

《人工智能预训练模型 价值对齐技术框架》（送审稿）编制说明

一、项目背景

党的十八大以来，以习近平同志为核心的党中央高度重视我国新一代人工智能发展。习近平总书记深刻把握世界科技发展大势，深刻洞察人工智能的战略意义，指出：“人工智能是引领这一轮科技革命和产业变革的战略性技术，具有溢出带动性很强的‘头雁’效应”“加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题”。

随着人工智能预训练模型及相关系统的能力日益增强，它们被逐渐广泛应用于教育、医疗、金融等社会领域，赋能推进“人工智能全域全时场景应用”。然而，预训练模型在实际应用过程中仍存在诸多现实问题，如模型幻觉问题、模型谄媚现象、数据隐私泄露、欺骗性奖励破解等等。这些问题代表了人工智能预训练模型在应用过程中的“对齐失败现象”，在当前行业中广泛普遍存在，且均属于“人工智能价值观对齐”底层技术研究的一部分。人工智能价值对齐旨在使人工智能模型的行为与人类的意图和价值观相一致，从而使人工智能模型能够准确无误、真实可信地完成人类指令，进一步与真实世界进行对齐，从而实现技术突破和场景创新。可以说，是否完成了价值观对齐，是人工智能模型能否走向“全域全时场景应用”的关键决定性因素。

从国家维度看，在工业和信息化部 2024 年 1 月发布的《国家人工智能产业综合标准化体系建设指南》中明确提出要加快形成人工智能“关键技术标准”和“安全治理标准”，如大模型关键技术要求、人工智能可解释性技术要求等，本标准提案契合该指南中的多个重点工作领域。从深圳市维度看，在《深圳经济特区人工智能产业促进条例》中明确指出：“要加快开展战略性、前瞻性、系统性人工智能基础研究和关键核心技术攻关，推动学科理论与前沿技术的突破和创新，发挥创新引领和支撑作用。”此外，在《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023—2024 年）》中明确指出要“增强关键核心技术与产品创新能力，研发基于国际主流大模型的创新产品，积极拓展国际市场。”从全球维度看，人工智能对齐技术作为目前国际上诸如 OpenAI、Google DeepMind 以及 Anthropic 等头部企业积极布局的战略性、前瞻性、系统性的关键共性技术，已经成为国际产业界及学术界的研究关注重点和产品研发热点。“是否实现了更好的价值对齐”，目前已经是国外主流市场选择人工智能模型产品时的基础要求、共性要求，成为人工智能产品或服务的准入门槛。

综上，推进研制人工智能价值对齐相关技术标准不仅已成为国家以及我市的重点标准化工作任务，更是与国际人工智能前沿一线研究和产业发展动向接轨，实现我市人工智能基础研究和关键技术国际领先的重要步骤。

二、工作过程概述

（一）任务来源

深圳市市场监督管理局、深圳市工业和信息化局于 2024 年 1 月 4 日发布了《关于征集 2024 年度深圳市人工智能领域标准项目的通知》，深圳赛西信息技术有限公司根据人工智能产业发展亟需和技术、产品创新实际需求，提出了《人工智能预训练模型 价值对齐技术框架》深圳市地方标准制定计划项目建议书。2024 年 7 月 23 日，深圳市市场监督管理局下达《2024 年第二批深圳市地方标准计划项目任务通知》，《人工智能预训练模型 价值对齐技术框架》等 15 项深圳市地方标准予以立项。

（二）主要起草过程

（1）2024 年 1 月，深圳赛西信息技术有限公司提交《深圳市地方标准制修订计划项目建议书》，2024 年 7 月深圳市市场监督管理局批准立项。

（2）2024 年 8 月，深圳赛西信息技术有限公司组织相关企事业单位及科研院所专家，组成标准核心编制组，对标准框架进行多轮讨论，形成初步达成共识的标准目录框架结构。深圳赛西信息技术有限公司发布通知，公开面向社会广泛征集相关单位参与标准编制工作。

（3）2024 年 9 月—11 月，标准核心编制组合作进行标准框架及初稿内容的编制，产出标准初稿内容。

（4）2024 年 12 月—2025 年 3 月，深圳赛西信息技术有限公司牵头组织标准启动会，正式成立标准编制组。编制

过程中共组织 4 次标准编制组全体会议及 10 余次专项研讨会，组织各单位专家围绕人工智能预训练模型价值对齐技术框架展开深度研讨，过程中产出 5 稿标准草案，最终形成标准草案。

（5）2025 年 3 月—4 月，深圳赛西信息技术有限公司通过电子邮件方式自行征求意见，面向 4 家深圳市人工智能产业内代表性企业、高校及科研院所征求意见。共收到反馈意见 14 条，其中采纳 11 条，不采纳 3 条。

（6）2025 年 4 月至 5 月，标准编制组根据收到的反馈意见进行修改，形成新的征求意见稿。

三、标准主要内容的依据以及与国内领先、国际先进标准的对标情况

（一）标准主要依据

本文件核心聚焦人工智能领域内“价值对齐”的前沿技术，制定的主要依据为《国家人工智能产业综合标准化体系建设指南（2024 版）》中（三）关键技术标准及（七）安全/治理标准，参考 3 项已发布的通用大模型国家标准 GB/T 45288.1-2025《人工智能 大模型 第 1 部分：通用要求》GB/T 45288.2-2025《人工智能 大模型 第 2 部分：评测指标与方法》GB/T 41867-2022《信息技术 人工智能 术语》GB/T 45081-2024《人工智能管理体系》，编写相关的技术框架及标准内容。

（二）与国内领先、国际先进标准的对标情况

国家标准层面，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/TC28/SC42)当前已发布《人工智能 大模型 第1部分：通用要求》《人工智能 大模型 第2部分：评测指标与方法》《信息技术 人工智能 术语》等大模型领域国家标准，从基础共性维度对人工智能大模型相关定义与指标进行标准化明确；在人工智能安全治理方面，全国网络安全标准化技术委员会已发布 TC260-003《生成式人工智能服务安全基本要求》，明确了人工智能大模型安全相关的标准要求。总体来看，当前已发布的国家标准主要集中在人工智能相关的术语定义及通用要求方面，在人工智能预训练模型价值对齐的前沿技术领域尚处于空白状态，企业在实际开展价值对齐工作的过程中仍缺乏标准技术指导。

行业标准方面，当前工业和信息化部立项了《人工智能关键技术 语言大模型对齐能力评估》，该项标准目前处于启动阶段，主要从对齐能力评估的维度进行标准化明确。当前在行业标准领域，尚未对人工智能价值对齐的技术框架进行标准化明确，产业内需求仍较大。

地方标准方面，目前暂无价值对齐技术框架相关的地方标准，该领域属于前沿亟需。

四、主要条款说明以及主要技术指标、参数、试验验证

(一) 编制原则

由于人工智能领域技术仍处于快速迭代发展的过程中，编制组在编制标准的过程中以“充分适用”为基本编制思路，结合当前人工智能大模型在各行各业应用推广过程中的实际问题及解决方案，兼顾国际上人工智能价值对齐技术领域

的前沿技术趋势和共识，进行标准的研究与编制工作。具体遵循以下 2 个原则：

1. 前沿技术前瞻性：本标准在编制过程中始终紧跟人工智能预训练模型价值对齐技术领域前沿趋势及动向，灵活整合前沿技术，确保价值对齐技术框架符合前沿技术发展趋势。

2. 行业适用性：本标准在编制过程中重点关注人工智能预训练模型全生命周期内各利益相关方，涵盖人工智能模型应用的重点场景，确保本标准所产出的技术框架对各行各业开展人工智能价值对齐技术研发应用具有参考价值。

（二）主要条款说明、技术指标参数

《人工智能预训练模型 价值对齐技术框架》针对当前人工智能产业内前沿技术趋势及实际应用情况，结合深圳市人工智能产业需求，提出可供产业界、学术界参考应用的技术框架。本文件主要包括：

1. 范围：

本文件确立了对人工智能预训练模型进行价值对齐的技术参考架构和相关方活动。

本文件适用于人工智能价值对齐技术的研究、开发、应用、治理。

2. 规范性引用文件：对本文件规范性引用进行阐述说明。本文件的规范性引用文件包含 GB/T 45081-2024 人工智能管理体系。

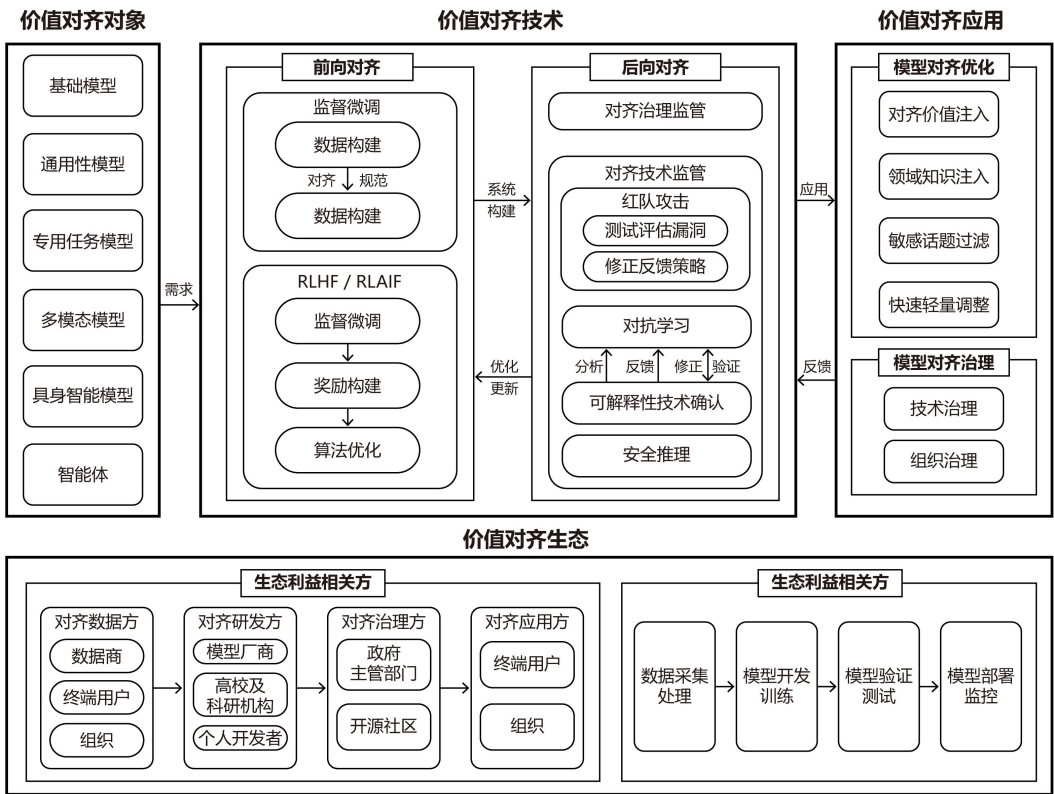
3. 术语和定义：

对人工智能系统、预训练模型、价值对齐、前向对齐、后向对齐、提示词、微调、训练数据、公平性、可信赖、伦理、鲁棒性、透明性等进行了标准化界定和说明。

4. 价值对齐概念: 价值对齐贯穿人工智能预训练模型及系统的生命周期，旨在实现技术目标的同时，在伦理与安全之间保持平衡，以确保技术实现与人类价值观的有效协调。

5. 价值对齐准则: 价值对齐的基础准则是人工智能系统在技术实施与伦理保障中的指导原则，包含包容性、隐私性、鲁棒性、可解释性、可控性等。

6. 价值对齐技术框架: 本文件明确了人工智能预训练模型价值对齐涉及的利益相关方、技术活动及对象的技术框架，技术活动及对象包括：价值对齐对象、价值对齐技术、价值对齐应用和价值对齐生态等四类。



（1）价值对齐对象：包括基础模型、通用型模型、专用任务模型、多模态模型、具身智能模型、智能体等。

（2）价值对齐技术：包含前向对齐和后向对齐两部分，给出了在价值对齐过程中应该考虑的技术框架性构成。

（3）价值对齐应用：包括模型对齐优化及模型对齐治理，通过技术迭代与系统性管理结合的方式，确保人工智能系统在全生命周期中遵循人类伦理准则、法律法规及社会价值规范。

（4）价值对齐生态：包括数据采集处理、模型开发训练、模型验证测试和模型部署监控等活动。

（三）主要试验情况分析

当前，在人工智能产业界内已形成对人工智能预训练模型价值对齐的技术基础，前向对齐、后向对齐等技术已经较为成熟。目前参与编写的单位中，已开展了对人工智能预训练模型价值对齐的落地应用，并形成了开源 AI 对齐工具，具备研制相关标准的基础。本标准相关内容经过与各重点企业的对接和沟通，已被证明确实可行。

五、知识产权情况说明

本文件不涉及专利及知识产权问题。

六、重大意见分歧的处理依据和结果

无。

七、实施标准的措施建议

为更好地发挥标准实施效力，建议标准发布后，开展《人工智能预训练模型价值对齐技术框架》标准宣贯会，提升标

准的宣传权威性和受众针对性。利用多种渠道、多种方式加强本文件的宣贯，并对本文件的执行情况进行跟踪调查，对标准实施效果进行评估，及时发现并解决标准实施过程中存在的问题，适时开展修订完善工作，提升本文件的科学性和适用性。