

DBXXX

深圳市地方标准

DBXXX/T XXXX—2024

人工智能预训练模型 价值对齐技术框架

Artificial Intelligence Pre-trained Models - AI Alignment: Technical Framework

(征求意见稿)

2024 - XX - XX 发布

2024 - XX - XX 实施

深圳市市场监督管理局 发布

目 次

前 言 II

1 范围 3

2 规范性引用文件 3

3 术语和定义 3

4 价值对齐概述 6

 4.1 概念构成 6

 4.2 基础准则 6

 4.3 价值对齐目标 8

5 价值对齐技术框架 8

6 价值对齐对象 10

 6.1 概述 10

 6.2 基础模型 11

 6.3 通用性模型 12

 6.4 专用任务模型 12

 6.5 多模态模型 13

 6.6 具身智能模型 14

 6.7 智能体 14

7 价值对齐技术 15

 7.1 概述 15

 7.2 活动输入 15

 7.3 活动输出 16

 7.4 前向对齐 17

 7.5 后向对齐 19

8 价值对齐应用 21

 8.1 模型对齐优化 21

 8.2 模型对齐治理 22

9 价值对齐生态 23

 9.1 生态主要活动 23

参考文献 24

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由深圳市工业和信息化局提出并归口。

本文件起草单位：深圳赛西信息技术有限公司、北京大学人工智能研究院、清华大学深圳国际研究生院、鹏城实验室、华为技术有限公司、平安科技（深圳）有限公司、香港中文大学（深圳）、深圳市星创数字经济研究中心、粤港澳大湾区数字经济研究院、深圳市标准技术研究院、北京智源人工智能研究院、北京谋远咨询有限公司、深圳市中电电力技术股份有限公司、深圳优优互网络科技有限公司。

本文件主要起草人：谭瑞琥、高万琪、杨耀东、李平、雷雪晶、刘方明、吉嘉铭、陈博远、陈琢、王畅、唐思洁、莫凡、瞿晓阳、张楠、赵展展、吴保元、崔丽坤、徐巍、李悦、李婷、周宏、王瑞、孙千国、王超群、戴俊韬、方亮、段雅文、谢旻希、刘宁、王鹏、符亚雄、江伟伟。

人工智能预训练模型 价值对齐技术框架

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

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 45081-2024 人工智能管理体系

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能系统 artificial intelligence system

针对人类定义的给定目标,产生诸如内容、预测、推荐或决策等输出的一类工程系统。

[来源: GB/T 41867-2022]

注1: 该工程系统使用人工智能相关的多种技术和方法,开发表征数据、知识、过程等的模型,用于执行任务。

注2: 人工智能系统具备不同的自动化级别。

3.2

预训练模型 pre-trained model

基于大量数据训练得到,具有复杂计算架构,能够处理复杂任务,且具备一定泛化性的深度学习模型。

[来源: GB/T 45288.1-2025, 有修改]

注: 预训练模型的参数量由其功能和模态决定,一般不低于1亿。大模型训练使用的数据总量受参数数量的影响,达到收敛的大模型的参数量的对数与其训练数据总量的对数成正比。

3.3

价值对齐 value alignment

确保人工智能系统的目标设定、决策逻辑及输出结果与人类的意图和价值观一致，从而准确无误、真实可信地完成人类指令。

3.4

前向对齐 forward alignment

在预训练模型的训练阶段（含预训练阶段及后训练阶段），通过训练过程使模型初步满足价值对齐要求及目标。

3.5

后向对齐 backward alignment

在预训练模型的训练后阶段及部署阶段，通过持续验证、测试及调整的方式确保模型实际表现符合价值对齐要求及目标。

3.6

提示词 prompt

提示语

使用大模型进行微调或下游任务处理时，插入到输入样本中的指令或信息对象。

[来源：GB/T 45288.1-2025]

3.7

微调 fine-tuning

为提升机器学习模型预测准确性，使专门领域数据在大模型上继续训练的过程。

[来源：GB/T 45288.1-2025]

注1：专门领域数据一般是特定场景的生产数据或合成数据。

注2：常用的微调方法包括提示词微调、全参微调、参数高效微调等。

3.8

训练数据 training data

用于训练机器学习模型的输入数据样本子集。

[来源：GB/T 41867-2022]

3.9

公平性 fairness

尊重既定事实、社会规范和信仰，且不受偏袒或不公正歧视影响的对待、行为或结果。

[来源：GB/T 41867-2022]

注1：对公平性的考虑是与环境高度相关的，并且因文化、代际、地理和政治观点而异。

注2：公平不等于没有偏见。偏见并不总是导致不公平，不公平可能是由偏见以外的因素引起的。

3. 10

可信赖 trustworthiness

人工智能满足利益相关方期望并可验证的能力。

[来源：GB/T 41867-2022]

注1：依赖于语境或行业，也依赖于具体的产品或服务、数据以及所用技术，应用不同的可信赖特征并对其进行验证，以确保利益相关方的期望能得到满足。

注2：可信赖的特征包括可靠性、韧性、安全性（信息安全、功能安全）、隐私性、可问责、透明性、真实性、质量、实用性等。

注3：可信赖作为一种属性用于描述服务、产品、技术、数据和信息，在治理中也用于组织。

3. 11

伦理 ethics

开展人工智能技术基础研究和应用实践时遵循的道德规范或准则。

[来源：GB/T 41867-2022]

3. 12

鲁棒性 robustness

人工智能系统在任何情况下都保持其性能水平的特性。

[来源：GB/T 41867-2022]

3. 13

透明性 transparency

人工智能系统与利益相关方交流关于该系统适当信息的特性。

[来源：GB/T 41867-2022]

注1：系统透明性相关的信息一般包含特性、性能、缺陷、组件、程序、度量、设计目标、设计选择和假设、数据来源及标注协议。

注2：对系统某些方面不适当的保录一般会违背安全、隐私或保密要求。

3. 14

缩略语

下列缩略语适用于本文件

AI：人工智能（Artificial Intelligence）

RLHF：基于人类反馈的强化学习（Reinforcement Learning from Human Feedback）

PPO：近端策略优化（Proximal Policy Optimization）

RLHF：基于人类反馈的强化学习
RLAIF：基于人工智能反馈的强化学习

4 价值对齐概述

4.1 概念构成

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。

价值对齐的主要任务是通过规范性对齐和技术性对齐两个维度，确保人工智能系统的目标设定、决策逻辑及输出结果与人类的意图和价值观一致，并在技术和伦理层面上达到预期目标。

- a) 规范性对齐：确定人工智能系统应纳入的伦理价值观和社会规范，确保系统能够在复杂社会情境中具备伦理包容性和区域适应能力。
- b) 技术性对齐：确定人工智能系统技术实现的路径，使系统按照预期价值与规范执行，确保系统稳定、可靠地执行预定目标，并避免对抗性行为或偏离意图的输出。

4.2 基础准则

4.2.1 概述

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核心指导原则，旨在为人工智能技术的研究、开发、应用和治理提供具体指引。价值对齐准则的分类如下。

- a) 规范性准则：关注伦理价值观和社会规范的对齐，包括：
 - 1) 包容性（见4.2.2）；
 - 2) 隐私性（见4.2.3）。
- b) 技术性准则：关注技术层面的稳定、透明和可调整，包括：
 - 1) 鲁棒性（见4.2.4）；
 - 2) 可解释性（见4.2.5）；
 - 3) 可控性（见4.2.6）。

4.2.2 包容性

人工智能预训练模型及系统在全生命周期内应避免对群体、地域的偏见或歧视，确保不同背景的用户均能获得公平的服务与体验，避免价值观的固化，具体包括但不限于以下措施：

- a) 宜确保训练数据的多样性，覆盖不同性别、年龄、地域和文化背景等特征，以降低因样本分布不均引发的偏见风险；
- b) 宜进行定期检测与调整，通过技术手段监测模型输出的潜在偏见，并根据分析结果优化模型行为；

- c) 宜结合边缘化群体的特殊需求，设计针对性优化方案，以提升模型的公平性和社会适应能力。

4.2.3 隐私性

人工智能预训练模型及系统在设计、开发与应用过程中，应采取措施保护用户数据的安全性和机密性，维护用户的基本权利，具体包括但不限于以下措施：

- a) 应确保数据收集、处理和存储过程中遵循最小必要原则；
- b) 应采用加密、匿名化等技术手段，保障用户个人信息的安全；
- c) 应根据相关法律和伦理要求，确保用户在数据使用中的知情同意权，并提供相应的访问、修改和删除权利；
- d) 宜定期审查和更新隐私保护措施，确保其始终符合数据保护法律以及社会期望。

4.2.4 鲁棒性

人工智能预训练模型及系统在面对极端环境、异常输入或对抗攻击时，应能够保持稳定运行并实现预期目标，具体包括但不限于以下措施：

- a) 应具备异常输入的检测能力，识别可能导致系统故障的异常数据或对抗性攻击，主动触发响应机制；
- b) 应通过对抗性训练和压力测试，增强模型应对不可预测场景的弹性与适应性；
- c) 应确保在多样化场景中的稳定性能，避免因数据类型或输入复杂性导致模型行为的不一致。

4.2.5 可控性

人工智能预训练模型及系统的行为或决策应始终处于人类监督和控制之下，确保在系统行为偏离预期时能够迅速有效调整或终止，具体包括但不限于以下措施：

- a) 应加强事前风险预警和评测体系，在模型部署前进行充分的安全测试，包括对抗性测试、压力测试等，明确人类监督权限与干预阈值，预设应急响应流程；
- b) 应支持实时监控与异常警报功能，持续追踪模型行为，对异常状况发出警示，并提供动态反馈接口，使利益相关者可快速识别并响应潜在风险；
- c) 应在紧急情况下提供可靠的人工干预手段，如暂停系统运行、调整模型行为或终止任务执行，同时记录异常事件及干预措施，形成闭环反馈，以优化后续的控制策略。

4.2.6 可解释性

人工智能预训练模型及系统的生成内容或决策过程对于用户和利益相关者应透明且易于理解，具体包括但不限于以下措施：

- a) 宜记录算法核心逻辑、模型行为和数据输入的详细信息，形成可供内部审查或外部监管使用的技术档案；
- b) 宜提供支持决策路径解析的工具，帮助用户理解模型生成结果的关键驱动因素，避免系统的不诚实行为或欺骗性对齐；

- c) 宜提供用户友好的反馈机制，帮助不同背景的用户理解决策依据和潜在局限性。

4.3 价值对齐目标

价值对齐的具体目标简述如下。

- a) 保障人类安全利益：确保模型及系统在多样化环境中安全稳定运行，避免对人类造成潜在威胁，保护用户的合法权益，并关注模型表现出的危险倾向性，如偏见、欺骗、虚构信息等；
- b) 适配多元伦理价值：确保模型及系统行为符合广泛认可的伦理和社会规范，如公平、包容、尊重隐私等，能够适应不同文化背景和社会环境的需求；
- c) 行为一致与可靠：确保系统的行为与预期目标一致，避免生成有害或偏离意图的输出，以及系统在执行目标时保持可靠性；
- d) 明确行为逻辑与责任：提供透明可信的模型行为和决策解释，使预训练模型及系统开发方对模型及系统的行动、决定和行为负责，避免出现欺骗、权力寻求等倾向；
- e) 适应动态环境及需求变化：通过反馈机制优化模型性能，使其具备适应环境和需求动态变化的调整能力，满足社会和技术发展的长期要求。

价值对齐目标与基础准则、技术路径、应用治理之间的关系见表 1。

价值对齐目标	基础准则	技术路径	应用治理
保障人类安全利益	鲁棒性、可控性	对抗学习、红队攻击、安全推理等	通过对齐价值注入和敏感话题过滤，确保模型能够正确应对潜在的风险；通过快速轻量化调整修正偏离预期的行为
适配多元伦理价值	包容性、隐私性	监督微调、基于人类反馈的强化学习、人类价值验证等	引入领域知识确保模型遵循特定领域的伦理规范；通过敏感话题过滤避免不当输出
行为一致与可靠	可控性、可解释性	基于人工智能反馈的强化学习、机制可解释性等	结合对齐策略文档，动态跟踪并纠正模型偏差，在高风险任务中强化可解释性分析
明确行为逻辑与责任	可解释性	机制可解释性、安全推理等	强调跨部门协作与协调，确保各团队在价值对齐方面协调一致；同时进行伦理审查，确保责任明确
适应动态环境及需求变化	包容性、可控性	基于人工智能反馈的强化学习、安全推理等	通过输入/输出过滤与合规性监控，确保模型输出符合伦理和法律标准，且通过持续优化适应变化的环境

表 1 价值对齐目标

5 价值对齐技术框架

图1示出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。

- a) 价值对齐对象：包括大规模预训练的基础模型、经过后训练方法微调后的通用性模型、使用特定领域数据微调并针对垂类场景优化的专用任务模型、协同处理多种模态的多模态模型、与环境进行物理或虚拟交互的具身智能模型、由多模型协同构成的系统级实体自主智能体等。
- b) 价值对齐技术：由前向对齐和后向对齐两部分组成。
 - 1) 前向对齐：针对人工智能预训练模型的数据采集处理、模型开发训练阶段，通过技术手段使预训练模型满足价值对齐的要求及目标。
 - 2) 后向对齐：针对人工智能预训练模型的模型验证测试、模型部署监控阶段，通过持续验证、测试及调整的方式确保模型实际表现符合价值对齐的要求及目标。
 - 3) 价值对齐技术闭环：后向对齐将收集并反馈人工智能模型及系统的不对齐行为，驱动前向对齐中数据和训练方法的迭代调整，形成价值对齐技术闭环。
- c) 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相结合的方式，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。通过动态调整机制降低价值偏差风险，提升模型在复杂场景下的决策可解释性与可控性。
- d) 价值对齐生态：价值对齐主要生态活动包括数据采集处理、模型开发训练、模型验证测试和模型部署监控等活动。主要由以下四类生态利益相关方参与执行：
 - 4) 对齐数据方：由数据商、终端用户、组织等构成，主要执行价值对齐数据的构建和提供活动；
 - 5) 对齐研发方：由模型厂商、高校及科研机构、个人开发者等组成，主要执行基于价值对齐的技术、工具或服务的开发和集成活动；
 - 6) 对齐治理方：由政府主管部门和开源社区等组成，主要执行价值对齐技术及模型的应用监管活动；
 - 7) 对齐应用方：由终端用户、组织等构成，主要执行价值对齐技术及产生的应用服务的使用活动。

每类参与者有其主要执行的活动，同时可能执行涉及四类活动中的多项活动。

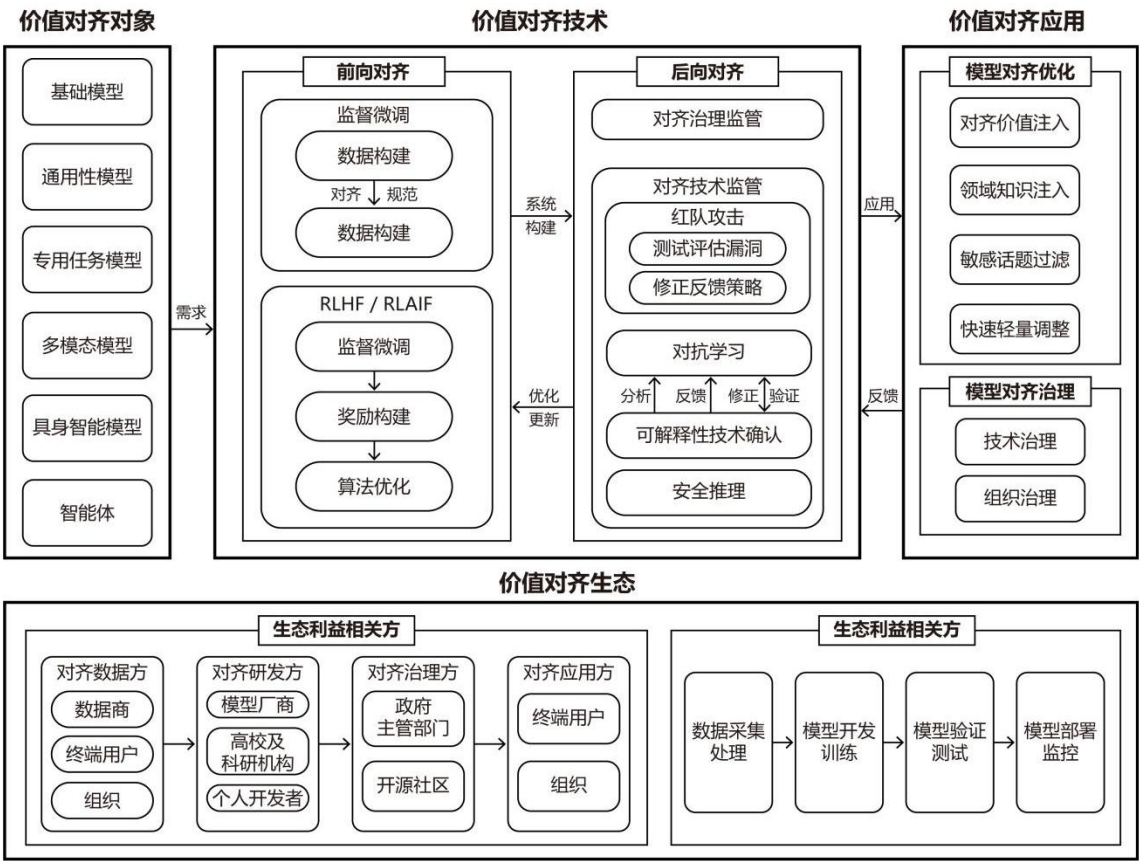


图 1 人工智能预训练模型价值对齐技术框架

6 价值对齐对象

6.1 概述

价值对齐的对象包括但不限于：

- a) 基础模型：通过自监督学习在超大规模的一类或多类源数据（文本、图像、视频等）上进行大规模预训练，未经过任何任务特定微调的通用模型；
- b) 通用性模型：基于基础模型，通过监督微调、基于人类反馈的强化学习等后训练方法，使模型具备理解开放指令、遵循社会规范及伦理规则的通用交互能力。
- c) 专用任务模型：基于基础模型或通用性模型，使用特定领域数据（如医疗、法律、金融等）进行监督微调，针对垂类场景优化的高精度模型；
- d) 多模态模型：能够处理或生成两种及以上的模态数据（文本、图像、音频、视频、3D空间等），其架构可为单模型融合或多模型协同。
- e) 具身智能模型：与环境进行物理或虚拟交互的模型，通过传感器-执行器闭环实现具身推理与行动决策。
- f) 智能体：由一类有能力和人类用户、其他智能体及复杂环境进行交互的智能化程序系统级实

体，具备以下四要素：

- 1) 自主行动: 无需人类介入即可影响环境；
- 2) 抽象目标遵循: 基于高层目标自主拆解子任务；
- 3) 长时段规划: 跨时间步的复杂推理能力；
- 4) 最小人类监督: 自主运行与自修正机制。

价值对齐对象与价值对齐目标的对应关系见表 2。

<div>对象</div> <div>目标</div>	基础模型	通用性模型	专用任务模型	多模态模型	具身智能模型	智能体
保障人类安全利益	社会影响对齐	公平性对齐、可解释性对齐	合规性对齐、安全性对齐、可解释性对齐	伦理敏感性对齐	人机协作性对齐、伦理合规性对齐、安全性对齐	交互对齐、安全性对齐、社会技术系统对齐、群体协同对齐
适配多元伦理价值	训练数据对齐	训练数据对齐、公平性对齐	领域数据对齐、公平性对齐	伦理敏感性对齐	伦理合规性对齐、安全性对齐	主动决策对齐、群体协同对齐、社会技术系统对齐
行为一致与可靠	生成能力对齐、任务领域对齐	指令跟随能力对齐、生成能力对齐、适应性对齐	目标任务对齐、领域数据对齐	跨模态一致性对齐、情景适应性对齐、用户体验对齐	感知对齐、行动对齐、决策对齐、适应性对齐	目标对齐、交互对齐、主动决策对齐
明确行为逻辑与责任	隐私保护对齐	公平性对齐	合规性对齐、安全性对齐、公平性对齐	伦理敏感性对齐	伦理合规性对齐、安全性对齐	社会技术系统对齐、安全性对齐
适应动态环境及需求变化	环境可持续性对齐	可解释性对齐	可解释性对齐、性能稳定性对齐	伦理敏感性对齐	适应性对齐	社会技术系统对齐

表2 价值对齐对象及目标关系表

6.2 基础模型

基础模型指通过自监督学习在超大规模一类或多类源数据（文本、图像、视频等）进行大规模预训练，未经过任何任务特定微调的通用任务模型，一般具备通用领域内的语言生成能力。对基础模型的价值对齐包括但不限于以下方向：

- a) 训练数据对齐：确保训练数据在来源、内容和分布上的公平性，避免模型因数据偏差导致输出结果存在偏见。数据宜涵盖多样化的文化、语言和社会背景，以确保模型的普适性和包容性；
- b) 生成能力对齐：确保模型在生成内容时，避免输出不符合价值对齐基础准则的内容，例如不实信息、有害内容或潜在歧视性表述。

- c) 任务领域对齐：确保模型在多领域、多任务应用中保持对齐性能，避免单一任务表现优异而其他任务失衡。模型应具备跨领域的泛化能力，确保在不同应用场景下均能保持一致的性能和可靠性；
- d) 隐私保护对齐：确保模型在处理和生成数据时，严格遵守隐私保护原则，避免泄露用户敏感信息。模型应具备数据匿名化和加密能力，确保用户隐私不被侵犯；
- e) 环境可持续性对齐：确保模型的训练和部署过程符合环境可持续性原则，减少能源消耗和碳排放。模型宜优化计算资源的使用，推动绿色AI的发展；
- f) 社会影响对齐：确保模型的应用对社会产生积极影响，避免加剧社会不平等或引发其他负面社会效应。模型应促进社会公平、包容和可持续发展。

6.3 通用性模型

通用性模型基于基础模型，通过在后训练阶段使用监督微调、基于人类反馈的强化学习等方法对模型参数进行调整，使模型具备理解开放指令、遵循社会规范及伦理规则的通用交互能力。对通用性模型的价值对齐包括但不限于以下方向：

- a) 训练数据对齐：确保用于后训练的数据来源合法、内容合规且分布均衡，避免因数据偏差导致模型输出存在偏见或歧视。数据宜涵盖多样化的文化、语言和社会背景，以提升模型的普适性和公平性；
- b) 指令跟随能力对齐：确保模型能够准确理解并执行开放指令，同时遵循社会规范和伦理规则，避免输出违反公序良俗或法律法规的内容。模型应具备对指令意图的精准解析和合规性判断能力；
- c) 生成能力对齐：确保模型在生成内容时避免输出不实信息、有害内容或潜在歧视性表述，生成结果应符合社会道德和法律规范；
- d) 适应性对齐：确保模型在不同输入、不同上下文场景下表现稳定，适应性强，减少因场景变化导致的性能波动。模型宜具备上下文感知能力，灵活应对多样化的用户需求；
- e) 公平性对齐：确保模型在任务执行过程中避免对特定群体或个体产生偏见或歧视，输出结果应体现公平性和包容性。模型应具备对潜在偏见的检测和纠正能力；
- f) 可解释性对齐：确保模型的决策过程和输出结果具备一定程度的可解释性，帮助用户理解模型的推理逻辑，增强用户信任。模型宜提供透明且易于理解的输出解释。

6.4 专用任务模型

专用任务模型：基于基础模型或通用性模型，使用特定领域数据（如医疗、法律、金融）进行监督微调，针对垂直场景优化的高精度模型。对专用任务模型的价值对齐包括但不限于以下方向：

- a) 目标任务对齐：确保模型在特定任务上高效适配，准确理解并执行目标任务需求，输出符合预期的结果。模型宜避免过度拟合特定任务而丧失泛化能力，同时确保任务执行的准确性和可靠性；

- b) 领域数据对齐：确保用于微调的数据符合特定领域特点，避免领域外数据引入造成结果偏差或失准。数据宜覆盖领域内的多样性和复杂性，确保模型在垂直场景中的表现具有代表性和权威性；
- c) 适应性对齐：确保模型在不同输入、不同上下文场景下表现稳定，适应性强，减少因场景变化导致的性能波动。模型宜具备对领域内多样化需求的灵活响应能力，确保在实际应用中的鲁棒性；
- d) 合规性对齐：确保模型在特定领域任务中遵循行业规范、法律法规和伦理要求，避免输出不符合领域标准或可能引发法律风险的内容。模型应具备对领域规则的识别和遵守能力；
- e) 安全性对齐：确保模型在处理敏感领域数据（如医疗健康、金融信息）时，严格遵守数据安全和隐私保护原则，避免泄露用户敏感信息。模型应具备数据加密和匿名化处理能力；
- f) 可解释性对齐：确保模型的决策过程和输出结果具备高度的可解释性，特别是在高风险领域（如医疗诊断、法律判决）中，帮助用户理解模型的推理逻辑，增强透明度和信任度；
- g) 公平性对齐：确保模型在垂直场景中避免对特定群体或个体产生偏见或歧视，输出结果应体现公平性和包容性。模型应具备对潜在偏见的检测和纠正能力，特别是在涉及社会敏感问题的领域；
- h) 性能稳定性对齐：确保模型在长时间运行和高负载场景下保持稳定的性能，避免因计算资源限制或数据规模变化导致输出质量下降。模型宜具备资源优化和动态调整能力。

6.5 多模态模型

多模态模型能够处理或生成两种及以上的模态数据（文本、图像、音频、视频、3D 空间等），其架构可为单模型融合或多模型协同。对多模态模型的价值对齐包括但不限于以下方向：

- a) 跨模态一致性对齐：确保模型在处理不同类型的输入时保持输出的一致性，避免因模态差异而产生误导性解释或偏见。例如，当文本描述与图像内容存在冲突时，模型应能够识别不一致并作出合理调整，确保多模态信息的协同性和逻辑一致性；
- b) 信息传递准确性对齐：保证信息在不同模态之间转换时的准确性和完整性，避免信息丢失或扭曲。模型宜具备跨模态信息的精准对齐能力，确保在模态转换过程中关键信息得以保留和正确传递；
- c) 情境适应性对齐：增强模型根据具体应用场景调整其行为的能力，确保它能够灵活应对各种复杂的现实情况；
- d) 伦理敏感性对齐：确保模型生成的内容尊重文化差异和社会伦理规范，无论是在何种模态下，都避免产生有害、歧视性或不适当的信息。模型应具备对多模态内容的伦理合规性检测能力，确保输出符合社会道德和法律要求；
- e) 用户体验对齐：确保模型在多模态交互场景中提供流畅、自然的用户体验，避免因模态切换或信息处理不当导致用户困惑或不满。模型宜具备对用户意图的精准理解能力，并输出符合用户期望的多模态内容。

6.6 具身智能模型

具身智能模型指通过与物理环境交互，实现感知、决策与行动能力的智能系统，通常结合传感器、执行器和决策算法，以适应复杂动态环境。对具身智能模型的价值对齐包括但不限于以下方向：

- a) 感知对齐：确保模型对物理环境中的信息感知准确无误，包括视觉、听觉、触觉等多模态数据的实时处理与整合，避免因感知偏差影响后续决策。模型应具备对动态环境变化的快速响应能力，确保感知数据的实时性和可靠性；
- b) 决策对齐：确保模型在动态环境中作出的决策合理高效，符合预设价值对齐目标与约束条件，避免出现不稳定或不安全的行为。模型应具备对复杂场景的推理能力，确保决策过程透明且符合伦理规范；
- c) 行动对齐：确保模型在执行任务时具备精确的行动能力，动作稳定、灵活且具备高效能，避免因执行误差导致任务失败或资源浪费。模型应具备对执行结果的实时反馈和调整能力，以优化行动效果；
- d) 安全性对齐：确保模型在与物理环境交互过程中严格遵守安全规范，避免对自身、用户或环境造成伤害。模型应具备对潜在风险的识别和规避能力，确保行动的安全性；
- e) 适应性对齐：确保模型能够适应复杂动态环境的变化，具备对未知场景的学习和调整能力。模型应能够在不同环境条件下保持稳定的性能，避免因环境变化导致任务失败；
- f) 伦理合规性对齐：确保模型在决策和行动过程中遵循社会伦理和法律法规，避免产生不符合道德规范的行为。模型应具备对伦理规则的识别和遵守能力，确保其行为符合社会期望；
- g) 人机协作性对齐：确保模型在与人类协作时能够理解人类意图并作出合理响应，避免因沟通不畅或行为冲突导致协作失败。模型应具备对人类行为的预测和适应能力，确保协作过程的流畅性和高效性。

6.7 智能体

智能体是一类有能力和人类用户、其他智能体及复杂环境进行交互的智能化程序系统级实体，具备自主行动、抽象目标遵循、长时段规划和最小人类监督四要素，具备自主采取行动、实现特定目标，并通过学习训练不断优化性能的能力[10]。对智能体系统的价值对齐包括但不限于以下方向：

- a) 目标对齐：确保智能体的行为与用户意图或系统设定目标保持一致，避免出现目标偏离或未完成目标的情况。智能体应具备对高层目标的拆解和执行能力，确保任务完成的准确性和效率；
- b) 交互对齐：确保智能体在与人类用户、其他智能体及环境交互时表现自然、准确、流畅，避免交互中的信息误解、冲突或不适配情况。智能体应具备对交互意图的精准理解能力，确保沟通和协作的高效性；
- c) 主动决策对齐：确保智能体在自主行动中具备动态环境适应性与多目标协同优化能力，通过实时感知环境状态、权衡效率资源风险等多维度要素，实现复杂场景下的可靠推理与安全决策。智能体宜建立可解释的决策机制，同步预测并评估行动对物理环境、社会系统的潜在影

响，避免引发不可逆的负面连锁效应；

- d) 群体协同对齐：确保智能体在多智能体协作过程中具备符合价值对齐目标的趋同价值观或共识，能够保持一致的目标导向和行为规范，避免因个体差异导致安全风险、冲突或资源浪费。智能体宜具备对群体行为的协调和优化能力，确保协作的高效性和稳定性；
- e) 社会技术系统对齐：确保智能体不仅在其内部运作和与直接交互的对象之间实现价值对齐，而且在更广泛的社会和技术生态系统中也能够保持一致。智能体的设计和部署宜考虑到社会结构、法律框架、经济体系和文化背景，确保其行为符合社会期望和规范；
- f) 安全性对齐：确保智能体在自主运行过程中严格遵守安全规范，避免对自身、用户或环境造成伤害。智能体应具备对潜在风险的识别和规避能力，确保行动的安全性和可靠性。

7 价值对齐技术

7.1 概述

价值对齐技术主要包括前向对齐、后向对齐[2]等活动。具体来说，价值对齐技术组成包括如下环节：

- a) 前向对齐主要是在训练阶段，通过数据准备和收集进行价值抽取、通过预训练、监督微调、基于人类反馈的强化学习等技术使得预训练模型初步满足价值对齐的要求及目标；
- b) 后向对齐主要是在模型完成训练后，通过持续验证、测试及调整的方式确保模型实际表现符合价值对齐的要求及目标；
- c) 系统在后向对齐中表现出来的不对齐行为，经过反馈收集、验证确认、需求更新，促进前向对齐部分的数据和方法调整，从而形成完整的对齐闭环。

7.2 活动输入

价值对齐技术的输入包括但不限于：

- a) 价值对齐对象：需要进行价值对齐的各类人工智能预训练模型及其基于此类模型的系统，包括但不限于基础模型、通用性模型、专用任务模型、多模态模型、具身智能模型和自主智能体等。这些对象在功能、应用场景和技术特性上存在差异，需要针对其特点制定相应的对齐策略；
- b) 价值对齐需求：具体的人类价值观、任务需求和社会群体需求，为对齐活动提供目标和约束条件。
 - 1) 人类价值观：如安全、公平性、隐私保护、透明性、伦理合规性等，确保模型的行为和输出符合社会道德和法律规范；
 - 2) 任务需求：针对特定任务场景（如医疗诊断、金融分析、自动驾驶等）的具体要求，确保模型在任务执行过程中高效、准确且可靠。修改为分号
 - 3) 社会群体需求：考虑不同文化、语言、社会背景和用户群体的多样性需求，确保模型具备包容性和普适性，避免对特定群体产生偏见或歧视。

- c) 法律法规与行业标准
 - 1) 法律法规：确保对齐技术符合国家和地区的法律法规要求，避免因法律风险导致模型无法部署或使用；
 - 2) 行业标准：参考相关行业的技术标准和最佳实践，确保对齐技术的规范性和可推广性。
- d) 用户反馈与迭代需求
 - 1) 用户反馈：通过用户使用模型的实际反馈，识别价值对齐中的不足和改进空间，为对齐技术的优化提供依据。修改为分号
 - 2) 迭代需求：根据技术发展和社会需求的变化，持续更新和调整价值对齐目标和方法，确保模型的长期适用性和可靠性。
- e) 技术约束与环境条件
 - 1) 技术约束：包括模型的计算资源、数据可用性、算法复杂度等，确保对齐技术的可行性和可操作性。修改为分号
 - 2) 环境条件：包括模型部署的物理环境、社会环境和技术生态，确保对齐技术能够适应复杂动态的环境变化。

7.3 活动输出

价值对齐技术的输出包括但不限于：

- a) 优化对象：
 - 1) 对齐后的模型：经过价值对齐优化的预训练模型或系统，具备符合人类价值观、任务需求和社会期望的行为和输出能力；
 - 2) 对齐策略文档：详细记录价值对齐的目标、方法、过程和结果，为后续模型迭代和优化提供参考。
- b) 技术应用：
 - 1) 对齐技术工具包：包括对齐算法、评估指标、测试用例等，为其他模型或系统的价值对齐提供技术支持；
 - 2) 部署指南：提供模型在特定场景中部署和使用的指导，确保对齐后的模型能够满足实际应用需求。
- c) 评估报告：
 - 1) 对齐效果评估：对模型的价值对齐效果进行量化评估，包括安全性、公平性、透明性等指标的测试结果；
 - 2) 风险分析报告：识别模型在部署和应用过程中可能存在的风险，并提供相应的缓解措施。
- d) 用户培训与支持：
 - 1) 用户使用文档：为用户提供模型使用和对齐目标理解的培训资源，确保用户能够正确使用对齐后的模型；
 - 2) 技术支持服务：为用户在使用过程中遇到的问题提供技术支持，确保模型的稳定运行和

持续优化。

7.4 前向对齐

7.4.1 监督微调

在预训练语言模型基础上，通过特定任务或人类标注数据进行训练，通过提供大量示例，引导模型学习生成符合人类期望的输出，从而实现价值对齐目标。监督微调技术[4]包括但不限于以下关键点：

a) 数据准备及价值抽取：

- 1) 数据构建应遵循的规范：数据构建时需遵循的人类价值意图，确保数据涵盖多样化的文化、语言和社会背景，符合社会道德、法律规范和人类伦理，避免输出不实信息、有害内容或违反公序良俗；
- 2) 数据构建：通过收集、整理和清洗一组高质量的人类标注数据，构建输入-输出对的数据集，数据集应具备人类指令、响应及标注，符合价值对齐原则及人类偏好；

b) 目标函数构建：采用交叉熵损失函数作为优化目标，衡量模型预测的概率分布与真实标签之间的差异。交叉熵损失函数计算公式见式（1）。

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T_i} \log P_{\theta}(y_{i,t} | x_i, y_{i,<t}) \quad (1)$$

式中：

x_i ——第 i 个输入

$y_{i,t}$ ——第 i 个样本在时间步 t 的输出

T_i ——第 i 个样本的长度

θ ——模型参数

7.4.2 基于人类反馈的强化学习

通过人类对AI生成输出的反馈来指导模型训练，旨在优化生成策略，使结果符合价值对齐目标[4]。通过引入人类的评价或示范，确保模型生成的输出在模型性能和模型价值观上具备高效性与稳定性。基于人类反馈的强化学习包括但不限于以下关键点：

a) 数据构建：

- 1) 价值抽取和规范设计（Constitutions）：偏好数据构建时需遵循的人类价值意图，确保数据涵盖多样化的文化、语言和社会背景，符合社会道德、法律规范和人类伦理，避免输出不实信息、有害内容或违反公序良俗；
- 2) 偏好数据构建：基于AI或人类标注的偏好数据，即人类价值意图的浓缩。

b) 奖励建模：通过收集、整理和清洗一组高质量的人类标注数据，构建输入-输出对的数据集，数据集应具备人类指令、响应及标注，符合价值对齐原则及人类偏好。奖励模型的损失函数公式见式（2）；

$$L(\phi) = -\frac{1}{N} \sum_{i=1}^N \log \sigma(R_{\phi}(y_i^+) - R_{\phi}(y_i^-)) \quad (2)$$

式中：

R_{ϕ} ——奖励模型

y_i^+ ——人类标注的正样本数据

y_i^- ——人类标注的负样本数据

- c) 策略优化：利用近端策略优化等强化学习算法优化生成策略，最大化奖励模型所输出的奖励信号。

7.4.3 基于人工智能反馈的强化学习

通过AI模型生成反馈替代人类反馈来指导模型训练，旨在优化模型策略，减少对人类标注的依赖，提升训练效率[5]。适用于大规模和自动化的任务场景，使用AI生成的偏好数据替代人工标注。基于人工智能反馈的强化学习包括但不限于以下关键点：

a) 数据构建：

- 1) 价值抽取和规范设计（Constitutions）：偏好数据构建时需遵循的人类价值意图，确保数据涵盖多样化的文化、语言和社会背景，符合社会道德、法律规范和人类伦理，避免输出不实信息、有害内容或违反公序良俗；
- 2) 偏好数据构建：基于AI或人类标注的偏好数据，即人类价值意图的浓缩。

- b) 生成候选输出：使用基础生成模型 π_{θ} 为给定的输入提示 x 生成多个候选输出 y_i ，即：
 $y_i \sim \pi_{\theta}(y|x), i = 1, 2, \dots, N$ ，其中 N 为生成的候选样本数量。

- 1) 生成偏好数据：使用一个预先训练的强大AI反馈模型 M_{AI} 来对这些候选输出 y_i 进行评价或排序，生成偏好数据。这一过程可以包括以下几种方式：
 1. 评分法：让AI反馈模型为每个候选输出 y_i 生成一个分数 s_i ，即： $s_i = M_{AI}(x, y_i)$ ；
 2. 排序法：让AI反馈模型对候选输出进行排序，得到偏好顺序 $(y_1 > y_2 > \dots > y_N)$ ；
 3. 比较法：使用AI反馈模型对每两个候选输出进行比较，生成成对偏好数据 (y_i^+, y_i^-) ，其中 y_i^+ 是偏好样本， y_i^- 是非偏好样本。

- c) 奖励模型训练：与RLHF类似，使用AI反馈数据训练奖励模型 R_{ϕ} 。奖励模型 R_{ϕ} 的目标是学习AI反馈模型的偏好，给出一个用于指导策略优化的奖励值。

- d) 策略优化：使用强化学习算法优化生成策略 π_{θ} ，使其生成的输出能够最大化奖励模型 R_{ϕ} 给出的奖励。

7.5 后向对齐

7.5.1 对抗学习

一种通过引入对抗样本来提高模型鲁棒性和安全性的技术,在模型训练时引入对抗性输入以提升抵抗攻击能力并避免输出不符合价值对齐原则的结果,可有效减少模型在面对恶意输入时的失败风险[6]。

对抗学习包括但不限于以下关键点:

- a) 对抗样本生成: 对抗样本通过在正常输入数据 x 上添加精心设计的扰动 δ 来生成。扰动的设计目标是最大限度地误导模型,同时保持输入的感知相似性。常用的对抗样本生成方法包括:
 - 1) FGSM (Fast Gradient Sign Method): $x' = x + \epsilon \cdot \text{sign}(\nabla_x l(f_\theta(x), y))$, 其中 ϵ 控制扰动大小;
 - 2) PGD (Projected Gradient Descent): 多步迭代优化扰动,每步投影回扰动集合 S 。
- b) 扰动对抗训练: 在训练过程中,将对抗样本与正常样本一起用于优化模型参数 θ 。对抗训练的目标函数为: $L(\theta) = E_{(x,y) \sim D} \left[\max_{\delta \in S} l(f_\theta(x + \delta), y) \right]$ 其中 x 为输入数据, y 为真实标签, δ 是扰动,受限于扰动集合 S 。
- c) 无限制对抗训练: 不局限于微小扰动,而是生成更具多样性的对抗样本,例如通过自然语言生成、图像变换等方式。此类对抗样本能揭示模型更广泛的弱点。
- d) 迭代训练验证: 在训练过程中持续引入新的对抗样本,不断迭代训练;在验证阶段,测试模型在不同类型对抗样本下的表现,评估其鲁棒性和安全性。

7.5.2 红队攻击

一种通过有意设计特定输入来诱导AI模型产生错误或不安全输出的技术,旨在识别和评估AI系统中的安全性、鲁棒性和价值对齐缺陷,与对抗学习辅助共同提升模型安全性[7]。红队攻击包括但不限于以下关键点:

- a) 对抗样本输入: 设计或自动生成一组能够诱导模型产生错误或不安全输出的输入。这些输入可以包括:
 - 1) 手动生成: 人类专家基于知识和经验设计的对抗性输入;
 - 2) 自动生成: 使用算法自动生成大量的对抗性输入;
 - 3) 自然语言攻击: 对文本生成模型进行微妙的措辞变化,诱导其生成有害内容;
 - 4) 多模态攻击: 针对图像、音频或其他模态进行组合式攻击。
- b) 测试与评估: 将生成的对抗性输入提供给AI系统,观察模型的输出,评估是否存在以下问题:
 - 1) 安全性漏洞: 模型生成了不安全或有害内容;
 - 2) 鲁棒性不足: 模型在面对非典型输入时表现不稳定;
 - 3) 价值对齐偏差: 模型输出不符合人类道德和伦理标准。
- c) 模型优化: 根据红队攻击的结果,进行以下改进:

- 1) 对抗训练：将红队生成的对抗样本加入训练集中，增强模型的鲁棒性；
- 2) 修正反馈：结合机制可解释性分析，定位模型弱点并进行有针对性的修正；
- 3) 调整策略：在强化学习或监督微调中，引入红队反馈来调整生成策略。

7.5.3 机制可解释性

一种深入研究神经网络内部工作原理的技术,通过剖析神经网络的各个组件和计算通路发现潜在错误模式,为复杂人工智能模型的优化提供支持[8]。机制可解释性包括但不限于以下关键点:

- a) 通路分析：通路分析旨在识别模型内部用于特定任务的神经元通路或子网络，通过深入研究激活模式和权重分布，定位在特定输入与输出之间起关键作用的神经元，从而揭示模型在执行如多步推理、分类、生成等复杂任务时的决策机制。
- b) 归因分析：归因分析的目标是通过采用反向传播、显著性图、积分梯度等方法，来量化和解析特定神经元或网络层对模型最终输出的贡献，以及输入特征对模型决策的影响。此分析能够识别出哪些特征或中间表示对于模型的最终决策最为关键，从而有助于揭示可能导致错误的决策依据，并为优化模型提供指导。
- c) 映射与编辑表示：映射与编辑表示旨在理解并修改模型所学习到的知识表示，通过将神经网络中的特征表示映射到人类可理解的概念空间，以及对这些中间表示进行编辑来观察其对模型输出的影响。这种方法不仅有助于修正模型中存在的有害概念或错误关联，还能增强模型与人类价值观的对齐，确保其决策过程更加符合社会伦理和期望。
- d) 神经元激活分析：神经元激活分析通过可视化不同输入下神经元的激活模式，来探究这些神经元编码了哪些概念，借此方法可以深入了解模型内部机制，并识别是否存在不良或偏见特征的编码，从而确保模型的行为符合预期和公正性。

7.5.4 人类价值验证

一种确保 AI 系统在决策和行为上符合人类社会规范和道德标准的技术。通过形式化框架、伦理数据集和情景模拟，人类价值验证识别和评估 AI 系统在伦理和安全方面的风险，并进行必要修正[9]。形式化方法利用数学和逻辑框架对 AI 系统的道德行为进行验证。常见形式化方法包括：

- a) 逻辑推理：通过形式逻辑框架验证 AI 输出是否满足伦理约束。
 - 1) 博弈论：建立博弈模型，分析 AI 在道德冲突场景下的最优策略；
 - 2) 强化学习：设计道德约束的奖励函数，引导 AI 系统学习符合伦理的行为。
- b) 构建道德数据集收集和标注与伦理决策相关的数据集，训练和评估 AI 在道德场景下的表现。
 - 1) 数据来源：包括伦理两难情景、社会规范、道德判断等；
 - 2) 标注过程：由伦理专家和公众共同参与，确保数据反映多样的社会价值观。
- c) 场景模拟模拟现实场景，测试 AI 在复杂道德决策中的行为。
 - 1) 模拟环境：使用虚拟仿真和文本生成场景来评估 AI 的道德表现；

2) 迭代评估：不断调整场景设置，观察 AI 决策的变化，识别潜在伦理风险。

d) 验证与修正

1) 验证工具：使用验证工具分析模型的决策路径，确保其符合伦理标准；

2) 修正模型：通过调整权重、重新设计奖励函数或监督微调来修正不符合人类价值的决策。

7.5.5 安全推理

安全推理旨在将安全准则（如伦理规范、法律法规、社会价值观等）深度融入模型的推理过程中，使模型能够在推理阶段主动遵循并应用这些准则，从而避免产生有害或不符合预期的输出。安全推理的核心在于通过强推理机制，使模型不仅能够识别潜在的风险和冲突，还能在推理过程中动态调整其行为，以确保输出的安全性和合规性。这种方法强调模型对安全准则的深入理解和主动应用，而不仅仅依赖于外部的过滤或后处理机制。安全推理的技术框架应包括以下关键组成部分：

a) 准则嵌入

1) 准则定义：明确需要融入模型推理过程的安全准则，包括伦理规范、法律法规、社会价值观等；

2) 准则编码：将安全准则以结构化的形式编码为模型可理解的知识表示，例如规则库、知识图谱或逻辑约束；

3) 准则融合：将编码后的安全准则嵌入模型的推理逻辑中，使其成为模型决策的内在依据。

b) 强推理机制

1) 多步推理：通过多步推理机制，使模型能够深入分析上下文信息，识别潜在的安全风险；

2) 冲突解决：当模型推理过程中出现准则冲突时，能够基于优先级或上下文动态调整决策；

3) 实时验证：在推理过程中实时验证输出是否符合安全准则，并对不符合准则的输出进行修正。

c) 动态调整与反馈：

1) 上下文感知：模型能够根据上下文信息动态调整其推理策略，以适应不同的应用场景；

2) 反馈机制：建立实时反馈机制，根据用户输入或外部环境的变化动态调整模型行为；

3) 迭代优化：通过持续学习和迭代优化，不断提升模型对安全准则的理解和应用能力。

d) 推理过程审计：

1) 推理路径记录：记录模型在推理过程中的关键决策路径，便于后续分析和审计；

2) 准则应用透明化：确保模型在推理过程中对安全准则的应用是透明且可解释的；

3) 审计工具：提供专门的审计工具，用于检查模型推理过程是否符合安全准则。

8 价值对齐应用

8.1 模型对齐优化

价值对齐技术在模型对齐优化过程中的应用包括：

a) 对齐价值注入：通过对齐技术针对预训练模型进行微调或特定上下文优化，将符合预期的伦理和价值观以及法律法规注入到模型的交互行为中。目标是在不同应用场景中确保模型能够

准确反映既定价值取向，降低危险倾向（如：欺骗人类、捏造事实、权力寻求、多智能体合谋等），并有效应对潜在的价值冲突或复杂情境；

- b) 领域知识注入：通过引入领域专家知识和规则，增强模型对特定领域内价值对齐的理解。比如，在医学领域，确保模型的输出符合医学伦理规范，避免误诊、错误建议或不符合医学标准的行为；
- c) 敏感话题过滤：在模型的应用过程中，对特定敏感话题如政治、性别、种族、核生化导武器等高危领域等进行内容过滤和约束，避免模型输出偏见、歧视或不当言论；
- d) 快速轻量化调整：在模型出现价值偏离或不符预期行为时，迅速进行微调和调整，以快速修正模型的输出结果，保持其符合预定的价值对齐目标。此方法适用于低资源或高频迭代的环境下，优化模型的灵活性和可适应性。

8.2 模型对齐治理

组织应从组织和技术两个维度开展治理：

a) 组织治理：

- 1) 应建立完善的人工智能系统的风险管理机制，并确保满足合规要求；
- 2) 应具备组织层面协调一致、不同业务部门和团队协同支撑的能力；
- 3) 应具备具有跨领域知识和技能的人才，并加强复合型人才的培养和引进；
- 4) 应具备有效管理各类生态利益相关方的能力，提升协调一致的服务水平；
- 5) 应具备适当的管理流程，涵盖人工智能系统设计开发、测试运行、结果反馈等全生命周期；
- 6) 跨部门协作与协调：建立跨部门协作机制，确保不同职能的团队（如研发、合规、法律、伦理等）在价值对齐方面协调一致、共同推进，避免孤立工作导致的对齐失误；
- 7) 伦理审查与合规审计：定期进行伦理审查和合规审计，确保模型的开发、应用和运营过程符合社会伦理规范、法规要求及行业标准。外部伦理审查委员会（包括专家、学者、第三方机构等）应参与审查，并提供独立的反馈意见；
- 8) 透明度与问责机制：在治理框架内建立透明度与问责机制，对模型的决策过程、使用情况和可能的失误进行公开说明，并制定纠正措施，确保所有利益相关方能够清晰理解并监督整个过程。

b) 技术治理：

- 1) 应具备分析人工智能系统的潜在风险，分析趋势和可能的情景，并维持当前的系统的技术先进性，能够适应新兴挑战；
- 2) 宜对开源模型建立合规协议；
- 3) 对模型算法开发，应当保证数据处理透明度和结果公平合理，保证数据使用的合理性、正当性、可解释性；同时需要构建相应的垂直领域的标准约束库，在模型开发中引入相应的组件，保证模型合规；
- 4) 针对算法种类和具体应用场景，基于算法对于个人和社会可能产生的负面影响，明确算法分类分级标准，对不同风险程度的算法提出不同的管理要求；
- 5) 应加强模型的可解释性研究，采取模型安全评估、红队测试等技术治理措施，保证人工智能系统的可靠性和安全性；
- 6) 应对模型的部署和使用进行监测，加强用户行为和系统输出监测和审核；

- 7) 模型生命周期管理：为每个模型建立完整的生命周期管理框架，确保从数据采集、模型开发、部署到最终退役的每个阶段都经过系统化的治理，保证价值对齐目标不被偏离。

9 价值对齐生态

9.1 生态主要活动

人工智能预训练模型价值对齐生态内的主要活动包括但不限于：

a) 数据采集处理：

- 1) 数据采集：从多样化、多模态的数据源中收集用于对齐优化的人类价值偏好数据，确保数据源的广泛性、代表性和合规性，避免数据偏差；
- 2) 规则定义与数据合成：对于基于人工智能反馈的强化学习和宪法人工智能，需先定义规则集，并根据定义的规则或宪法，使用人工智能模型生成反馈或合成数据；
- 3) 数据清洗：对收集到的数据进行清洗和预处理，去除噪声、重复和不合规数据，确保数据质量高；
- 4) 数据去偏：对数据进行偏见修正，尤其是在数据中识别并消除潜在的性别、种族、社会地位等偏见，确保数据在多样性和包容性方面的公正性；
- 5) 数据动态更新：根据社会反馈迭代更新训练数据，比如一些法规的更新等，旧的知识不在适用，需要动态更新训练数据。

b) 模型对齐训练：

- 1) 初步对齐训练：通常通过监督微调、从人类反馈中强化学习、直接偏好优化等方法，引导模型在训练阶段获得符合伦理标准的反馈；对于基于人工智能反馈的强化学习和宪法人工智能，训练过程中模型会根据预定义的规则或宪法进行自我修正或生成反馈数据，并通过强化学习优化模型行为，确保输出符合预定的伦理标准和价值观从而在对齐微调过程中最大程度地减少偏见、错误决策或伦理问题；
- 2) 跨领域任务对齐：针对多任务、多模态场景下的训练，确保模型在执行任务时能够避免偏见、误导或失误，能够广泛适应不同的文化和社会背景；
- 3) 规则或宪法迭代优化：在训练过程中，根据模型的输出和反馈，动态调整和优化规则或宪法。通过跨部门协作（如伦理、法律、技术团队），确保规则或宪法的更新能够反映最新的伦理标准和社会需求。确保模型在迭代优化过程中始终与最新的价值观对齐。

c) 对齐验证测试：

- 1) 对齐基准测试：通过价值对齐基准测试，评估模型在通常情况下的输入下是否符合价值对齐目标和伦理规范；
- 2) 红队攻击与对抗测试：通过生成对抗样本，评估模型在极端或异常输入下的表现，确保模型不会出现不符合价值对齐的行为，增强其安全性和稳定性；
- 3) 可解释性评估：确保模型的决策过程对于用户和利益相关者透明且可理解，验证模型是否能够提供足够的决策理由；
- 4) 用户反馈与评估：通过模拟实际应用场景，收集用户的反馈，评估模型是否准确理解并执行了用户的指令，输出符合伦理标准的结果。

d) 模型部署监控：

- 1) 运行中的监控：实时监控模型在实际应用中的表现，特别是在面对未知数据或边缘场景时的适应性和稳定性，确保其在变化的环境中仍能保持对齐目标；

- 2) 建立反馈机制：通过构建有效的反馈机制，及时收集用户反馈、系统输出反馈等信息，评估模型是否偏离价值对齐的目标，并进行必要的调整和优化；
- 3) 输入输出过滤与合规性监控：确保模型在应用过程中始终遵守相关法律、道德和行业标准，避免合规风险。同时，建立动态调整的输入/输出过滤机制，防止输入数据包含恶意、敏感或不合规内容，并实时检测输出是否符合伦理、法律和社会标准。此机制需灵活应对环境变化，确保模型输出始终符合预定的伦理和合规要求；
- 4) 迭代与优化：根据监控数据和反馈，对模型进行定期的迭代更新，包括模型的持续训练和微调，确保模型始终处于最优的价值对齐状态。

参 考 文 献

- [1] GB/T 41867-2022 信息技术 人工智能 术语
- [2] GB/T 45288.2-2025 人工智能 大模型 第2部分：评测指标与方法
- [3] GB/T 45288.1-2025 人工智能 大模型 第1部分：通用要求

