多次警告称，人工智能可能对人类构成威胁，甚至可能导致人类灭亡。
2	亚马逊	Facebook 创始人兼 CEO，对 AI 技术持积极态度。他认为 AI 将带来许多创新和便利，但同时也需要关注其潜在的隐私和安全问题。
3	微软	著名计算机科学家，被称为“深度学习之父”。他对 AI 技术充满信心，认为 AI 将彻底改变人类生活。
4	IBM	斯坦福大学教授，谷歌大脑创始人。他对 AI 技术持乐观态度，认为 AI 将带来巨大的经济和社会价值。
5	Facebook	牛津大学哲学家，研究人工智能伦理问题。他认为 AI 技术的发展将对人类社会产生深远影响，需要制定相应的伦理规范。
6	苹果	以色列历史学家，《人类简史》作者。他对 AI 技术持谨慎态度，认为 AI 可能导致大量失业和社会不平等问题。
7	三星	PayPal 创始人，风险投资家。他对 AI 技术持积极态度，认为 AI 将带来巨大的商业机会。
8	华为	著名发明家和未来学家，对 AI 技术充满信心。他预测，到 2045 年，人工智能将达到与人类智能相当的水平。
9	阿里巴巴	利用 AI 技术优化电商平台的推荐系统、物流系统等。2019 年，其 AI 相关业务的收入为 56.6 亿美元，占公司总收入的 3.3%。
10	腾讯	发展游戏、社交等领域的 AI 技术，提升产品质量和用户体验。2019 年，其 AI 相关业务的收入为 77.5 亿美元，占公司总收入的 5.6%。

表 4 全球 10 大头部企业市场态势

综上所述，全球著名企业在考虑 AI 伦理的基础上制定的策略在推动技术创新、提升企业形象和促进社会和谐方面发挥了积极作用，在市场占有和收入上都有不俗的表现。未来，随着 AI 技术的进一步发展和应用，这些企业有望在伦理规范的引导下实现更加可持续和负责任的发展。

三、居安思危如履薄冰，杜绝风险伦理安全

(一) 伴随社会和技术进步，AI 伦理安全理念与时俱进

AI 伦理安全的发展变化是一个随着技术进步和社会需求不断演进的过程。以下是 AI 伦理观点随时间顺序发展的简要概述：

1、早期阶段（20 世纪 50 年代 -80 年代)

在人工智能的早期阶段，AI 系统主要局限于简单的规则和逻辑运算。此时的伦理讨论主要集中在计算机程序的道德责任问题上。例如，如果一个程序导致了意外的结果，责任应该归咎于谁——程序员、用户还是机器本身？这一时期的讨论为后来 AI 伦理的发展奠定了基础。

2、专家系统时代（1980 年代 -90 年代)

随着专家系统的兴起，AI 开始被用于具体的决策支持。这一时期的伦理关注点转向了算法的偏见和歧视问题。专家系统往往依赖于大量的专家知识和数据，而这些数据和知识可能存在隐性偏见。因此，伦理学家开始呼吁对 AI 系统的输入数据进行审查和验证，以确保决策的公正性。

3、机器学习时代（21 世纪初至今)

进入 21 世纪，随着机器学习特别是深度学习的发展，AI 系统的能力得到了显著提升。这一时期的伦理讨论变得更加广泛和深入：

数据隐私：随着大数据技术的应用，个人数据的收集和分析成为 AI 训练的关键。这引发了关于数据隐私和用户授权的广泛讨论，推动了如欧盟《通用数据保护条例》(GDPR) 等法规的制定。

透明度和可解释性：深度学习模型因其“黑箱”特性而受到批评，

因为它们的工作原理对人类来说是不透明的。这促使研究者开发新的算法和技术，以提高 AI 决策的透明度。

自动化与就业：随着自动化程度的提高，AI 对就业市场的影响成为热议话题。一些人担心 AI 将取代大量工作，而另一些人则认为 AI 将创造新的就业机会。

安全性与对抗性攻击：随着 AI 系统在各个领域的部署，如何防止恶意攻击和确保 AI 的安全性成为关注焦点。对抗性样本攻击的研究揭示了 AI 系统的脆弱性，促使研究者探索更强的防御机制。

人机关系：随着 AI 技术的发展，人们开始更深入地思考人类与机器的关系。这包括讨论 AI 的自主性、人类的角色以及 AI 在决策过程中的道德地位。

(二) 加强监管，防范失控性风险

失控性风险指的是AI系统失去控制并产生不可预测的行为的风险。解决办法是通过技术创新和伦理规范，加强对 AI 系统的监管和管理，以确保其在人类的控制之下运行。

随着人工智能 (AI) 技术的快速发展，AI 系统在各个领域的应用越来越广泛，如自动驾驶、医疗诊断、金融投资等。然而，AI 系统的发展也带来了一系列风险，其中失控性风险是最为严重的一个。这种风险可能导致严重的后果，如人身伤害、财产损失、社会不安等。为了确保 AI 系统在人类的控制之下运行，我们需要通过技术创新和伦理规范，加强对 AI 系统的监管和管理。

1、失控性风险的表现

1）不可预测的行为：AI 系统在运行过程中可能出现无法预测的行为，这些行为可能是由于算法的缺陷、数据的不完整或者训练过程的问题导致的。例如，在自动驾驶领域，如果 AI 系统无法正确识别交通信号灯的颜色，可能会导致交通事故的发生。

2）自我学习与进化：AI 系统具有自我学习和进化的能力，这意味着它们可以不断地从数据中学习新的知识和技能。然而，这种能力也可能导致 AI 系统产生意想不到的行为。例如，AI 系统可能会学会如何在没有人类干预的情况下进行攻击，从而导致安全风险。

3）恶意攻击：AI 系统可能成为黑客攻击的目标，黑客可以利用 AI 系统的漏洞对其进行操控，使其产生恶意行为。例如，黑客可以通过对 AI 系统进行训练，使其学会如何破解密码、窃取信息等。

4）误判与歧视：AI 系统在处理数据时可能会出现误判和歧视现象。例如，在招聘领域，AI 系统可能会因为训练数据的偏见而拒绝某些求职者的申请。这种误判和歧视不仅会导致个人的损失，还可能加剧社会的不公平现象。

2、失控性风险的原因

1）算法缺陷：AI 系统的运行依赖于算法，算法的缺陷可能导致 AI 系统产生不可预测的行为。例如，深度学习算法中的梯度消失问题可能导致 AI 系统无法正确地学习知识。

2）数据不完整：AI 系统的训练需要大量的数据，数据的质量直接

影响到 AI 系统的性能。如果训练数据不完整或者存在偏差，可能导致 AI 系统产生错误的判断和行为。

3）训练过程的问题：AI 系统的训练过程可能存在问题，如过拟合、欠拟合等。这些问题可能导致 AI 系统无法正确地学习知识，从而产生不可预测的行为。

4）人为因素：AI 系统的设计和开发过程中可能存在人为因素，如设计者的偏见、技术水平不足等。这些因素可能导致 AI 系统产生不可预测的行为。

3、解决失控性风险的办法

1）技术创新：通过技术创新，提高 AI 系统的稳定性和可靠性。例如，研究新的算法、优化训练过程等。此外，还可以通过技术手段对 AI 系统进行监控和限制，确保其行为在可接受的范围内。

2）伦理规范：制定和完善 AI 领域的伦理规范，引导 AI 系统的设计和开发遵循道德和法律的要求。例如，制定关于数据隐私、算法公平等方面的规定，确保 AI 系统的使用不会侵犯他人的利益。

3）监管和管理：加强对 AI 系统的监管和管理，确保其在人类的控制之下运行。例如，建立专门的监管机构，对 AI 系统的开发和应用进行审查和监督；制定相关的法律法规，对违法行为进行处罚；加强国际合作，共同应对 AI 带来的挑战。

4）公众教育：提高公众对 AI 风险的认识和理解，培养公众的安全意识和防范能力。例如，开展 AI 安全教育，让公众了解 AI 系统的工作

原理和潜在风险；鼓励公众参与 AI 系统的监督和管理，共同维护社会的安全和稳定。

4、案例分析

1）AlphaGo：AlphaGo 是一款由谷歌 DeepMind 公司开发的围棋 AI 程序。在 2016 年和 2017 年的比赛中，AlphaGo 分别击败了世界围棋冠军李世石和中国围棋冠军柯洁，展示了 AI 在复杂游戏中的强大实力。然而，AlphaGo 的胜利也引发了关于失控性风险的讨论。一些人担心，如果将 AlphaGo 的技术应用于其他领域，如军事、经济等，可能会导致不可预测的后果。为了确保 AI 技术的可控性，DeepMind 公司在 AlphaGo 的开发过程中采取了一系列措施，如限制其计算资源、设置时间限制等。

2）自动驾驶汽车：自动驾驶汽车是 AI 技术在交通领域的应用之一。虽然自动驾驶汽车有望提高交通效率、减少交通事故等，但其失控性风险也引起了广泛关注。例如，2018 年 3 月，一辆 Uber 自动驾驶汽车在美国亚利桑那州发生了一起事故，导致一名行人死亡。事故调查发现，自动驾驶汽车的传感器未能正确识别行人和自行车骑手。为了防止类似事故的发生，各国政府和企业纷纷加强了对自动驾驶汽车的监管和管理，如制定严格的测试标准、要求驾驶员随时准备接管等。

总之，失控性风险是 AI 发展过程中面临的一个重要挑战。为了确保 AI 系统在人类的控制之下运行，我们需要通过技术创新、伦理规范、监管和管理等多种手段，降低失控性风险的可能性。同时，我们还需要

提高公众对 AI 风险的认识和理解，共同应对 AI 带来的挑战。

（三）完善法规，杜绝侵权性风险

侵权性风险是关于 AI 可能侵犯个人或群体权益的风险。随着人工智能技术的不断发展和应用，这种风险也日益凸显。为了解决这个问题，我们需要建立完善的法律体系，保护个人隐私和权益。

1、侵权性风险的表现

1）侵犯个人隐私：AI 技术在处理大量数据的过程中，可能会收集、存储和使用个人的敏感信息，如身份证号、电话号码、家庭住址等。这些信息一旦泄露，可能会导致个人隐私受到侵犯，甚至引发诈骗、勒索等犯罪行为。

2）侵犯知识产权：AI 技术在创作、翻译、音乐制作等领域的应用，可能会涉及到知识产权的问题。例如，AI 创作的音乐、文章等作品，其版权归属问题尚未明确；AI 翻译软件可能会侵犯原作者的翻译权等。这些问题可能导致原创者的权益受到损害。

3）侵犯肖像权：AI 技术在图像识别、视频监控等领域的应用，可能会涉及到肖像权的问题。例如，AI 监控系统可能会误识别他人为犯罪嫌疑人，导致无辜者名誉受损；AI 图像识别技术可能会将他人的肖像用于商业广告等。这些问题可能导致个人肖像权受到侵犯。

4）侵犯言论自由：AI 技术在内容审查、舆论引导等领域的应用，可能会涉及到言论自由的问题。例如，AI 内容审查系统可能会误删合法言论，导致言论自由受到限制；AI 舆论引导技术可能会操控公众舆论，

影响民主决策等。这些问题可能导致言论自由受到侵犯。

2、侵权性风险的原因

1）法律法规滞后：目前，针对 AI 技术的法律法规尚不完善，很多问题尚无明确规定。这导致了 AI 技术在实际应用中，容易产生侵权性风险。

2）技术不成熟：虽然 AI 技术取得了很大的进步，但在某些领域仍存在技术不成熟的问题。例如，AI 图像识别技术在识别人脸时，可能会出现误识别的情况；AI 翻译软件在翻译过程中，可能会出现语义理解错误等问题。这些问题容易导致侵权性风险的产生。

3）监管缺失：由于 AI 技术的快速发展，监管部门很难跟上技术发展的步伐，导致监管缺失。这使得一些企业和个人在开发和应用 AI 技术时，容易忽视侵权性风险。

3、解决侵权性风险的措施

1）完善法律法规：政府应加快制定和完善针对 AI 技术的法律法规，明确 AI 技术在应用过程中可能涉及的侵权性问题，为侵权行为设定明确的法律责任。

2）加强技术研发：企业和研究机构应加大研发投入，提高 AI 技术的准确性和可靠性，降低侵权性风险的发生概率。

3）建立监管机制：政府应建立健全 AI 技术的监管机制，加强对 AI 技术应用的监督和管理，确保其在合法合规的范围内使用。

4）提高公众意识：政府和企业应加强对公众的宣传教育，提高公

众对 AI 技术侵权性风险的认识和防范意识。

4、案例分析

1）2018 年，一款名为“DeepFakes”的 AI 换脸技术在网络上引起了广泛关注。这款技术可以将一个人的脸部特征替换到另一个人的脸上，生成逼真的虚假视频。这种技术被恶意利用后，可能会导致个人肖像权受到侵犯，甚至引发诽谤、敲诈等犯罪行为。为了应对这一问题，各国政府和企业纷纷加强了对 AI 换脸技术的监管和管理。

2）2019 年，一款名为“Replika”的 AI 聊天机器人在全球范围内受到了广泛关注。这款机器人可以与用户进行深度交流，了解用户的情感需求和生活经历。然而，这款机器人在与用户交流过程中，可能会收集到用户的敏感信息，如家庭住址、工作单位等。这些信息一旦泄露，可能会导致个人隐私受到侵犯。为了解决这一问题，Replika 的开发团队表示将加强数据安全保护措施，确保用户信息安全。

3）2020 年，一款名为“ZAO”的 AI 换脸应用在中国引发了广泛关注。这款应用可以将用户的脸部特征替换到电影、电视剧等影视作品中的角色上，生成逼真的换脸视频。然而，这款应用在收集用户个人信息时，并未明确告知用户其用途和范围，引发了用户隐私泄露的担忧。为了应对这一问题，中国政府要求相关部门加强对 AI 换脸应用的监管和管理，确保其合法合规运营。

侵权性风险是关于 AI 可能侵犯个人或群体权益的风险。为了解决这个问题，我们需要建立完善的法律体系，保护个人隐私和权益。同时，

政府、企业和社会各界也需要共同努力，加强技术研发、监管和宣传教育，降低侵权性风险的发生概率。

（四）确保透明公正，规避歧视性风险

歧视性风险是 AI 技术在应用过程中可能面临的一个重要问题。由于 AI 系统的决策过程往往是基于大量的数据和算法，因此，如果这些数据和算法存在偏见，那么 AI 系统就可能对某些特定群体产生歧视。这种歧视可能表现为就业、信贷、保险等方面的不公平待遇。为了应对这一风险，我们需要提升 AI 系统的透明度和可解释性，确保算法公正公平。

1、歧视性风险的表现

1）就业歧视：AI 技术在招聘领域的应用可能导致某些特定群体的就业机会受到限制。例如，一些企业在招聘过程中使用 AI 技术进行简历筛选，但由于算法可能存在性别、年龄等偏见，导致某些特定群体的求职者难以获得面试机会。

2）信贷歧视：AI 技术在信贷领域的应用可能导致某些特定群体的信贷需求受到限制。例如，一些银行和金融机构使用 AI 技术进行信贷审批，但由于算法可能存在种族、地域等偏见，导致某些特定群体的借款人难以获得贷款。

3）保险歧视：AI 技术在保险领域的应用可能导致某些特定群体的保险需求受到限制。例如，一些保险公司使用 AI 技术进行保险定价，但由于算法可能存在性别、年龄等偏见，导致某些特定群体的投保人难

以获得合适的保险产品。

2、歧视性风险的原因

1）数据偏见：AI 系统的决策过程是基于大量的数据，如果这些数据存在偏见，那么 AI 系统就可能对某些特定群体产生歧视。例如，在招聘领域，如果企业使用的数据集存在性别、年龄等偏见，那么 AI 系统在进行简历筛选时就可能对某些特定群体产生歧视。

2）算法偏见：AI 系统的决策过程是基于算法的，如果这些算法存在偏见，那么 AI 系统就可能对某些特定群体产生歧视。例如，在信贷领域，如果银行和金融机构使用的算法存在种族、地域等偏见，那么 AI 系统在进行信贷审批时就可能对某些特定群体产生歧视。

3）人为偏见：AI 系统的开发和应用过程中，开发者和维护者的人为偏见也可能导致歧视性风险的产生。例如，开发者在设计算法时可能有意无意地引入了性别、年龄等偏见，从而导致 AI 系统对某些特定群体产生歧视。

3、解决歧视性风险的措施

1）提升 AI 系统的透明度：为了确保 AI 系统的公正公平，我们需要提升其透明度，让使用者了解 AI 系统的决策过程和依据。这可以通过公开算法原理、提供决策依据等方式实现。

2）提高 AI 系统的可解释性：为了确保 AI 系统的公正公平，我们还需要提高其可解释性，让使用者能够理解 AI 系统的决策

逻辑。这可以通过提供详细的决策过程说明、使用易于理解的语言等方式实现。

3）确保算法公正公平：为了确保 AI 系统的公正公平，我们需要确保算法的公正公平。这可以通过使用无偏数据、消除人为偏见等方式实现。

4）加强监管和审查：为了确保 AI 系统的公正公平，我们还需要加强对 AI 系统的监管和审查。这可以通过建立专门的监管机构、制定相关的法律法规等方式实现。

4、案例分析

1）亚马逊公司的人工智能招聘系统“Recruit AI”曾因涉嫌性别歧视而受到关注。据报道，该系统在筛选简历时，可能会将女性候选人排除在外。亚马逊公司后来对此进行了调查，并表示将改进算法，确保招聘过程的公正公平。

2）美国一家名为“Truepic”的图片版权保护公司曾因涉嫌种族歧视而受到关注。该公司使用 AI 技术进行图片识别和版权保护，但在实际应用中，其算法被指控对非洲裔美国人的照片识别准确率较低。Truepic 公司后来对此进行了调查，并表示将改进算法，确保图片识别的公正公平。

3）中国一家名为“水滴筹”的互联网医疗平台曾因涉嫌性别歧视而受到关注。据报道，该平台在为患者筹款时，可能会对女性患者的筹款成功率产生影响。水滴筹公司后来对此进行了调

查，并表示将改进算法，确保筹款过程的公正公平。

歧视性风险是 AI 技术在应用过程中可能面临的一个重要问题。为了应对这一风险，我们需要提升 AI 系统的透明度和可解释性，确保算法公正公平。同时，政府、企业和社会各界也需要共同努力，加强监管和审查，确保 AI 技术的公正公平应用。

（五）加强培训教育，缓解人类失业风险

人类失业风险是 AI 技术发展过程中的一个重要问题。随着 AI 技术的不断进步，越来越多的工作岗位可能会被机器取代。这一现象在全球范围内都有所体现，尤其是在制造业、服务业和金融业等领域。为了应对这一问题，我们需要推动教育和培训，培养新的技能和能力，以适应新的工作环境。

1、人类失业风险的表现

1）制造业：在制造业领域，AI 技术的发展已经导致了许多工作岗位的减少。例如，自动化生产线可以替代大量的装配工人，而机器人则可以完成一些复杂的操作任务。根据国际劳工组织 (ILO) 的数据，到 2025 年，全球约有 8500 万工作岗位面临被自动化取代的风险。

2）服务业：在服务业领域，AI 技术的应用也可能导致部分岗位的消失。例如，智能客服系统可以替代传统的人工客服，而自动驾驶技术则可能影响到出租车司机、货车司机等职业。根据麦肯锡全球研究所的报告，到 2030 年，全球约有 14% 的工作岗

位可能被自动化取代。

3）金融业：在金融业领域，AI 技术的应用也可能导致部分岗位的减少。例如，智能投顾系统可以替代传统的投资顾问，而区块链技术则可能影响到银行柜员等职业。根据高盛的研究，到 2025 年，全球约有 1700 万金融从业人员可能面临被自动化取代的风险。

2、人类失业风险的原因

1）技术进步：AI 技术的发展使得机器可以完成越来越多的工作任务，从而取代了部分人类的工作。这种技术进步是不可避免的，因为人类社会的发展总是伴随着新技术的出现和应用。

2）成本考虑：对于企业来说，使用 AI 技术可以降低生产成本，提高生产效率。因此，许多企业选择采用 AI 技术来替代部分人力。

3）政策因素：政府在推动 AI 技术发展的过程中，可能会出台一些鼓励企业采用 AI 技术的政策，从而加速了 AI 技术在各个领域的应用。

3、解决人类失业风险的措施

1）教育和培训：为了应对 AI 技术带来的失业风险，我们需要加强教育和培训，培养新的技能和能力。这包括提高人们的计算机素养、编程能力、数据分析能力等，以适应新的工作环境。

2）政策引导：政府应该制定相应的政策，引导企业在使用

AI 技术的过程中充分考虑到人力资源的培养和利用。例如，政府可以提供税收优惠、补贴等政策支持，鼓励企业开展员工培训和再就业服务。

3）产业升级：通过推动产业升级，将 AI 技术与传统产业相结合，创造新的就业岗位。例如，可以将 AI 技术应用于农业、医疗、教育等领域，提高这些领域的生产效率和服务质量。

4）社会保障体系完善：政府需要完善社会保障体系，为失业人员提供一定的生活保障和再就业服务。这包括提高失业保险金的支付水平、扩大再就业培训的覆盖范围等。

4、案例分析

1）亚马逊：亚马逊是一家典型的应用 AI 技术的公司。在亚马逊的仓库中，大量的机器人负责货物的搬运和分拣工作。这使得亚马逊的仓库工作效率大大提高，但同时也导致了部分仓库员工的失业。为了应对这一问题，亚马逊推出了“机器人 + 人”的工作模式，即让机器人和人类员工共同完成工作任务。这样既提高了工作效率，又保证了员工的就业机会。

2）优步：优步是一家提供网约车服务的公司。随着自动驾驶技术的发展，优步计划在未来几年内推出自动驾驶汽车。这意味着优步的部分驾驶员可能会面临失业的风险。为了应对这一问题，优步已经开始与汽车制造商合作，研发自动驾驶汽车。同时，优步也在积极寻求与政府合作，制定自动驾驶汽车的相关

政策和法规。

3）IBM：IBM 是一家致力于人工智能技术研发的公司。在 IBM 的实验室中，研究人员正在开发一种名为“沃森”的人工智能系统。该系统可以模拟人类的思考和决策过程，广泛应用于金融、医疗、教育等领域。然而，随着沃森系统的不断完善，部分从事相关工作的人员可能会面临失业的风险。为了应对这一问题，IBM 已经开始与高校合作，培养具有人工智能专业知识的人才。同时，IBM 也在积极推动人工智能技术与传统产业的融合，创造新的就业岗位。

人类失业风险是 AI 技术发展过程中的一个重要问题。为了应对这一问题，我们需要推动教育和培训，培养新的技能和能力，以适应新的工作环境。同时，政府和企业也需要采取相应的措施，如政策引导、产业升级、完善社会保障体系等，以确保 AI 技术的发展能够惠及整个社会。只有这样，我们才能确保 AI 技术的健康发展，为人类社会带来更多的福祉。

(六) 促进政策公平和广泛合作，抹平数字鸿沟风险

数字鸿沟风险：AI 技术的普及可能加剧数字鸿沟，使得资源分配更加不均。解决方法是推动公平的数字政策，缩小数字鸿沟。随着人工智能 (AI) 技术的快速发展，其在各个领域的应用越来越广泛，为人们的生活带来了诸多便利。然而，AI 技术的普及也可能带来一些负面影响，其中之一就是加剧数字鸿沟，使得资源分

配更加不均。本文将通过数据和案例详细解释这一风险点，并提出相应的解决方法。

1、数字鸿沟风险的表现

1）地域差异：AI 技术的普及程度在不同地区存在明显差异。发达地区由于经济条件较好，基础设施建设完善，人们对新技术的接受度较高，因此 AI 技术在这些地区的普及程度较高。而在欠发达地区，由于经济条件较差，基础设施建设滞后，人们对新技术的接受度较低，因此 AI 技术在这些地区的普及程度较低。这种地域差异导致了数字鸿沟的产生。

2）教育水平差异：AI 技术的普及程度与人们的教育水平密切相关。受过高等教育的人群对新技术的接受度较高，更容易掌握和应用 AI 技术。而教育水平较低的人群对新技术的接受度较低，难以掌握和应用 AI 技术。这种教育水平差异也加剧了数字鸿沟。

3）收入差距：AI 技术的普及程度与人们的收入水平密切相关。高收入人群有更多的资金投入到新技术的学习和培训中，从而更容易掌握和应用 AI 技术。而低收入人群由于资金有限，难以承担新技术的学习和培训费用，因此难以掌握和应用 AI 技术。这种收入差距也加剧了数字鸿沟。

2、数字鸿沟风险的影响

1）社会不公平：数字鸿沟的存在使得资源分配更加不均，导致社会不公平现象加剧。在教育领域，由于教育资源的不均衡

分配，贫困地区的学生难以获得高质量的教育资源，从而影响了他们的学习和发展。在就业领域，由于 AI 技术的普及程度不同，不同地区、不同教育背景和收入水平的人群在就业市场上的竞争地位也不同，从而导致就业机会的不公平分配。

2）经济发展不平衡：数字鸿沟的存在影响了欠发达地区的经济发展。由于 AI 技术的普及程度较低，这些地区的企业和产业难以充分利用 AI 技术提高生产效率和降低成本，从而影响了其经济发展速度。同时，由于人才流失，这些地区的创新能力和竞争力也受到影响。

3）社会不稳定：数字鸿沟的存在可能导致社会不稳定现象加剧。一方面，由于资源分配不均，低收入人群的生活水平难以提高，可能导致社会不满情绪的积累；另一方面，由于教育水平和就业机会的差距，年轻人对未来的信心可能受到影响，从而导致社会动荡。

3、解决数字鸿沟风险的方法

1）推动公平的数字政策：政府应制定公平的数字政策，确保 AI 技术的普及不受地域、教育水平和收入差距的影响。这包括加大对欠发达地区基础设施建设的投入，提高这些地区的网络覆盖率和通信质量；加强教育投入，提高欠发达地区的教育水平；实施扶贫政策，提高低收入人群的收入水平等。

2）缩小数字鸿沟：政府和企业应共同努力，通过技术创新

和合作，降低 AI 技术的成本和门槛，使其更容易被欠发达地区和低收入人群所接受和使用。此外，还可以通过培训和教育项目，提高欠发达地区和低收入人群的 AI 技术应用能力。

3）促进公平竞争：政府应加强对 AI 技术领域的监管，防止市场垄断和不公平竞争现象的出现。同时，鼓励企业开展技术研发和创新，提高我国在 AI 技术领域的国际竞争力。

4）强化国际合作：政府应积极参与国际交流和合作，引进先进的 AI 技术和管理经验，推动我国 AI 技术的发展和应用。同时，通过国际合作，帮助欠发达地区和低收入人群提高 AI 技术应用能力，缩小数字鸿沟。

4、案例分析

1）中国农村地区的数字鸿沟：在中国农村地区，由于基础设施落后、教育水平较低和收入水平不高等原因，AI 技术的普及程度相对较低。为了缩小农村地区的数字鸿沟，中国政府采取了一系列措施。例如，加大农村地区基础设施建设投入，提高农村地区的网络覆盖率和通信质量；实施农村教育改革，提高农村地区的教育水平；实施扶贫政策，提高农村地区低收入人群的收入水平等。这些措施在一定程度上缓解了农村地区的数字鸿沟问题。

2）非洲国家的数字鸿沟：非洲国家由于经济发展水平较低、基础设施落后和教育资源不足等原因，数字鸿沟问题较为严重。为了缩小非洲国家的数字鸿沟，国际社会采取了一系列措施。例

如，联合国教科文组织 (UNESCO) 发起了“全球教育联盟”，旨在通过国际合作提高非洲国家的教育水平；世界银行等国际金融机构提供了大量资金支持非洲国家的基础设施建设；非政府组织 (NGO) 和企业也积极参与非洲国家的教育和科技项目，帮助当地人民提高 AI 技术应用能力等。这些措施在一定程度上缓解了非洲国家的数字鸿沟问题。

数字鸿沟风险是 AI 技术普及过程中需要关注的一个重要问题。通过推动公平的数字政策、缩小数字鸿沟、促进公平竞争和强化国际合作等措施，我们可以有效应对这一风险点，确保 AI 技术在我国的健康发展和应用。同时，通过借鉴国内外的成功经验和教训，我们还可以为其他国家和地区提供有益的参考和借鉴。

总之，AI 技术的普及为人类社会带来了巨大的发展机遇，但同时也带来了数字鸿沟风险。数字鸿沟的存在使得资源分配更加不均，导致社会不公平现象加剧、经济发展不平衡和社会不稳定等问题。为了应对这一风险点，我们需要推动公平的数字政策、缩小数字鸿沟、促进公平竞争和强化国际合作等措施。通过这些努力，我们可以确保 AI 技术在我国的健康发展和应用，为人类社会的发展做出更大的贡献。

(七) 构建指导原则，减少人类道德价值观受损风险

人类道德价值观受损风险：AI 的决策过程可能与人类的道德价值观不符，影响社会的道德风气。这需要我们在设计和应用 AI

时，充分考虑其对人类道德价值观的影响。

随着人工智能 (AI) 技术的快速发展，其在各个领域的应用越来越广泛，为人们的生活带来了诸多便利。然而，AI 技术的普及也可能带来一些负面影响，其中之一就是人类道德价值观受损风险。本文将通过数据和案例详细解释这一风险点，并提出相应的解决方法。

1、人类道德价值观受损风险的表现

1）伦理冲突：AI 技术的决策过程可能与人类的道德价值观产生冲突。例如，在自动驾驶汽车中，当面临一个无法避免的事故时，AI 系统需要在保护乘客和行人之间做出选择。在这种情况下，AI 系统的决策可能与人类的道德价值观不符，导致伦理冲突。

2）歧视问题：AI 技术在数据处理和分析过程中，可能会因为训练数据的偏见而产生歧视性结果。例如，在招聘过程中，如果使用 AI 技术进行简历筛选，由于训练数据中存在性别、年龄等歧视性信息，可能导致 AI 系统对某些人群产生不公平的待遇。

3）隐私侵犯：AI 技术在数据收集和处理过程中，可能会侵犯个人隐私。例如，在智能家居设备中，AI 系统需要收集用户的生活习惯和偏好信息，以便为用户提供更加个性化的服务。然而，在这个过程中，用户的隐私信息可能会被泄露或滥用。

2、人类道德价值观受损风险的影响

1）社会信任危机：由于 AI 技术的决策过程可能与人类的道

德价值观不符，导致公众对 AI 技术的信任度下降。这可能会影响 AI 技术在各个领域的应用和发展，甚至引发社会恐慌和抵制。

2）法律和道德困境：AI 技术的伦理冲突和歧视问题可能导致法律和道德困境。例如，在自动驾驶汽车事故中，如何界定责任和赔偿问题；在招聘过程中，如何确保 AI 系统的公平性和透明度等。这些问题需要我们重新审视现有的法律体系和道德规范，以适应 AI 技术的发展。

3. 道德风气恶化：由于 AI 技术的决策过程可能与人类的道德价值观不符，导致社会的道德风气恶化。例如，在社交媒体上，由于 AI 算法的推荐机制，用户可能更容易接触到低俗、恶俗的内容，从而影响社会的道德风气。

3、解决人类道德价值观受损风险的方法

1）设立伦理指导原则：政府、企业和研究机构应共同制定 AI 技术的伦理指导原则，明确 AI 技术在设计和应用过程中应遵循的道德规范。这些原则应涵盖公平性、透明性、隐私保护等方面，以确保 AI 技术与人类的道德价值观相符。

2）加强监管和审查：政府应加强对 AI 技术的监管和审查，确保其在各个领域的应用符合伦理要求。这包括对 AI 技术的研发、应用和推广过程进行监督，以及对涉及伦理冲突和歧视问题的 AI 产品和技术进行审查和处罚。

3）提高公众道德意识：政府、企业和教育机构应加强对公

众的道德教育，提高公众的道德意识和道德判断能力。这有助于公众更好地理解和接受 AI 技术，减少因伦理冲突和歧视问题引发的社会矛盾。

4）促进跨学科合作：政府、企业和研究机构应鼓励跨学科合作，加强伦理学、社会学、心理学等领域的研究力量。这有助于我们从多角度审视 AI 技术对人类道德价值观的影响，为制定合理的伦理指导原则提供理论支持。

5）建立多方参与的伦理审查机制：政府、企业、研究机构和公众应共同参与 AI 技术的伦理审查过程，确保各方的利益得到充分保障。这可以通过建立多方参与的伦理审查委员会、开展公众征求意见等方式实现。

4、案例分析

1）谷歌 AlphaGo 事件：2016 年，谷歌的人工智能程序 AlphaGo 在围棋比赛中战胜了世界冠军李世石。这一事件引发了关于 AI 技术与人类道德价值观的讨论。一方面，人们认为 AlphaGo 的胜利证明了 AI 技术的强大；另一方面，也有人担忧 AI 技术在未来可能取代人类的智慧和创造力，导致人类道德价值观的衰退。

2）亚马逊 Alexa 事件：2018 年，亚马逊的智能语音助手 Alexa 误将一段恶意录音识别为笑话并播放给用户。这一事件引发了关于 AI 技术隐私保护的讨论。人们担忧 AI 技术在收集和处

理用户数据过程中可能存在隐私泄露的风险，从而影响用户的道德权益。

3）美国自动驾驶汽车事故：近年来，美国发生了多起自动驾驶汽车事故。在这些事故中，自动驾驶汽车在面临无法避免的事故时，需要做出选择保护乘客还是行人的决策。这些事故引发了关于 AI 技术伦理冲突的讨论。人们担忧 AI 技术在决策过程中可能与人类的道德价值观不符，导致伦理冲突和社会矛盾。

5、结论

AI 技术的普及为人类社会带来了巨大的发展机遇，但同时也带来了人类道德价值观受损风险。人类道德价值观受损风险表现为伦理冲突、歧视问题和隐私侵犯等现象。这些现象可能导致社会信任危机、法律和道德困境以及道德风气恶化等问题。为了应对这一风险点，我们需要设立伦理指导原则、加强监管和审查、提高公众道德意识、促进跨学科合作和建立多方参与的伦理审查机制等措施。通过这些努力，我们可以确保 AI 技术在我国的健康发展和应用，为人类社会的发展做出更大的贡献。

(八) 展望未来伦理演进，探索 45°平衡律

随着 AI 技术的不断进步，未来的伦理讨论可能会集中在更加复杂的问题上，如 AI 的道德决策能力、跨文化伦理问题、以及 AI 在全球治理中的角色等。同时，随着技术的成熟和社会对 AI 伦理的认识提高，可以预见将会有更多的国际标准和法规出台，

以指导 AI 的负责任发展。

AI 伦理已经成为了 AI 领域不可或缺的组成部分，对于大模型而言，其伦理、安全、人机对齐等问题之应对和解决，将需要政府、业界、学界等利益相关方进行持续的探索。

其一，数据安全、隐私泄露、抗攻击能力提升等问题是现有大模型应用面临的真实挑战，解决这些问题的技术手段还存在一定的局限性。对抗性人工智能技术与防御策略之间的竞赛将加剧，为了应对对抗性攻击和操纵等恶意行为，模型需要被设计为更加具有鲁棒性。

其二，从整体上对 AI 大模型的安全风险进行建模，系统化地构建安全评估系统是大模型安全领域的未来发展方向。这将最大程度地确保大模型应用是在符合社会价值与应用价值方面同步进行。

其三，增强模型透明度和可解释性。研究模型的可解释性，提高模型的透明度既是未来 AI 的发展方向，也能帮助提升 AI 模型的安全性。未来的人工智能模型可能会融入更先进的 XAI 技术。

其四，人机价值对齐和伦理嵌入设计 (ethics by design) 的理念将变得越来越重要。无论是 AI 价值对齐还是伦理嵌入设计，都需要人们发展新的更加务实的 AI 伦理框架及其实践指南。

其五，人工智能监管立法和国际治理合作将得到进一步推进。未来立法和监管措施将给大模型安全和伦理的研究和实践提供更

进一步的指导。

其六，在伦理研究中，跨学科合作是一个重要趋势。未来伦理研究需要吸收多领域的知识和技术，形成跨学科的研究团队，共同解决不断升级的复杂伦理问题，确保负责任的、安全可控的 AI 发展应用。

2024 年 7 月 4 日，2024 世界人工智能大会暨人工智能全球治理高级别会议 (WAIC 2024) 在上海开幕。上海人工智能实验室主任、首席科学家，清华大学惠妍讲席教授，衔远科技创始人周伯文提出一个技术主张：探索人工智能 45°平衡律——Towards AI-45° Law， AI-45°平衡律是指从长期的角度来看，我们要大体上沿着 45 度安全与性能平衡发展，平衡是指短期可以有波动，但不能长期低于 45° (如同现在)，也不能长期高于 45 度 (这将阻碍发展与产业应用)。这个技术思想体系要求强技术驱动、全流程优化、多主体参与以及敏捷治理。像安全可控的核聚变技术为全人类带来清洁、丰富的能源一样，我们希望通过深入理解 AI 的内在机理和因果过程，从而安全且有效地开发和使用这项革命性技术，我们坚信未来 AI 的风险控制将带给全球性的公共福祉。

四 、金砖国家扩展加盟，伦理安全展望新篇

在 2023 年金砖国家峰会之后，该联盟宣布从 2024 年 1 月起增加六个新成员。它们依次是阿根廷、埃塞俄比亚、伊朗、埃及、阿拉伯联合酋长国和沙特阿拉伯。

2023 年金砖国家峰会在南非举行，主题为“金砖与非洲 : 深化伙伴关系 , 促进彼此增长 , 实现可持续发展 , 加强包容性多边主义”。此次峰会取得了多项进展和成就，包括加强包容性多边主义、推动全球治理变革、深化经贸、财金合作、拓展政治安全合作、加强人文交流、坚持公平正义、完善全球治理等。

金砖国家需要做什么，取决于金砖国家战略选择，在战略选择的目标 AI 应用中，透析出潜在和现存的伦理风险，金砖国家有可能后发先至，决胜未来。

(一) 金砖国家应正视 AI 伦理风险

随着人工智能 (AI) 的飞速发展，金砖国家面临着许多挑战和伦理问题。这些问题不仅涉及到技术层面，还涉及到社会、经济、法律等多个方面。因此，我们需要积极寻找解决方案，以确保 AI 的发展能够造福人类社会，而不是带来负面影响。

首先，AI 发展所带来的挑战之一是就业问题。随着 AI 技术的不断进步，许多传统的工作岗位可能会被自动化取代。这可能导致大量的失业和社会不稳定。为了应对这一挑战，我们需要加强对失业人员的培训和教育，帮助他们适应新的工作环境。此外，

政府和企业也应该加大对创新产业的支持，以创造更多的就业机会。

其次，AI 发展可能加剧数字鸿沟。在金砖国家，许多人仍然无法享受到先进的互联网服务和智能设备。这使得他们在获取信息、接受教育和参与经济活动方面处于劣势地位。为了解决这一问题，我们需要加大对基础设施建设的投入，提高网络覆盖率和速度。同时，我们还需要推广低成本的智能设备，让更多的人能够享受到 AI 带来的便利。

再者，AI 发展可能引发隐私和安全问题。随着大数据和云计算技术的发展，越来越多的个人信息被收集和存储。这使得个人隐私面临前所未有的威胁。为了保护公民的隐私权，我们需要建立完善的法律法规，规范企业和政府部门对个人信息的收集、使用和传播。此外，我们还需要加强网络安全技术的研究和应用，防范黑客攻击和数据泄露等风险。

最后，AI 发展可能引发伦理道德问题。例如，自动驾驶汽车在紧急情况下应该如何选择行动？人工智能是否应该拥有与人类相同的权利和义务？这些问题需要金砖国家在技术发展的同时，进行深入的伦理道德探讨。金砖国家可以借鉴国际上的经验和做法，制定相应的伦理准则和法律法规，引导 AI 技术的健康发展。

(二) 金砖国家需合作应对共同挑战

金砖国家为解决上述问题，可以加强合作，共同应对 AI 带来的全球性挑战。

1、建立合作机制：

成立金砖国家 AI 合作小组，定期召开工作会议，共商合作事宜。

建立信息共享平台，促进金砖国家在 AI 领域的经验交流和技术合作。

2、共同制定标准：

在人工智能伦理治理方面缺乏国际合作会创造一个支离破碎的法律环境，一定程度上会阻碍创新，同时会引发对不同司法管辖区的潜在差异和不一致的担忧。这种缺乏协调可能会造成监管漏洞，并在追究参与者对人工智能相关危害的责任方面构成挑战。金砖国家之间携手制定 AI 伦理准则和技术标准，有利于推动金砖国家间的互操作性和互信。

同时，金砖国家应积极参与全球 AI 治理体系的构建，争取更多话语权和影响力。

3、加强联合研发：

鼓励金砖国家在 AI 领域开展联合研发项目，共同攻克关键技术难题。

促进产学研用深度融合，加速 AI 技术的产业化进程。

4、扩大人才交流：

建立金砖国家 AI 人才交流机制，促进人才互访和合作研究。举办 AI 领域的研讨会、工作坊等活动，增进金砖国家间的相互

了解和合作意愿。

5、应对共同挑战：

携手应对 AI 发展带来的全球性挑战，如数据安全、隐私保护、就业市场变动等。

在国际事务中发挥积极作用，推动构建更加公正合理的全球 AI 治理体系。

通过这些措施，金砖国家可以在 AI 时代加强合作，共同应对伦理挑战，推动 AI 技术的健康、可持续发展，并为全球科技进步做出金砖国家应有的贡献。

智能如箭逐风飞， 伦理风险伴影随。

人机和谐需共筑， 金砖之路更辉煌。

参考文献

[1]《数字中国发展报告 (2023 年)》[R]. 国家数据局，2024 年 6 月 .

[2]《2023 年通信业统计公报》[R]. 工业和信息化部，2024 年 1 月 .

[3]《中国数字经济发展研究报告 (2023 年)》[R]. 中国信息通信研究院，2023 年 4 月 .

[4] 黄长武 .《人工智能伦理概述》[R].IEEE，2022-07-28.

[5] 席丹 .《寻求价值对齐之路：人工智能面临的课题与挑战》[J]. 传媒，2024(11).

[6]《2024 年人工智能政策预测》[R].CSIS，2024-01-30.

[7] 韩冰意、苏中 .“信息茧房”、隐私外泄，如何应对人工智能带来的伦理风险？ [N]. 新华网，2023-01-19.

[8] 李洋，梁正 . 道德行动体视角下生成式人工智能伦理问题解析 [J]. 中国人工智能学会通讯，2024,14(1).

[9] 普华永道中国 .《解码生成式 AI 引发的安全风险与挑战》，2023 年 10 月 .

[10] 冯姣 .《人工智能面临的风险及伦理治理、法律规制》，民主与法制时报 .

[11]《人工智能伦理问题：AI 技术应用的道德与法律挑战》，

搜狐网，2023 年 12 月 13 日 .

[12]《2023 全球人工智能伦理治理报告》(案例篇)：人类的“造物主”地位正在受到动摇 , 发布于北京钛媒体 APP, 2023 年 12 月 7 日 .

[13]《AI 伦理》，IBM，2024 年 7 月

[14]《人工智能伦理 (AI Ethics)》，Tech Ethics China, 2023 年 6 月 4 日 .

[15] 俞鼎 .《“有意义的人类控制”：一项新的人工智能伦理治理原则》，中国社会科学网 ,2022 年 12 月 27 日 .

[16]《任重道远：一文全览 AI 风险审计的定义、方法、实践和展望》，数据化审计，2024 年 07 月 08 日 .

[17] 梁刚健 .《<AI 风险声明 > 的背后之问：灾难性 AI 风险可能如何出现？》 2023 年 11 月 6 日 .

[18] Kate Yuan.《人工智能的三大风险以及如何避免》，2022 年 11 月 22 日 .

[19] 李水青、香草 .《AI 让全球 GDP 翻 10 倍！但风险比核问题还严峻，三大图灵奖大佬对话，WAIC 干货看尽》，凤凰网，2024 年 07 月 04 日 .

[20]《2024 年的 AI 风险管理：您应该知道的几件事》，Splunk 大数据公众号， 2024 年 03 月 18 日

[21] 德勤中国 .《创新与信心偕行 – 德勤人工智能与风险管理报告》，2024 年 06 月 .

[22] Ravikumar Ram.《人工智能治理：从理论到实践——NIST AI 风险管理框架解析》，ISACA Blog，2024 年 05 月 06 日 .

[23]《2023 年中国人工智能产业研究报告》，艾瑞咨询，2023 年。

附录

2023 年 AI 大事记：

2023 年 9 月 13 日，阿里云宣布其最新的人工智能大模型“通义千问”已经通过备案，并表示将开源一个更大参数规模的大模型版本，供全社会免费商用。

2023 年 8 月，广汽集团宣布正式推出 AI 大模型技术的最新研发成果——广汽 AI 大模型平台。

2023 年 7 月，Meta 公司宣布，将开始向部分用户推送其 Meta Ray-Ban 智能眼镜的多模态 AI 功能。

2023 年 6 月 8 日，微软宣布推出集成 AI 技术的搜索引擎“Bing”。

2023 年 4 月 28 日，特斯拉宣布推出全自动驾驶技术。

2023 年 4 月 15 日，苹果宣布推出 AI 芯片。

2023 年 4 月，特斯拉在美国德克萨斯州首次向公众展示了其全自动驾驶技术，并开始进行大规模的测试。

2023 年 3 月 21 日春季新品发布会上，苹果推出了第二代 AI 芯片——M2，并将其应用于新款 MacBook Air 和 13 英寸 MacBook Pro 等产品中。苹果的 AI 芯片此前已经应用于 iPhone、iPad 等设备。

2023 年 3 月 15 日，亚马逊宣布推出 AI 语音助手“Alexa”。

2023 年 3 月 15 日，Meta 宣布推出 AI 绘画工具“DALL-E”，用户只需输入文字描述，AI 便能自动生成相应的图像。

2023 年 2 月 8 日，微软推出了集成 AI 技术的搜索引擎“Bing”，以与谷歌搜索展开更激烈的竞争。

2023 年 2 月 7 日，华为宣布推出 AI 芯片。

2023 年 2 月 6 日，谷歌推出了 AI 聊天机器人产品“Bard”，旨在与 OpenAI 的 ChatGPT 展开竞争。

2023 年 1 月 18 日，百度宣布推出 AI 语音助手“小度助手”。