“平衡发展”的人工智能治理白皮书

——商汤人工智能伦理治理年度报告

(2022 年)

人工智能伦理与治理委员会

2022 年 9 月

目录

【关于商汤】

作为人工智能软件公司，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 2022 年 9 月以中英文版本面向全球发布，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：

电子邮箱：aiethics.committee@sensetime.com

一、人工智能发展与治理概述

“2010 年前后，人工智能步入新发展阶段，算力和数据成为主要驱动。此阶段的人工智能，某种程度上不再以人的认知为依据，其规则跳出了当前人的认知边界，引发更多有关治理的探讨” ——商汤科技 CEO，徐立

过去十年，人工智能技术在深度学习、大数据，以及“摩尔定律”的支撑下取得一个又一个突破性进展，在计算机视觉、自然语言处理、语音识别等不同细分领域实现商业化落地。截至目前，人工智能已在城市治理、教育、金融、医疗、零售、交通、文娱、制造等众多场景获得广泛应用，并正加速向科学研究、文艺创作等知识拓展类场景渗透。人工智能作为一项通用目的技术，正在被广泛的认可和采用，一个泛在智能的时代正在加速到来。

纵观历史，通用目的技术在推动社会生产力实现跃迁的同时，不可避免地会对原有的社会生产关系带来革命性影响。这一规律同样适用于人工智能，尤其是，当数据驱动的人工智能突破人类既有认知的边界时，其对原有社会生产关系形成的冲击将比以往任何一次技术革命都要强烈。即便在尚处于“弱人工智能”阶段的当下，人们对自动化决策系统鲁棒性、公平性的关切，对算法推荐、深度合成以及生物特征信息识别技术滥用的担忧，无不清晰地表明，如果不能尽快就人工智能的发展目标、发展方式以及应用规范等问题达成广泛共识，人工智能技术走向规模应用将持续面临信任挑战。

正因如此，人工智能治理在过去十年里同样获得了企业、政府机

构、国际组织、社会团体等多利益相关方的高度重视，并取得显著进展。**目前，人工智能治理已经进入落地实践的阶段。从发展历程看，人工智能治理至今已经走过了三个阶段：**

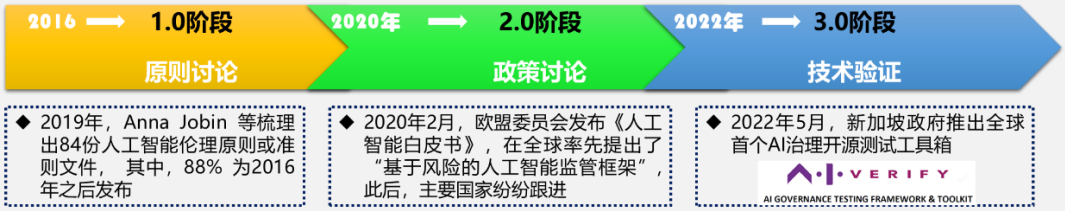

图 1 全球人工智能治理发展历程

资料来源：商汤智能产业研究院

- **人工智能治理的 1.0 阶段：起于 2016 年，以原则讨论为主。**哈佛大学 Jessica Fjeld 等[1]在研究中将人工智能治理 1.0 阶段的开端定位于 2016 年 9 月，由 Google、Facebook、IBM、亚马逊和微软等美国企业发布的《Principles of Partnership on AI》。Anna Jobin 等[2]通过梳理来自全球 84 份人工智能伦理相关的原则或准则文件同样发现，2016 年之后发布的占比高达 88%；其中，私营企业和政府部门发布的文件数量分别占比 22.6%和 21.4%。
- **人工智能治理的 2.0 阶段：起于 2020 年，以政策讨论为主。**2020 年 2 月，欧盟委员会发布《人工智能白皮书》，在全球率先提出了“基于风险的人工智能监管框架”。此后，主要国家纷纷跟进，均在不同程度开展了监管人工智能相关技术及应用的探索。因此，2020 年也常被称为“人工智能监管元年”。

[1] https://cyber.harvard.edu/publication/2020/principled-ai
[2] https://doi.org/10.1038/s42256-019-0088-2

据经合组织（OECD）的统计[3]，全球已有 60 余个国家提出了 700 余项人工智能政策举措，而德勤全球(Deloitte Global)则进一步预测[4]，2022 年将会有大量关于更系统地监管人工智能的讨论。

- **人工智能治理的 3.0 阶段：起于 2022 年，以技术验证为主。**进入 2022 年，随着全球人工智能治理进程的持续推进，以及可信、负责任人工智能等相关理念的持续渗透，有关验证人工智能治理如何落地实施的倡议日益增多。在政府侧，2022 年 5 月，新加坡政府率先推出了全球首个人工智能治理开源测试工具箱-“AI.Verify”；2022 年 6 月，西班牙政府与欧盟委员会发布了第一个人工智能监管沙箱的试点计划。在市场侧，美国人工智能治理研究机构 RAII 发布了“负责任人工智能认证计划”，向企业、组织、机构等提供负责任 AI 认证服务。

面向 3.0 阶段，我们认为，人工智能治理的技术验证应当包括两个层面：一是通过实践验证原则、政策要求和标准的可落地性，二是通过采用技术工具验证各方对人工智能治理规范的落实程度。未来，深入推进人工智能治理还应重点处理好以下“三组关系”：

[3] https://www.pymnts.com/news/regulation/2022/oecds-principles-can-guide-governments-to-design-ai-regulatory-frameworks/

[4] https://www2.deloitte.com/global/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/ai-regulation-trends.html

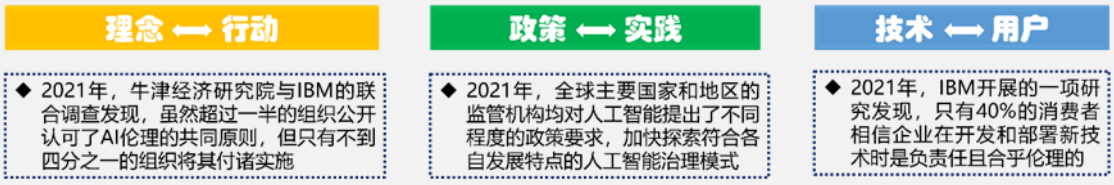

图 2 人工智能治理落地应重点关注“三组关系”

资料来源：商汤智能产业研究院

- **理念与行动的关系**。IBM 的一项研究[5]发现，虽然超过半数接受其调研的组织发布了或公开认可了人工智能伦理的共同原则，但只有不到四分之一的组织将其付诸实施。人工智能伦理治理在实践过程中仍面临诸多现实挑战：一是人工智能治理与现有组织结构的融合问题。人工智能治理工作涉及信息安全、数据治理等与现有组织结构重叠的领域，职责交叉、工作范围难以厘清等问题在组织层面对推动落地实践形成了一定的制约；二是伦理治理尚未真正融入企业的业务价值闭环。在推动人工智能伦理治理工作时，业务方可能会因看不到短期收益，出现对伦理治理工作重视程度降低、治理工作落实缺位等情况；三是伦理治理落实缺乏共识性标准的问题。
- **政策与实践的关系**。政策制定者与技术开发者的视角不同、立场不同，对政策意涵的理解也不尽相同，在推动人工智能治理落地的过程中，还需要将政策要求转化为技术和业务团队可执行的实践标准。处理好政策与实践的关系，应重点考虑以下四方面：一是政策制定应考虑产业、技术应用的动态性和多样性，为行业发展提供一定宽松的环境；二是不同机

[5] https://www.ibm.com/downloads/cas/VQ9ZGKAE

构、部门、国家应努力推动AI治理标准的互联互通；三是应加强AI治理实践部门在政策制定过程中的参与程度，提升政策要求的可落地性；四是政策层面与产业层面在AI治理相关问题的定义方面应当寻求更多共识。

- **技术与用户的关系**。目前，AI 治理仍主要局限于专业讨论、企业内部治理领域，尚未将最终用户纳入AI治理的工作体系之中。因此，市场或者社会上对AI相关的治理问题还存在不少的误解，有些用户会将暂时的技术问题定义为长期性的治理问题，例如，将新部署的AI系统暂时性的识别效果不佳的问题理解为算法歧视问题。IBM的研究[6]发现，只有40%的消费者相信企业在开发和部署新技术时是负责任且合乎伦理的，用户对AI治理缺乏深度认识。因此，在推动AI治理落地的过程中，还应处理好技术与用户间的关系。为此，技术提供方应使用用户语言讲解技术，AI治理机构应主动厘清治理挑战的根源、深化用户对AI治理问题的理解，并且提升用户参与AI治理的能力。此外，企业还应加大对治理技术工具的投入，提升AI治理的可验证性。

二、商汤的人工智能治理思考

“今天，科技与人类活动深度结合。我们需要科技伦理的规范和人文精神的熏陶，促进科技朝着有益于人类的方向健康发展。因此，

[6] https://www.ibm.com/downloads/cas/VQ9ZGKAE

我们十分重视人工智能伦理治理” ——商汤科技 CEO，徐立

长期以来，商汤始终高度重视人工智能的治理问题，自 2019 年起，便在内部建立并持续维护“全球 AI 伦理风险库”，涵盖上百项全球人工智能伦理正面及负面事件、案例分析等。在密切跟踪全球人工智能治理发展趋势的同时，我们不断深化自身对人工智能治理的认识和理解，并逐步形成对人工智能治理问题的系统性思考。

基于对全球风险案例的深入分析，**我们认为，人工智能时代的风险挑战主要来自数据、算法和应用管理三个层面，具体来说：**

- **数据层面**的风险主要涉及个人信息保护、数据治理与数据质量三个方面。其中，个人信息保护风险是指在人工智能的开发、测试、运行过程中存在的隐私侵犯问题，这一类问题是当前人工智能应用需要解决的关键问题之一。数据质量风险是指用于人工智能的训练数据集以及采集的现场数据可能会存在的质量问题，以及导致相应的不良影响，这也是人工智能特有的一类数据风险问题。数据安全风险是指人工智能开发及应用过程中，企业对持有数据的安全保护问题，其涉及数据采集、传输、存储、使用、流转等全生命周期，以及人工智能开发和应用等各个环节。
- **算法层面**的风险主要涉及算法决策、算法“黑箱”与算法安全三个方面。其中，算法决策风险是指因算法推理结果的不可预见性与人类自身的认知局限，导致无法预测智能系统做出的决策原因与产生的效果，归责问题便是此类风险的典型

代表。算法“黑箱”风险主要指因采用复杂神经网络的算法导致决策不透明与无法被充分解释，而造成的可解释性风险。算法安全风险是指因模型参数泄露或被恶意修改、容错率与弹性不足引发的风险。

- **应用管理层面**的风险涉及算法歧视、伦理冲突、劳动竞争等多个方面。例如，因人为原因、原始训练数据存在偏见性、机器在自我学习过程中输入了数据的多维不同特征，而造成算法将偏见引入决策过程的算法歧视风险；因算法设计者出于自身的利益，对用户进行不良诱导、过度依赖算法本身、盲目扩大算法的应用范围而导致的算法滥用风险；以及人工智能在长期对于就业的负面影响，加剧不正当竞争与市场支配地位，引起责任界定的困境与风险。

基于对全球人工智能治理诉求的分析，并结合自身对技术和市场的理解，**我们认为，人工智能治理应当是一个价值牵引、技术先行、多方参与、分层推进的动态进程。从治理效果看，人工智能治理的实现由低至高，可分为可用、可靠、可控、可信四个层次。**

- “可用”是指人工智能系统在功能、性能层面应当能够满足应用场景需求。
- “可靠”是指人工智能系统在安全性、鲁棒性层面应当能够满足部署环境和可持续运营要求。
- “可控”是指人工智能系统在功能层面应当能够充分保护人类的自主意志和权利、保障人类对系统的控制权。

- “可信”是指人工智能系统的设计和应用应当符合人类的价值理念和伦理道德。

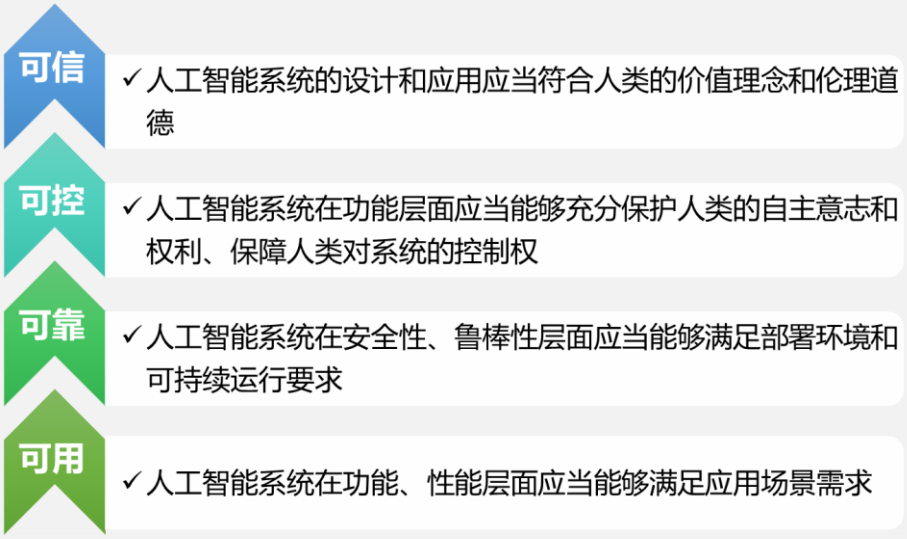

图 3 人工智能治理的“分层目标”

资料来源：商汤智能产业研究院

三、商汤的人工智能治理理念

“人工智能的发展和治理要并行。治理如果过早出现，可能会限制人工智能的发展；但若过于滞后，亦可能带来灾难性后果，且抬高后续治理成本” ——商汤科技 CEO，徐立

从技术成熟度及行业普及度的角度观察，人工智能技术及其应用仍然处于发展初期，与人工智能相关的经济形态、产业生态、商业模式还处在探索阶段。与历史上涌现的其他通用目的技术一样，人工智能技术及其相关产业的健康和可持续发展，既需要符合技术产业发展特点的创新空间，也需要规则层面的正确引导和保障。因此，**我们认为，人工智能的发展与治理应当并行，两者的功能和关系正如汽车的“发动机”和“方向盘”一样，相辅相成、缺一不可。**

基于以上对于人工智能发展与治理的认识，**我们于 2021 年正式提出了“平衡发展”的人工智能伦理观，并进一步明确了“以人为本、技术可控、可持续发展”的伦理原则**。具体来说，“平衡发展”的伦理观就是倡导，统筹推进人工智能的治理和发展，以人工智能治理促进人工智能产业的健康可持续发展和经济社会的数字化转型、智慧化升级。

- “以人为本”就是要追求不同文化之间的道德共识，尊重、包容、平衡全球不同国家地区的历史、文化、社会、经济等方面的发展差异，确保人权和个人信息保护，以及无偏见地应用技术。
- “技术可控”就是要确保人工智能由人类开发、为人类服务、受人类控制，相应地，其人工智能应用导致的伦理责任也应由其控制者（人类）承担。
- “可持续发展”就是要促进社会、经济、文化及环境的可持续发展，崇尚开放及包容合作，积极探索创新及可持续的人工智能治理模式的应用。

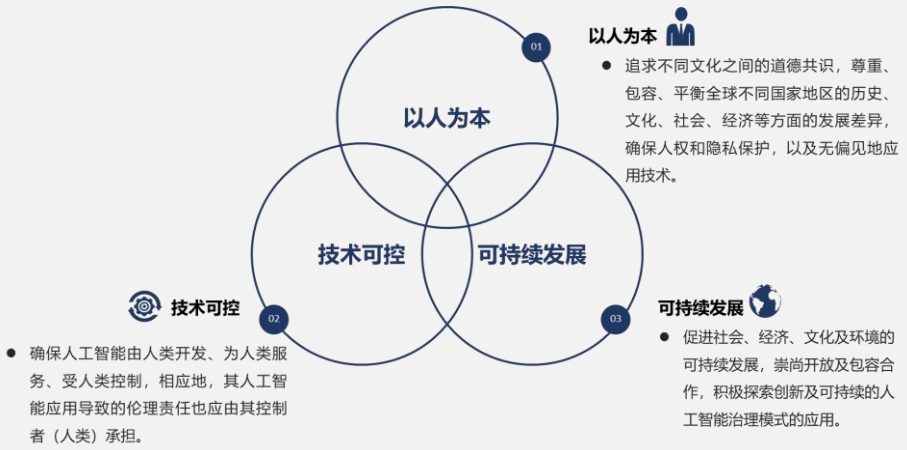

图 4 商汤人工智能伦理原则的核心意涵

资料来源：商汤智能产业研究院

同时，基于“平衡发展”的伦理观，**我们围绕“以人为本、技术可控、可持续发展”的伦理原则，进一步提出了合乎伦理的产品设计要求：**

- 尊重人权。应保护人类自由、尊严等基本权利，维护全球公认的道德伦理，提升人类职业发展与生活体验，不得损害人类的主体地位。
- 向善。应促进人类可持续发展，保护弱势群体利益，不应违背人类伦理道德的基本方向，在使用过程中合理合法合规。
- 无偏见。使用的数据应该保持相对的客观中立、具有整体的代表性，应兼顾普遍适用性和特殊人群需求。
- 保护隐私。对个人信息的收集与使用应坚持最小必要原则，特别是对个人敏感信息的处理应基于个人信息主体的明示同意，或法律规定的特定情形。
- 可靠可控。在一定时间内、一定条件下可以无故障地实现特定的功能；并且，在故障情况下，能实现有效停止和人类接

管。

- 透明可解释。应优先采用可解释性强的算法模型，应向用户提供清晰、易懂、充分的产品运行机制说明，并明确告知用户产品的局限性及潜在的风险。
- 可验证。算法模型以及结果均应可以在同等或近似的条件下被重复验证。
- 可问责。应明确研发、设计、制造、运营和服务等各环节主体的权利义务，应具备相关机制实现对输出结果背后的模型和数据进行追溯。

四、商汤的人工智能治理目标

“随着新技术的兴起，科技伦理面临许多新的课题，需要全社会各界联合加强研究。商汤科技作为业内领先企业，做好人工智能伦理相关工作责无旁贷”——商汤科技副总裁，人工智能伦理与治理委员会主席， 张望

“AI 伦理风险的边界和核心要求是发展负责任的人工智能。在治理和合规上企业要先想一步，先做一步”——商汤科技联合创始人、副总裁，人工智能伦理与治理委员会委员，杨帆

信任是商业社会运行的底层逻辑，也是新兴技术获得广泛认可的基本前提。商汤作为一家人工智能领域的科创企业，向来将市场与用户的信任视为自身发展的“生命线”。因此，我们从创业之初便将负责任地开发与部署人工智能作为指导一切行动的根本遵循。同时，我

们认为，负责任的人工智能不仅仅是一项原则要求，也应当是具体的、可实施的，而实现这项目标的关键就在于人工智能治理体系的建构。

随着对人工智能治理探索的深入，我们深刻地认识到，人工智能治理体系的建构不应停留在“口头上”、“纸面上”，还应做到有迹可循、有据可查。因此，**我们在业界首次提出，发展“负责任且可评估”的人工智能，并将其作为我们开展人工智能治理的愿景目标。具体来说，“负责任且可评估”的人工智能应当满足以下基本要求：**

- 对人负责。人工智能系统应当充分尊重和保护人的尊严与权利，有利于提升人的健康福祉。
- 对社会负责。人工智能系统应当充分尊重、适应不同社会文化的风俗习惯，有利于促进社会的健康可持续发展。
- 对环境负责。人工智能系统的开发应当充分考虑对自然环境的影响，其使用应当有益于环境的可持续发展。
- 责任可溯源。人工智能系统应当具备充分的可溯源能力，确保不同的子模块和环节具有明确的责任方。
- 风险有评估。人工智能系统在上线部署前应当通过充分的伦理风险评估审查。
- 过程可评估。人工智能系统的全生命周期及相应的治理过程应当具备完善的技术日志和文档记录。
- 效果可评估。人工智能系统在全生命周期中对人工智能治理要求的落实应当具备客观的支持凭据。

图 5 “负责任且可评估的人工智能”的核心特点

资料来源：商汤智能产业研究院

五、商汤的人工智能治理实践

（一）组织建设

商汤是业界最先组建人工智能伦理与治理委员会并将伦理治理工作作为公司战略级方向的科技创新企业之一。

为系统应对数据、算法及应用等不同层面的人工智能伦理风险，2020 年 1 月，商汤正式成立“人工智能伦理与治理委员会”，统筹推进人工智能伦理治理工作体系建设。在人员构成上，人工智能伦理与治理委员会由内、外部委员共同组成，其中，外部委员 2 名、内部委员 4 名，他们均来自技术、工程、法律、伦理等相关专业背景。同时，委员会下设秘书处、专家顾问组和执行工作组，保障伦理治理与审查工作的独立性、透明性、专业性和有效性。此外，为确保人工智能伦理委员会的高效运转、强化全体员工对人工智能伦理治理的贯彻落实，我们先后制定了《人工智能伦理与治理委员会管理章程》、《商汤集团伦理治理制度》等相关配套管理制度。

图 6 “人工智能伦理与治理委员会”的工作职责及内容

资料来源：商汤科技

与此同时，为强化数据和算法层面的伦理风险治理，商汤全面打通内部组织结构和工作流程，赋予集团“信息安全管理委员会”开展个人信息及隐私数据保护评估和算法安全评估的工作职责。

（二）机制建设

目前，针对人工智能领域面临的数据、算法及应用管理风险，商汤建立了覆盖产品全生命周期的风险控制机制，初步形成了人工智能治理闭环：

在应对数据风险方面，我们建立了“个人信息及隐私数据影响评估机制”，并设置“数据安全与个人信息保护委员会”，对涉及个人信息及隐私数据获取、存储、传输、处理的产品开展个人信息及隐私数据影响评估，确保产品遵循默认符合个人信息保护的设计要求。

在应对算法风险方面，我们建立了“算法定级备案管理与安全评估机制”，成立了“算法安全管理工作组”，依据算法的数据类型、业务场景、伦理风险等级、数据质量、存储现状、用户规模、数据重要程度、对用户行为的干预程度将算法进行分类分级备案管理，并从技

术局限、算法设计、软件缺陷、数据安全、框架安全等相关维度对算法风险进行安全评估。

在应对应用管理风险方面，我们建立了“伦理风险分级分类管理机制”，设立了“伦理风险审核小组”，对产品设计、开发、部署、运营的全生命周期实施分阶段、分目标的伦理风险分级分类管理，并建立了配套的风险自查、评估、审查和跟踪审查流程。我们依据产品对最终产品安全、个人权益、市场公平、公共安全和生态安全的影响程度，将伦理风险由低至高划分为 E0 至 E4，五个等级：

- E4 级产品，即禁止类产品。指背离商汤伦理原则、违反法律法规要求的人工智能产品。
- E3 级产品，即伦理高风险产品。指直接关系最终产品安全、个人权益、市场公平、公共安全和生态安全的产品。
- E2 级产品，即伦理中风险产品。指对最终产品安全、个人权益、市场公平、公共安全和生态安全具有间接或潜在重要影响的产品。
- E1 级产品，即伦理低风险产品。指对最终产品安全、个人权益、市场公平、公共安全和生态安全不具备明显影响的产品。
- E0 级产品，即伦理无风险产品。指不包含机器学习算法、不具备人工智能功能的产品。

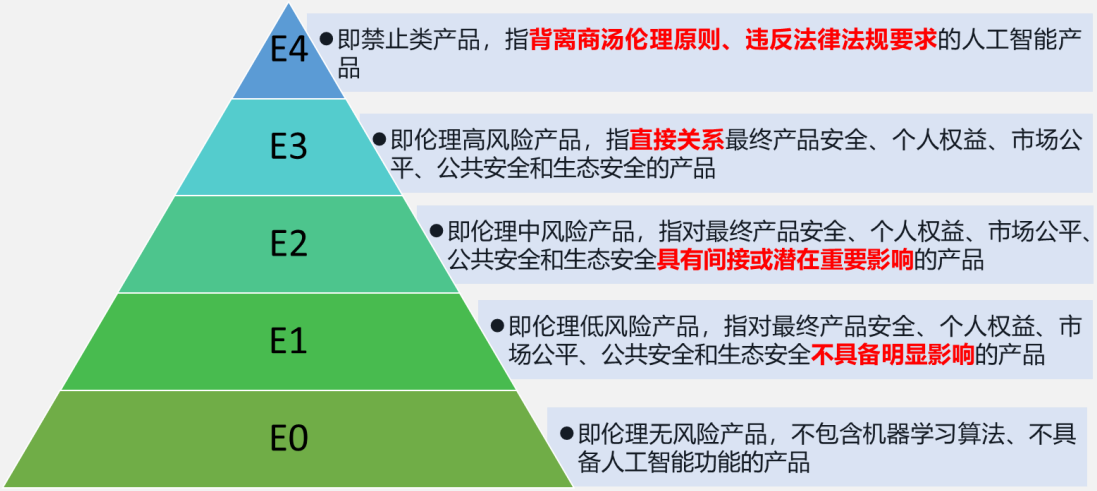

图 7 产品伦理风险分级分类管理评估标准

资料来源：商汤智能产业研究院

此外，为保障人工智能治理体系的有效落实、推动形成伦理治理“人人有责”的企业文化，我们还建立了伦理风险目标管理机制、伦理风险事件处置机制和伦理治理质量控制机制。

（三）工具建设

为推动发展“负责任且可评估”的人工智能，我们在内部开发了一系列覆盖数据治理、算法测评、模型体检、伦理审查的流程工具和技术平台。

在应对数据风险方面，我们开发了统一的数据治理平台以及规范的数据采集流程保证流入的数据准确性、均衡性、合理性，并依托于数据个人信息保护平台实现对于数据的隐私加密，全方位保证数据可用性、可靠性和安全性。同时，我们设计了一套产品研发全流程的个人信息保护评估清单，推动面向个人信息保护的产品功能设计，确保人工智能产品设计流程，使收集和处理（包括使用、披露、保留、传输和处置）限于所确定的必须的目的。

此外，我们在数据处理过程中，通过开发、使用自动化标注工具，

减少人工接触的数据量，在模型训练的源头降低引入人类偏见的风险，并且，数据标注平台具备访问控制和身份验证功能，仅能由认证的数据标注人员进行访问。

目前，我们通过了多项网络安全及数据安全国际性权威资质，包括：ISO/IEC 27001:2013 信息安全管理体系国际认证、ISO/IEC 27701:2019 个人隐私信息管理体系（PIMS）标准认证、ISO/IEC 29151:2017 个人信息保护实践指南、BS10012 个人信息安全管理体系认证，所售产品也获得了重要信息系统等级保护三级认证、可信人脸认证专项检测证书等。

在应对算法风险方面，算法黑箱是目前算法不可信任和不可解释的重要风险来源。在设计模型时，我们会将各种信息在代码中输出，当算法决策出错时，可以实现快速追溯原因。

同时，我们通过建立模型体检平台，对模型进行推理攻击和逆向攻击测试，可以检出算法模型对数字世界白盒对抗、数字世界黑盒查询对抗鲁棒准确率、数字世界迁移攻击对抗鲁棒准确率、物理世界对抗样本攻击成功率、模型后门攻击成功率等测试因子进行检测并评分，判定算法模型是否符合设计要求。在算法模型不符合设计要求的情况下，我们会启用相应的算法修复模块进行安全性的提升；同时，我们在系统层面建立 AI 防火墙，用于抵御对抗样本的攻击。在模型发布时，会在产品定义的测试集上做测试，并进行人工测试，保证测试集规模足够大，且在各类上的精度都可以达到要求的水平。

此外，我们基于真实场景下的数据集的算法验证与评测，开发了

“算法测评工具”：通过对主要场景和长尾场景的算法测评的全覆盖、计多元化的测评项目和丰富的指标体系，以及足够充分的数据集和全方面的评测方案对算法进行充分的评测，实现了所有商业化算法的可信与可控。

在应对应用管理风险方面，我们结合产品伦理风险审查的不同阶段设计了一整套伦理风险自查工具，并开发了产品伦理风险审查平台。目前，所有商汤人工智能产品在立项、发布，再到上线运营的不同阶段均需通过产品伦理风险审查平台进行伦理风险审查，并且，在进入审查程序之前，可以利用自查工具开展审查前的准备工作。在审查过程中，我们可选择驳回新产品方案、中断进行中的产品开发项目，或下线不符合我们原则及标准的现有产品。

凭借上述实践，我们的人工智能伦理治理体系和相关技术工具，多次获得第三方机构的积极评价，并先后入选哈佛商业评论及中国人工智能产业发展联盟“可信 AI”的优秀案例。我们发布的首份《AI 可持续发展白皮书》也获得联合国《人工智能战略资源指南》收录。

《哈佛商业评论》拉姆·查兰奖

图 8 商汤人工智能治理实践获得的外部认可

资料来源：商汤科技

（四）文化建设

我们深刻地认识到发展负责任且可评估人工智能的关键还在于在集团层面形成伦理治理的组织文化。

为了统一认识、强化执行，我们陆续发布了《商汤集团伦理治理制度》、《商汤集团人工智能伦理与治理委员会管理章程》、《商汤集团产品伦理风险评审指引》，为深化员工对伦理治理体系和实践的认识提供统一的指引和明确的标准。为加强员工对人工智能治理话题的关注和理解，我们建立了定期宣传和培训机制，每周都会向全体员工发送人工智能治理相关的重要动态，并会定期组织研讨会、邀请内外部专家开展伦理治理培训。

（五）生态建设

我们积极参与国家信安标委、IEEE 等人工智能伦理治理相关的标准组织，并在多个工作组担任组长或副组长。同时，我们也与清华大学、上海交通大学、新加坡“人工智能国际研究院”等国内外知名高校和专业研究机构建立起密切的伦理研究合作机制，并联合国内外的行业伙伴发起“亚洲科技促进可持续发展目标联盟”（Tech4SDG），稳步推动发展负责任且可评估的人工智能。目前，Tech4SDG 联盟涵盖亚洲地区 9 个国家和地区，机构成员超过 40 家，包括中国大陆、中国香港、中国澳门、新加坡、印度、沙特等地的多家知名高校和智库。

六、合乎伦理的产品设计实践

案例展示：“元萝卜”AI 下棋机器人

当手机、平板电脑越来越司空见惯，“让孩子放下电子设备”成了众多家长的需求和心声。2022 年 8 月 9 日，商汤正式推出了首个家庭消费级人工智能产品——“元萝卜 SenseRobot” AI 下棋机器人。“元萝卜 SenseRobot” AI 下棋机器人结合商汤领先的 AI 视觉技术和机械臂技术，将其浓缩到一个在家庭书桌上就能摆放的实体机器人产品中，让孩子从电子屏幕中跳脱出来，在学棋、下棋的同时，还能提升专注力，保护视力不伤眼。

图 9 “元萝卜 SenseRobot”使用场景展示

资料来源：商汤科技官网

“元萝卜 SenseRobot” AI 下棋机器人的外观简约、酷萌、科技感十足，以一个小小“宇航员”的形象呈现，如同家里来了一位新成员，与孩子“面对面”进行象棋教学和对弈。

商汤科技董事长兼 CEO 徐立在发布会中谈到：“我们希望通过创新和领先的人工智能技术，打造一个能够真正‘思考’和‘行动’的

机器人产品，让产业级 AI 技术走进千家万户，与孩子、长辈进行真实互动；不仅能陪伴儿童的成长，也让长辈轻松享受高科技，消除数字鸿沟，用科技搭建情感的桥梁，为全家人带来更多乐趣。”

元萝卜 SenseRobot”包含 AI 学棋、残局挑战、棋力闯关、巅峰对决等多种模式，可以从“0 基础”为孩子介绍和讲解象棋的文化、规则及每个棋子的使用技巧，在锻炼思维的同时，还能提升孩子的文化素养。此外，它还带来 100 多个残局设定和 26 个等级的棋力对战，让用户不仅可以体验“执子下棋”的真实感受，还能开动脑力享受高水平对弈的乐趣。

基于商汤原创的“AI 黑科技”，“元萝卜 SenseRobot” AI 下棋机器人可以做到“手眼协同”，实现毫米级的操作精准度，保证在下棋对弈过程中的运行顺畅和落子准确。不仅如此，它还“满腹经纶”，得到国家体育总局棋牌管理中心、中国象棋协会的权威认证和授权，为用户提供专业课程，并实现在家里足不出户就能完成 16-13 级的官方象棋考级评测，获得专业证书。

“好玩就要全家一起”这是“元萝卜 SenseRobot”首批测试用户的真实反馈。“元萝卜 SenseRobot” AI 下棋机器人通过 AI 深度学习和自我训练，棋力水平达到大师级，无论初学者还是已经具备一定象棋水平的玩家都能找到适合自己的对弈等级。不仅如此，它还能让全家人一起出谋划策，共同挑战，提升孩子与父母、长辈之间的互动感，成为全家人的情感纽带。

锻炼思维、全家陪伴，“元萝卜 SenseRobot” AI 下棋机器人通

过践行“以人为本”的伦理设计理念，为孩子的成长打开通向更高成就的阶梯。

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

孙大鹏　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

huzhengkun@sensetime.com

田丰

商汤科技

智能产业研究院院长

tianfeng@sensetime.com

鸣谢

梅莹、闫欣桐、李玥璐、刘志毅、宫超、吴晶彧、成瑾、龚柳婷、綦伟良、梁鼎、吴一超、王义飞、何聪辉、辛昱辰

更多信息，敬请关注商汤科技：

官网 https://www.sensetime.com/cn

领英 https://www.linkedin.com/company/sensetime-group-limited/

微信公众号