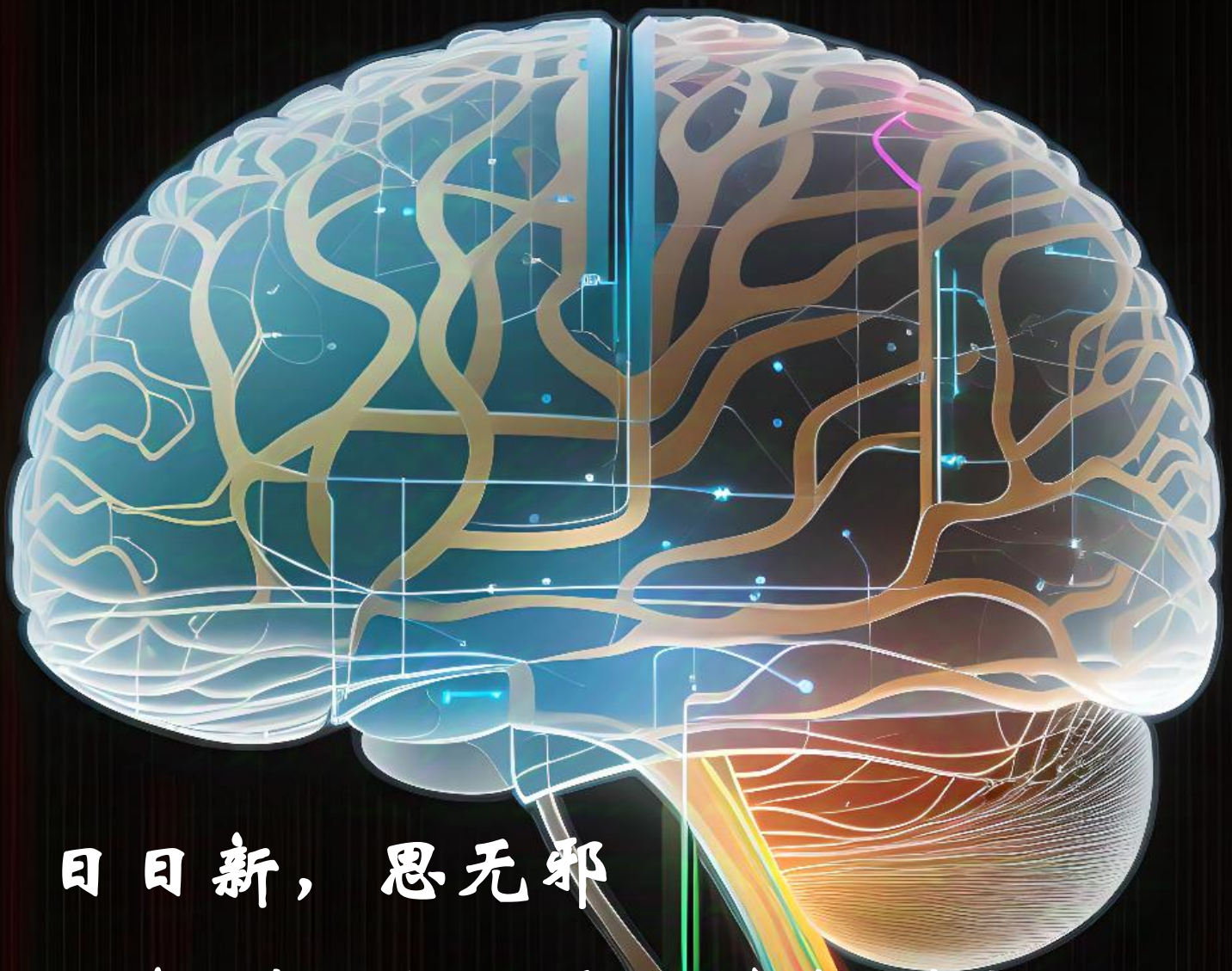


日日新，思无邪

——商汤大模型伦理原则与实践白皮书

商汤人工智能伦理与治理委员会

人工智能伦理治理年度报告（2023）

商汤智能产业研究院、上海交通大学计算法学与 AI 伦理研究中心联合出品

目录

致大模型从业者的一封信：

人类普适价值观，驱动 AI 伦理“三维对齐”

图 1： “三维对齐”科技伦理模式

人类经济文明之树上“低垂的果实”已被采摘一空，新一代创新科技犹如“新的进化梯子”，帮助人类采摘到“更高的果实”。联合国常务副秘书长阿明娜·穆罕默德女士在 2022 年联合国可持续发展高级别政治论坛上说：“新冠疫情、冲突以及环境危机造成的挑战已经影响到教育、医疗保健、性别平等以及经济发展。她指出，2030 年议程的时间表已经过半，但“我们还没有完成一半的任务”。为此，全球需要在可再生能源、粮食体系和数字连接领域进行转型，加快“人力资本投资，为机遇融资”的步伐，进而化危为机。”而以人工智能为代表的新一代科技，即能为人类持续提升环境治理能力，带来新能源、新农业、新制造与新商业，又能通过科技伦理在一定程度上弥补“数字鸿沟”，创造新兴就业市场，实现各国经济结构的转型升级，

正如丘吉尔的名言:“决不要浪费一场好的危机(Never waste a good crisis)”。

人机对齐，以保障AI任务目标与人类用户意图对齐、AI伦理与人类价值观、社会风序良俗、法律政策对齐。麦肯锡全球研究院（McKinsey Global Institute）说，人工智能正在促进社会发生转变，这种转变比工业革命“发生的速度快 10 倍，规模大 300 倍，影响几乎大3000倍。”为了保证AI大模型产品全生命周期的人机对齐，应持续识别发现训练数据集、应用场景的偏差风险，并设计风险内控检查表与管理流程，并在RLHF（基于人类反馈的增强学习）的测试、使用过程中，形成反馈闭环，修正问题、增补风险项、完善高阶伦理原则。正如人类的价值观是经过历史、文化、社会活动、产业革命逐步变化形成，从AIGC到AGI的伦理秩序同样需要持续改进与发展。

人人对齐，是充分考虑世界不同种族文化差异、区域经济差异、国家社会发展阶段的不同，跨越文明形态、地缘分歧，形成普惠全人类的互相尊重、包容、理解的统一价值观体系。哈佛大学教授塞缪尔•亨廷顿在《文明的冲突》一书中，根据历史发展将世界分为八大文明板块，分别拥有不同的文化价值观。农业时代、工业时代、信息时代，人类价值观的分歧长期存在、并变化演进，伴随人工智能技术进入千行百业、拥有了数亿用户群体，跨洲际AI2.0服务、跨国科研合作形成了很好的生态产业链、开放开源社群，急需一种普适全人类的价值观指引。联合国秘书长古特雷斯于2021年9月发布了《我们的

共同议程》报告，面向政府、联合国系统、私营部门（含科技公司）、民间社会、基层组织、学术界和个人，提出《全球数字契约》，该契约将成为“所有人共享开放、自由和安全的数字未来的共同原则”，涵盖的领域包括：数字连接、避免互联网碎片化、为人们提供将如何使用其数据的选择、网络人权，以及通过引入对歧视和误导信息问责标准促进可靠的互联网内容。该契约有望在 2024 年形成全球初步共识，并为人人对齐形成统一框架。

人类、科技与环境对齐，为避免环境恶化、灾难性气候为人类带来社会崩溃等恶劣影响，科技企业应注重并遵循环境伦理。1972 年，罗马俱乐部在《增长的极限》报告中提出：“一旦人口与经济超越了地球的物理极限，那么只有两条路可以返回正常：通过日益升级的短缺与危机而导致的非自愿崩溃；或者通过精心的社会选择而带来的生态足迹有控制的缩减。”并预测人类将在 21 世纪因资源瓶颈、环境恶化等客观因素带来经济衰退、社会崩溃等严重影响。基于全球 ESG 理念、碳达峰碳中和目标，AI 产业链、科研机构、私营机构应肩负起环境生态保护、能源可持续发展的社会责任，重新思考在满足了人类生存需求之后，人类该如何满足地球家园环境的保护要求与自然资源良性开发，为子孙后代留住流水青山。

新兴科技来自全人类，更应该造福全人类，所以对人类命运共同体负责、对全球环境负责的 AI 伦理风控、AI 治理机制，将成为大模型技术、生成式人工智能技术、通用人工智能技术的核心指引。

——商汤科技智能产业研究院院长 田丰

【关于商汤】

作为行业领先的人工智能软件公司，商汤集团以“坚持原创，让 AI 引领人类进步”为使命，“以人工智能实现物理世界和数字世界的连接，促进社会生产力可持续发展，并为人们带来更好的虚实结合生活体验”为愿景，旨在持续引领人工智能前沿研究，持续打造更具拓展性更普惠的人工智能软件平台，推动经济、社会和人类的发展，并持续吸引及培养顶尖人才，共同塑造未来。

商汤拥有深厚的学术积累，并长期投入于原创技术研究，不断增强行业领先的全栈式人工智能能力，涵盖感知智能、决策智能、智能内容生成和智能内容增强等关键技术领域，同时包含 AI 芯片、AI 传感器及 AI 算力基础设施在内的关键能力。此外，商汤前瞻性打造新型人工智能基础设施——SenseCore 商汤 AI 大装置，打通算力、算法和平台，大幅降低人工智能生产要素价格，实现高效率、低成本、

规模化的 AI 创新和落地，进而打通商业价值闭环，解决长尾应用问题，推动人工智能进入工业化发展阶段。

商汤业务涵盖智慧商业、智慧城市、智慧生活、智能汽车四大板块，相关产品与解决方案深受客户与合作伙伴好评。

商汤坚持“平衡发展”的伦理观，倡导“可持续发展、以人为本、技术可控”的伦理原则，实行严格的产品伦理风险审查机制，建设全面的 AI 伦理治理体系，并积极探索数据治理、算法治理相关的检测工具和技术手段，致力于将伦理原则嵌入到产品设计、开发、部署的全生命周期，发展负责任且可评估的人工智能。

目前，商汤集团（股票代码：0020.HK）已于香港交易所主板挂牌上市。商汤现已在香港、上海、北京、深圳、成都、杭州、南平、青岛、三亚、西安、台北、澳门、京都、东京、新加坡、利雅得、阿布扎比、迪拜、吉隆坡、首尔等地设立办公室。另外，商汤在泰国、印度尼西亚、菲律宾等国家均有业务。

【关于本报告】

商汤集团（以下简称“商汤”、“公司”或“我们”）主动向社会公众报告公司的人工智能伦理与治理情况，让全社会了解、监督商汤的人工智能伦理与治理工作。

商汤面向社会各界发布人工智能伦理与治理报告，旨在通过及时披露商汤的人工智能伦理治理理念和实践，促进商汤与利益相关方以及社会公众之间的了解、沟通与互动，推动发展负责任且可评估的人工智能。

作为商汤人工智能伦理与治理的年度报告，本报告于 2023 年 7 月以中文版本率先发布，英文版本将另行择期发布，如对本报告有任何建议和意见，请通过以下方式与商汤联系：

商汤 AI 伦理与治理委员会：AIethics.committee@sensetime.com

一、生成式人工智能浪潮下的范式变革

2022 年，是人工智能发展历程中极具里程碑意义的一年。以 ChatGPT 为代表的生成式人工智能工具迅速火爆全球，成为人类迈向通用人工智能（Artificial General Intelligence）时代的历史性节点之一。ChatGPT 基于 NLP 基础模型（NLP 即自然语言处理），体现出跨知识领域、跨语种、多模态为特征的海量知识挖掘、人机自然交流，能实现撰写代码、回答问题、书写论文、诗歌、剧本等指令，可以让 AI 生产力从重复性体力生产环节向认知和创造性生产环节延伸。上线仅两个月，ChatGPT 活跃用户便突破 1 亿大关，一举成为人类科技史上消费者增长速度最快的应用程序。

ChatGPT 这一现象级应用的成功，标志着人工智能正式进入以“基础模型+微调”为主要特征的生产范式，推动人工智能进入 2.0 阶段。2012 年，后来被誉为“人工智能教父”的 Geoffrey Hinton 带领团队凭借卷积神经网络（Convolutional Neural Networks, CNN）在 ImageNet 的比赛中获得冠军，标志着机器的视觉识别能力能够超越人眼识别准确率，开启了人工智能工业化的进程。由此，人工智能开始走进一个个应用场景。这一阶段，人工智能的生产范式属于典型的“手工作坊”模式，即人工智能厂商需要针对每个细分场景开发专属的模型，进而导致人工智能开发周期长，落地成本高，成为人工智能规模化应用亟待突破的制约。ChatGPT 等预训练大模型应用的成功打破了“手工定制”的生产范式，通过“基础模型+微调”的方式，使得一个基础模型能够快速适配海量的下游应用，为人工智能的规模化落

地提供了一条可行的路径。

如果以生产模式的差异作为分界线，我们大致可以将 2022 年之前的人工智能发展阶段定义为人工智能 1.0 阶段（AI1.0 阶段），将 2022 年之后的人工智能发展阶段定义为人工智能 2.0 阶段（AI2.0 阶段）。AI2.0 阶段相比 AI1.0 阶段有以下几点显著变化：一是任务类型由封闭场景转向开放任务；二是数据处理模态由单一模态转向多模态；三是模型类型由判别式模型转向生成式模型；四是生产模式由“手工作坊”转向“基础模型+微调”。

图 2： 人工智能领域的范式变革

进入 AI2.0 阶段，人工智能生产范式的变革同样引起了人工智能风险范式的转变。具体来说，一方面任务场景的开放性导致风险的潜在边界理论上被无限放大，风险来源防不胜防，而且风险评估标准更加难以界定；同时，跨模态数据交互能力的实现，在大幅降低 AI 工具应用门槛的同时，也使得 AI 滥用的风险呈指数上升。另一方面，

由于生产范式的转变，基础模型内含的风险也会随着下游应用的规模推广而被规模化扩散，风险的外溢性显著提升。此外，诸多安全机制的嵌入也会影响模型自身的表现，如何实现安全能力与模型性能之间的平衡，也成为业界持续面临的巨大挑战。

二、生成式人工智能风险管理——一项紧迫的议程

自 ChatGPT 发布以来，全球主要国家、国际组织、企业和研究机构纷纷提出人工智能治理举措和呼吁，强调加强人工智能风险管理，规范人工智能技术发展。

图 3：ChatGPT 之后，全球 AI 治理步伐加快

一方面，各国政府加快推进人工智能相关政策制定步伐。2023 年 4 月 3 日，中国科技部发布《科技伦理审查办法（试行）》（征求意见稿），提出涉及科技伦理敏感领域的，应设立科技伦理（审查）委员会，并建立伦理高风险科技活动的清单制度，对可能产生较大伦理风险挑战的新兴科技活动实施清单管理。4 月 11 日，中国网信办发布

《生成式人工智能服务管理办法（征求意见稿）》，旨在对生成式人工智能带来的风险做出及时应对。5 月 13 日，美国白宫总统科技顾问委员会（PCAST）成立生成式人工智能工作组，以帮助评估关键机遇和风险，并就如何更好地确保这些技术的开发和部署尽可能公平、负责任和安全提供意见。5 月 16 日，美国国会召开听证会，OpenAI CEO Sam Altman 应邀出席，讨论 “AI 监管：人工智能的规则”，并建议政府组建新机构、创建安全标准和指派第三方专家对 AI 系统进行审计等。5 月 16 日，法国 CNIL 发布行动计划，提出审计和监控人工智能系统。5 月 20 日，G7 领导人表示，“需要立即评估生成式人工智能技术的机会和挑战”，同意对人工智能采取“基于风险”的监管。日本政府召开“AI 战略会议”，讨论制定人工智能使用规则。6 月 1 日，澳大利亚官员表示，正计划加强对人工智能的监管，包括可能禁止深度造假和看似过于真实的虚假内容。6 月 6 日，国务院办公厅发布《国务院 2023 年度立法工作计划》显示，《人工智能法》已列入立法计划，草案预备年内提请全国人大审议。6 月 14 日，欧洲议会投票通过《人工智能法案》，标志着该法案朝着成为法律迈出关键一步；其中，法案增加了对生成式人工智能工具提出新的透明度要求。与此同时，联合国教科文组织呼吁各国应立即实施全球伦理框架，以应对人工智能挑战。世界卫生组织（WHO）也呼吁谨慎使用人工智能生成的大型语言模型工具。

另一方面，业界积极呼吁加强人工智能监管。2023 年 3 月，特斯拉首席执行官埃隆·马斯克（Elon Musk）、苹果联合创始人史蒂

夫·沃兹尼亚克（Steve Wozniak）以及其他上千名 AI 研究人员日前签署公开信，呼吁暂停研究比 GPT-4 更先进的 AI 技术。由微软等企业发起的商业软件联盟（BSA）公开发文，呼吁在国家隐私立法基础上制定管理人工智能使用的规则。同时，包括 OpenAI 在内的企业管理者认为需要建立国际人工智能监管机构，并且要尽快建立起来。对此，联合国秘书长安东尼奥·古特雷斯明确表示支持，并提出，计划在今年年底建立一个国际人工智能监管机构，定期审查人工智能治理工作，并对如何与人权、法治和公共利益保持一致提出建议。英国首相里希·苏纳克也对此积极回应，并表示希望英国成为全球人工智能安全监管的发源地。非营利组织人工智能安全中心呼吁将减轻人工智能带来的灭绝风险作为全球优先事项推进。

总体来看，以生成式人工智能治理为焦点的人工智能治理已经成为一项全球共同的紧迫议程。与一般性技术风险相比，生成式人工智能是一种能力强大、影响深远的变革性技术，并且许多潜在风险和问题的严重性已经需要引起足够重视。

三、生成式人工智能治理原则——基于现实的考量

当前，生成式 AI 正处于规模化落地的探索期，全球创新创业活跃、市场日新月异。然而，随着生成式人工智能的应用持续推进，一些具有现实意义的风险挑战也陆续显现，比如：

- “幻觉”现象，即一本正经的胡说八道，比如生成不真实存在的内容，已成为生成式人工智能应用进一步发展面临的重

点问题。

- 数据投毒风险，即通过对训练数据加入恶意数据的方式，在模型中植入后门，并通过特定输入触发后门，使模型输出错误的结果。
- 混淆攻击风险，即通过输入逻辑混淆的问题，使模型在回答时出现逻辑混乱。
- 诱导攻击风险，即通过情景对话、思维链引导等方式，绕过模型内置安全机制，进而使模型输出危险内容。
- 数据泄露风险，即通过特定的输入，诱导大模型输出训练数据集中个人身份识别信息、商业秘密等敏感数据。
- 用于网络诈骗和网络攻击等不法目的，不法分子可能利用自动生成诈骗话术、虚假语音、合成“虚拟角色”等实施电信诈骗。同时，ChatGPT 等工具降低攻击者的技术门槛。有机构表示，ChatGPT 将为网络钓鱼、虚假信息和网络犯罪提供便利。
- 版权保护、就业替代等社会性风险，例如，AI 利用 32 张图片便成功习得迪士尼画师的风格，以及好莱坞编剧抗议 AI 创作工具使其丧失就业机会等。

面对上述现实挑战，我们结合自身实践，认为生成式人工智能治理应当重点关注以下几项原则：

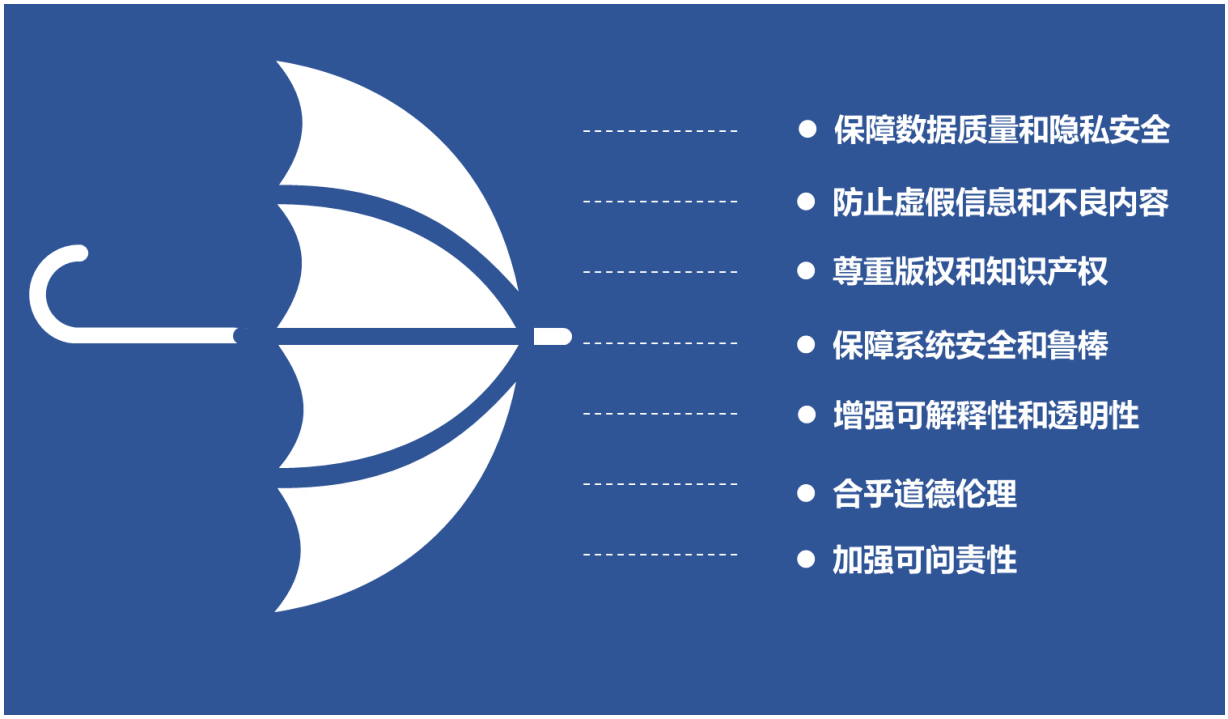


图 4：生成式人工智能治理的核心原则

- 保障数据质量和隐私安全

生成式人工智能需要大量的数据进行学习，然而，数据可能存在质量问题，例如数据缺失、数据噪声、数据偏见等。此外，数据获取也可能存在法律和道德问题，例如隐私泄露等。因此，为了确保生成式人工智能的质量和可靠性，我们建议采取以下治理措施：

一是保障数据质量。应通过数据清洗、数据标注、数据增强等方式，确保用于训练生成式人工智能的数据是高质量、全面且准确的。

二是保护数据安全与个人隐私。应通过加密技术、访问控制等技术方法，保护隐私数据、商业秘密，并确保隐私数据的使用合规、合乎道德规范。

- 防止虚假信息和不良内容

生成式人工智能可以生成类似人类创作的虚假信息、不良内容等，这些信息可能误导公众或造成不良影响，并且也会对后续模型的优化训练带来潜在危害。因此，我们建议采取以下治理措施：

一是审核和动态监测。应当建立健全生成内容质量评估体系和审核机制，对生成的内容进行审核和动态监测，确保其符合事实或预期。

二是识别虚假信息。应当采用技术手段和人工力量，例如通过训练模型、建立审核团队等方式，对生成的信息进行虚假识别和过滤。

- 尊重版权和知识产权

生成式人工智能可以生成类似人类创作的作品，例如音乐、图片、文本等，这些作品可能会涉及到版权问题。因此，我们建议采取以下治理措施：

一是加强版权保护。通过利用数字水印等相关技术，加强内容版权保护能力建设，保护生成作品的版权权益，并为版权溯源提供技术支撑。

二是建立许可和授权机制。对于生成的作品，应建立使用许可和授权机制，确保其符合知识产权法律和道德规范。

- 保障系统安全和鲁棒

生成式人工智能的应用会涉及处理大量敏感数据和任务，因此，确保人工智能系统的安全和鲁棒至关重要。为此，我们建议采取以下治理措施：

一是提升安全性。通过加强访问控制、加密通信、增量学习等方

式，建立完善的安全保障机制，保障生成式人工智能算法和系统的安全性。

二是提升鲁棒性和可靠性。通过容错机制、增量学习等方式，保证生成式人工智能的稳定性和可靠性，并建立健全监测和告警机制，及时发现和处理风险事件。

- 增强可解释性和透明性

生成式人工智能的算法模型结构复杂、参数庞大，因此，天然的具有黑盒属性。这可能导致人们对生成式人工智能的结果和决策产生疑虑和不信任。因此，我们建议采取以下治理措施：

一是提升可解释性。应通过设计更可解释的算法、可视化模型结构等方式使生成式人工智能的算法和模型更易于理解。

二是提升透明性。生成式人工智能的决策和结果应该透明、公开，例如通过记录模型训练和生成的日志、提供用户反馈机制等方式，让相关方能够了解其原理和过程。

- 合乎道德伦理

对于生成式人工智能的应用涉及到的偏见、歧视等道德和伦理问题，我们建议采取以下治理措施：

一是制定符合道德和伦理准则。制定道德和伦理准则，并通过建立伦理委员会和伦理审查机制等方式，规范生成式人工智能的研究和应用行为。

二是确保公平、公正。应当通过采取数据集偏见评估、标注人员管理、数据增强、公平性学习等方式，确保生成式人工智能算法和系

统的公平和公正性，避免偏见和歧视。

- 加强可问责性

考虑生成式人工智能具有的强大创造力，以及可能会参与辅助决策等相关场景，强化可问责性对于保障其可持续发展至关重要。因此，我们建议采取以下治理举措：

一是建立版本管理体系。应当建立健全模型版本管理，完整记录版本迭代信息，并持续进行跟踪监测和记录。

二是建立问责机制。应当健全产品全生命周期责任主体管理。对产品开发、测试、部署过程中相关的负责方进行记录，并通过数字水印等溯源技术，实现对文本、图片、代码、音频等数据责任方的溯源。

四、生成式人工智能治理实践——“商汤日日新 SenseNova”治理案例

2019 年，商汤在大模型领域和人工智能治理领域的工作同步启动。通过四年的持续努力，我们成功训练出全球规模最大的视觉大模型，并基于视觉大模型及多模态大模型的研发模式及技术积累，迁移到 NLP 领域的研发，完成了商汤自主研发的 NLP 中文大语言模型的训练和集成开发。我们在人工智能治理领域同样获得了多项重磅认可：

- 首份《AI 可持续发展白皮书》获得联合国《人工智能战略资源指南》收录；
- 获邀加入新加坡“人工智能验证基金会”（AI Verify Foundation），积极推动可信 AI 生态建设，树立具有国际社会

影响力的 AI 治理实践标杆；

- 荣获隐私信息管理体系认证全部三项认证，并获评中国人工智能产业发展联盟（AIIA）颁发的“可信 AI 2022 年突出贡献企业”。

2023 年 4 月 10 日，我们正式推出“商汤日日新 SenseNova”大模型体系，为行业提供自然语言处理、内容生成、自动化数据标注、自定义模型训练等多种大模型及能力。在“商汤日日新 SenseNova”大模型体系下，我们打造了商汤自研中文自然语言处理大模型应用“商量 SenseChat”，以及“秒画 SenseMirage”文生图创作平台、“如影 SenseAvatar”AI 数字人视频生成平台、“琼宇 SenseSpace”和“格物 SenseThings”3D 内容生成平台。

图 5：“商汤日日新 SenseNova”大模型体系

针对“商汤日日新 SenseNova”大模型体系，我们在传统“判别式模型”治理经验的基础上，结合生成式人工智能的风险特点，从数据、模型、内容三个层面初步构建起安全、可信的大模型治理“防护网”。

在数据层面，我们通过数据合规审查机制和严格的数据筛选及标注的业务逻辑，确保模型训练数据准确、合法、合规，并遵循相应的使用许可和版权规定。

- 在数据采集环节，我们建立了数据采集来源管理、数据采集业务评估、数据采集审批流程、采集合规审批等管理机制，确保数据采集的合规性、正当性和执行上的一致性。
- 在数据预处理环节，我们对收集到的原始数据进行清洗、去重、格式化等多步骤的预处理，以确保数据质量。并且，在此过程，我们会严格筛查，去除那些不完整、错误、带毒或含有敏感信息的数据。
- 在数据标注和筛选环节，我们通过自动化工具和人工相结合的方式，对预处理后的数据进行标注和筛选，以识别训练数据中是否包含敏感信息。此外，我们通过构建敏感内容反馈机制，利用内容生成本身特性，通过复用敏感内容的生成条件，丰富敏感鉴别模型的训练样本，持续提升模型性能。
- 在个人信息保护方面，“商汤日日新 SenseNova”大模型系列应用从设计、编码、测试、交付等阶段均设有个人信息保护的审核节点，对个人信息处理情况进行必要的检查，并严格按照法规规范要求实施个人信息保护。同时，我们根据 GB/T 35273-2020《信息安全技术个人信息安全规范》进行了全面的自评估对照、为指导并结合产品自身的技术特点及业务实现逻辑，将个人信息保护能力融入产品基本功能并面向最终

用户开放。

在模型层面，我们通过建构大规模测试数据集以及人工对抗测试的方式，对“商汤日日新 SenseNova” 大模型体系中的全部应用开展了系统的风险评估评测。

- 在商汤自研中文自然语言处理大模型应用“商汤商量 SenseChat”的评测过程中，我们重点对其准确性、鲁棒性、安全性和隐私性进行了测试评估。在准确性测试中，我们采用人工打分标注的形式，从整体评价、相关性、可读性、拟人性、专业性等五个指标对文本生成质量的进行评价；并从生成内容事实性错误，生成内容逻辑性错误，生成内容和问题相关性错误等三个方面对文本生成准确性进行评价。在鲁棒性测试中，我们通过同义替换、无关提示、引导性提示等方式，对生成内容的一致性、稳定性进行评价。在安全性测试中，我们通过引导性提示的方式对模型在政治敏感性和伦理敏感性两个方面的表现进行评测。在隐私测试中，我们采用引导性提示诱导模型输出隐私敏感信息。
- 在“商汤秒画 SenseMirage”文生图创作平台的评测过程中，我们重点对其准确性、鲁棒性、安全性和隐私性进行了测试评估。在准确性测试中，我们采用了 MS COCO、clip-g/14 等公开数据集对生成图片质量，以及图文匹配度进行了评测。在鲁棒性测试中，我们采用人工评价的方式，对模型在无语义意义的文本以及超长文本情景下的生成图像质量进行评估。

在安全性测试中，我们通过测试数据集，向模型提供政治敏感性和伦理敏感性输入，并基于模型的输出结果进行评估。在隐私测试中，我们使用训练数据的提示词进行输出查重的方式进行评估。

- 在“商汤如影 SenseAvatar”AI 数字人视频生成平台的评测过程中，我们重点对其准确性、鲁棒性和隐私性进行了测试评估。在准确性测试中，我们主要对关键表情匹配率以及嘴型幅度准确率两项关键指标做了测试评估。其中，关键表情匹配率指在测试音视频中如闭嘴、张嘴、嘟嘴等表情是否能在准确的时间点做出；嘴型幅度准确率指在全部测试音视频时间帧中预测的张嘴上下幅度、左右幅度、嘟嘴幅度是否与 GT 一致。在鲁棒性测试中，我们主要从模型对语速、音量大小、口音方言、语种等声音方面的鲁棒性，对环境噪声的鲁棒性，对输入画面的鲁棒性，以及对采集工具等设备的鲁棒性四个方面进行评测。在隐私安全性测试中，我们主要对其算法处理安全性，以及数据采集、传输、存储等方面的保密性做了评估。
- 在“商汤琼宇 SenseSpace”3D 内容生成平台的评测过程中，我们重点对其准确性、鲁棒性、安全性进行了测试评估。在准确性测试中，我们基于自采集数据集，对其场景的重建效果、清晰度和精度进行测试。在鲁棒性测试中，我们主要从场景重建的成功率、不同分辨率图像的重建质量影响、场景

中移动任务的干扰以及不同设备的渲染性能等维度进行了测评。在安全性测试中，我们主要考察了其对数据的加密和保护，以及对隐私场景的处理编辑能力。

- 在对“商汤格物 SenseThings”3D 内容生成平台的评测过程中，我们重点对其准确性、鲁棒性、安全性进行了测试评估。在准确性测试中，我们基于自采集数据集，主要从重建效果的逼真性和清晰度两个方向进行了测评。在鲁棒性测试中，我们主要从重建物体成功率进行了测评。在安全性测试中，我们主要考察了模型对政治敏感性和伦理敏感性输入的合理应对能力。

在上述评估中，“商汤日日新 SenseNova”大模型体系整体表现出与全球主要可比模型较为出色的风险应对能力。同时，出于安全考虑，我们将采取定向提供的方式披露相关评估结果。通过上述评估测试，我们对“商汤日日新 SenseNova”大模型体系的风险边界有了比较清晰的认识，并将相关测试结果反馈于模型强化学习的过程之中，帮助我们进一步提升模型风险防御能力。目前，相关模型体系仍处于邀约测试阶段，我们也将开展持续的跟踪测试，不断提升风险防御能力。

在内容层面，我们针对文本生成、图像生成、音频生成、视频生成等不同场景，建构了一套由深度学习算法驱动为核心的内容过滤工具，并通过自动化与人工相结合的方式对产品输入、输出的内容进行审核，确保生成内容的合规性与合乎伦理性。同时，以显式标记方式

告知用户该内容为利用深度合成技术生成或编辑得到的，并通过明确标记的方式有效告知合成内容，其中标识信息包括生成模型的信息、信息服务信息、合成标示；

此外，我们还建立了模型版本管理机制，并结合商汤自研的数字水印技术对生成的文本、图像，代码以及音视频内容进行溯源管理，系统提升“商汤日日新 SenseNova”大模型体系的可问责性。我们通过将传统频域技术以及深度学习技术相结合的水印算法将隐藏信息抽象为任意二进制信息流，嵌入到图片等多模态数据当中，在不对生成内容产生任何可感知影响的情况下，实现对生成内容的确权、溯源。

五、生成式人工智能治理基础设施——商汤“SenseTrust”工具体系

面对 AI2.0 阶段的风险挑战，加强全行业、全社会的人工智能风险治理能力已成为全球各方亟待解决的紧迫命题。**回顾人类科技发展史，我们从“保险丝”的发明中获得了启示。保险丝的存在能够避免因电流异常导致的电器损害；或许 人工智能走进千行百业也需要具备类似的“安全装置”。而这类安全装置，我们认为，其实就是覆盖数据处理、模型训练、模型部署，以及推理服务等 AI 系统全生命周期的治理工具。因此，我们正式推出“SenseTrust”——商汤可信人工智能基础设施，并将持续通过“商汤 AI 安全治理开放平台”等多种形式，为行业提供 AI 治理公益技术服务，推动建设安全可信的人工智能产业生态。**具体来看：

图 6： “SenseTrust”——商汤可信 AI 基础设施

在数据层面，商汤“SenseTrust”能够提供从数据脱敏、数据去毒、数据合规审查及偏见评估等治理工具。我们的敏感信息脱敏工具，能够面向活体检测、车牌检测、文字文档信息检测等广泛应用场景，提供高水平的数据脱敏技术，并且具备接口灵活，平台覆盖面广，支持实时脱敏等优势。数据脱敏模块对涉及个人信息的敏感数据提供数据脱敏服务。数据脱敏的范围包括但不限于生物特征数据，自动驾驶场景的敏感数据等。数据脱敏服务还可根据实际业务需求实现是否具备重标识的能力，在特定场景下可还原已去标识化的敏感数据。我们的数据祛毒工具能够在数据预处理环节对训练数据进行带毒性检测，判定数据是否存在异常，接着进行毒性判定，并根据投毒类别做不同的祛毒方案，同时进行溯源调查。

此外，面向数据要素可信流通，我们创新打造了“数据沙箱”工具。通过沙箱包装后，结合隐私计算集群协同调度，实现数据可用不可见，在保证数据隐私安全的前期下实现数据价值转化，促进数据要素流程利用。目前数据沙箱可面向两个应用场景：一是多用户拥有不

同场景分布的数据，提供联合训练方案，并且具有携带离线模型可以完成不泄露数据的反演；二是针对用户端拥有大量数据的场景，可使用数据加密训练方案，可以在保护隐私的前提下完成数据回流。

在模型层面，商汤“SenseTrust”基于自研的模型体检系列平台，能够针对传统“小模型”、生成式“大模型”，以及基础模型提供标准化和定制化的模型评测能力。我们针对传统“小模型”开发的模型体检平台，能够面向活体识别、图像分类、目标检测等商业化需求提供一键式评测，用户只需提供模型和评测数据即可进行。目前已在商汤的大量商业化模型检测方面获得验证。模型体检内容包括对抗安全、鲁棒安全、后门安全、可解释性和公平性评测。同时，我们针对生成式“大模型”和基础模型测评建构了百万体量的测试数据集，能够实现对大模型的伦理属性、安全属性，以及模型能力的评测评估。

针对模型体检出的问题，商汤“SenseTrust”还能够进一步提供模型加固解决方案，主要包括鲁棒性训练和 AI 防火墙两个部分。鲁棒性训练模块可以在不损失精度的情况下强化模型的安全性和鲁棒性，当前主要包括对抗训练和针对性的数据增强。鲁棒性训练模块是模型开发的代码插件，已融入商汤目前的模型开发流程。AI 防火墙模块主要用于过滤可疑攻击样本，可以在不重新训练模型的情况下提升模型部署的安全性。当前 AI 防火墙可以有效抵御主流的黑盒攻击和物理攻击方式。AI 防火墙和部署的质量模型相结合，在提升安全的同时不引入格外的计算开销。

在应用层面，我们在涉及数据保护、数字取证及伪造检测等技术

领域有着深厚的积累，并逐步开发了基于生成、鉴伪和溯源三位一体的综合解决方案。

在深伪鉴别方面，商汤“SenseTrust”提供包括数十种先进攻击手段的伪造生成平台，为鉴伪检测和溯源提供丰富多样的攻击案例和海量数据支持。并可通过持续集成先进伪造算法，在 zero/few-shot 场景下快速响应难例样本和长尾类型，帮助提升鉴伪算法的泛化性。

商汤“SenseTrust”伪造检测大模型，可充分利用面部表情一致性、动作序列连贯性，并结合频谱、声音和文字等多模态信息，准确鉴别包括 PS、换脸、活化以及各种先进扩散模型（如：Stable Diffusion）合成的高清人像。主流评测数据集上算法检测精度可达到 99%以上，在应对新技术复合伪造方法上（如：通过 MidJourney），检测能力也高出行业同类产品 20%以上。

同时，我们通过自研基于解耦-重建的伪造检测算法，能够从伪造数据中分离出真实内容及伪影痕迹。在针对 10 余种主流伪造算法溯源上，准确率超过 90%，同时还可给出数据中的相关伪造痕迹，提高检测算法的可解释性和可信度。这一技术为行业首创，并作为数字取证技术成功落地司法领域。

目前，商汤“SenseTrust”综合鉴伪解决方案已投入实战，为十余家银行的安全系统提供服务，对各类灰黑产攻击拦截成功率超行业同类产品 20%以上，有效防范了灰黑产身份盗取、支付盗刷等网络诈骗。

此外，针对当前各方关注的 AIGC 相关确权溯源和内容保护问题，商汤“SenseTrust”具备数字水印解决方案。商汤数字水印结合频域分

析、深度学习、扩散模型等技术，将特定信息嵌入到数字载体中，同时不影响载体的使用价值，也不易被人的知觉系统察觉，只有通过特定的解码器和专属密钥才能提取，可实现篡改内容的检测且水印不可窃取。数字水印技术可在 AIGC 相关产品发布时加入，能够有效增强深伪检测的可靠性，甚至进一步影响生成效果，从源头上遏制深度伪造，实现主动防御。

具体应用中，商汤数字水印技术可用于版权保护，防伪溯源等场景，支持图像、视频、音频、文本等各种模态的数字载体，在不同程度的干扰下(裁剪、压缩等)能保证 99%+的水印提取精度，且不影响数据本身质量(如高清图画质)，在保证水印信息容量大(256 位)以及安全性(通过密钥加密)的同时具备足够的隐蔽性以及鲁棒性。目前，我们的数字水印技术已服务于“商汤秒画 SenseMirage”、“商汤如影 SenseAvatar”等多个产品，以及内容创作、大数据客户。

六、生成式人工智能治理的发展——避免陷入“失控的竞赛”

2023 年 3 月，非营利性组织“生命未来研究所”（Future of Life Institute，FLI）发布公开信，信中指出，当前，全球领先的人工智能实验室正处于一场失控的竞赛之中，并呼吁暂停巨型 AI 实验。公开信一经公布，便引起全球各方广泛关注和激烈讨论，Elon Musk、Yoshua Bengio、Steve Wozniak、Emad Mostaque 等学术和行业领军人士纷纷签署表达支持。FLI 官方网站公布的数据显示，截至 2023 年 7 月 7 日，全球已有 33000 余人署名支持该倡议。

事实上，过去十年，人工智能领域一直处于两种持续加速，并且总体良性的“竞赛”之中。一种是技术方法上的“竞赛”，另外一种则是治理路径上的“竞赛”。每一次技术赛道上取得的突破都会推动一次治理“竞赛”的加速。例如，2016 年 3 月，DeepMind 开发的 AlphaGo 在围棋比赛中战胜韩国棋手李世石，成为第一个战胜围棋世界冠军的人工智能系统，引发各方对人机关系的深层审视。同年 5 月，欧盟便发布了第一份有关人工智能治理问题的文件——“机器人民法规则”报告草案；9 月，Google、Facebook、IBM、Amazon、Microsoft 等美国企业共同发布“Principles of Partnership on AI”；10 月，美国白宫发布报告——“为人工智能的未来做准备”，关注人工智能的长期影响，并在《美国国家人工智能研究与发展战略规划》中明确提出符合伦理道德的设计，以及确保安全可靠等治理要求。由此，全球人工智能治理进程正式开启。

当前，生成式人工智能的发展仍处于早期阶段，在全球范围内，生成式人工智能治理已经成为一个重要议题。**我们认为，加强人工智能治理，尤其是针对生成式人工智能存在的潜在风险采取更广泛的共识行动，是一项务实且事关长远的工作。我们原则上，对全球各方目前已采取或计划采取的行动表示赞赏。同时，鉴于生成式人工智能当前所处的发展阶段，我们在此诚挚地呼吁各方在推动人工智能发展与治理的过程中应保持“动态平衡”，无论是技术发展本身、还是对技术的风险治理，都应避免陷入“失控的竞赛”之中。为此，我们愿就生成式人工智能治理的推进提出以下几点倡议：**

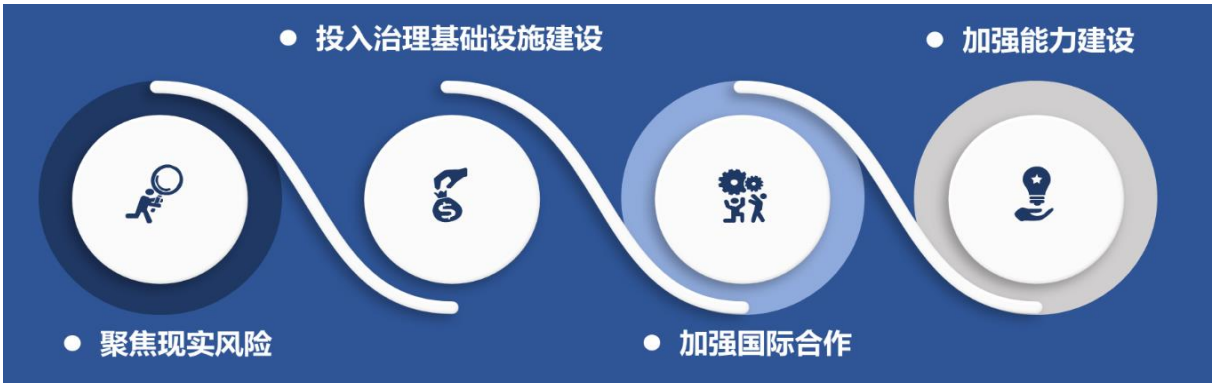


图 7：商汤科技关于生成式人工智能治理的倡议

一是准确定义问题，聚焦现实风险。针对实践中已发生或者观察到的风险问题，应加快推进相关政策制定工作，明确开发和应用的底线要求，特别是就数据隐私、知识产权、反垄断，以及模型的基本安全属性和规范应用做出明确、可操作的要求。

二是加大治理基础设施建设投入。科研机构和企业应进一步加强生成式人工智能治理技术研究和工具创新，提高生成式人工智能的安全性、可靠性，以及合乎伦理性。具体来说，可以通过研究新的算法和技术，减少模型偏见和歧视的可能性；开发更加安全的计算环境和工具，以防止模型窃取和数据泄露等相关风险；同时，探索新的商业模式和服务模式，以满足不同用户的需求和期望。

三是加强人工智能治理国际合作。各国监管机构应加强生成式人工智能监管协调，共同应对生成式人工智能带来的全球性挑战；确保在全球范围就生成式人工智能治理达成基线水平的同时，降低生成式人工智能跨境合作与治理成本。具体来说，可以通过分享最佳实践和技术经验，促进知识共享和技术转移；协调政策和法规制定，以确保

各国之间的一致性和互信；以及支持发展中国家和地区的技术能力和人才培养，以缩小数字鸿沟和发展差距。

四是加强全社会治理能力建设。具体来说，可以通过向公众普及有关生成式人工智能的基本知识和技能，帮助人们更好地使用和管理这些模型；加强对媒体和社会舆论的引导和监督，以避免虚假信息和误导性言论的传播。通过对公众开展有关生成式人工智能风险的培训教育和意识普及，持续提升全社会参与生成式人工智能风险治理的能力。

生成式人工智能治理将是一个复杂而长期的过程，需要政府、企业、学术界和公众的共同参与。在当前发展阶段，我们对生成式人工智能潜在风险的理解和应对还有许多可以提升的空间。商汤期待与全球利益相关方加强交流合作，推动发展负责任且可评估的人工智能，促进人工智能为人类社会的进步做出更大贡献。

报告编委会

薛澜　清华大学人工智能国际治理研究院院长

季卫东 上海交通大学中国法与社会研究院院长

徐立　商汤科技联合创始人、董事长兼首席执行官

张望　商汤科技副总裁、人工智能伦理与治理委员会主席

杨帆　商汤科技联合创始人、副总裁

骆静　商汤科技副总裁、首席运营官

金俊　商汤科技首席营销官

张少霆 商汤科技副总裁、研究院副院长

林洁敏 商汤科技副总裁

作者

胡正坤

商汤科技

人工智能伦理与治理研究主任

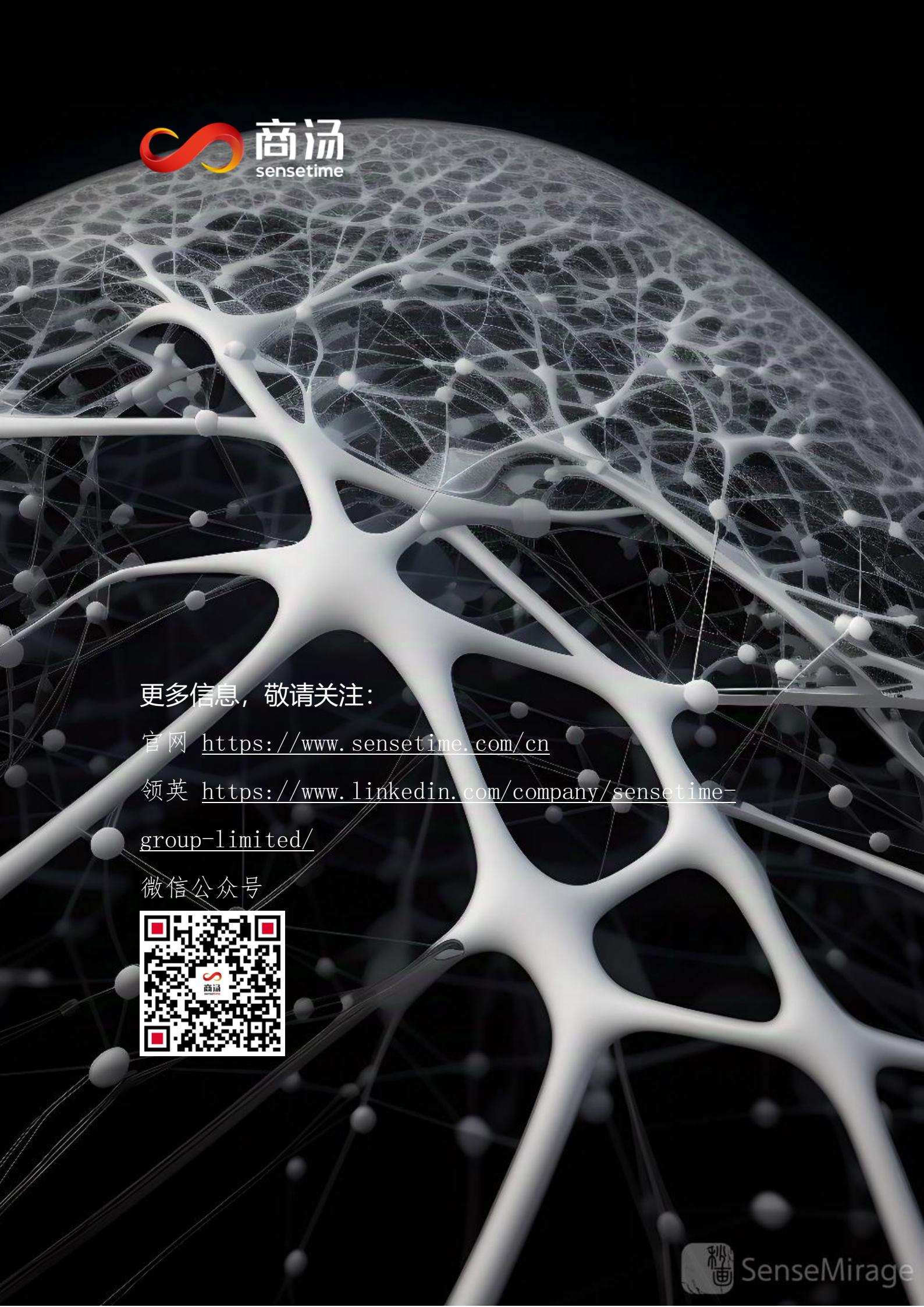
田丰

商汤科技

智能产业研究院院长

特别鸣谢

李学尧、林喜芬、吴一超、石华峰、秦昊煜、姚程元、王宇航、刘国帅、崔志超、胡琨、高梦雅、许晨晔、王义飞、梅莹、朱文辉、骆卓昱、路少卿、王岩桦、姜文睿、闫欣桐、吴晶彧、成瑾、梁蓉蓉、綦伟良


商汤
sensetime
更多信息，敬请关注：
官网 https://www.sensetime.com/cn
领英 https://www.linkedin.com/company/sensetime-group-limited/
微信公众号
SenseMirage