可信 AI 操作指引

（V0.5 版）

1 前言

人工智能（AI）技术和应用的发展正在呈现加速趋势，在大力推动经济社会创新发展的同时，人工智能暴露出难以解释、偏见歧视等普遍性风险，人工智能信任问题也日益受到重视。为进一步细化《人工智能行业自律公约》文件要求，构建健康的人工智能产业发展环境，中国人工智能产业发展联盟特组织开展《可信 AI 操作指引》的编写工作。

2 适用范围

本文件规定了可信 AI 的概念，详细阐述了 AI 系统的可信要求，旨在为人工智能产业健康发展提供参考。

本文件在现阶段适用于 AI 系统提供方，主要是指研究、设计、制造、运营和服务人工智能系统的企业、高校、科研院所、行业组织、个人以及其他实体。并且，在适用于 AI 系统提供方的同时，也要综合考虑 AI 系统部署情况以及运行环境等。

另外，本文件只适用于较成熟的商用 AI 系统，不适用于非商业化，及处于研究或实验阶段中的 AI 系统。

3 概念

本文件中所指的“可信 AI”，即人工智能在技术、产品、

应用、运营、管理等方面，应能遵循以人为本、公平公正、增进民生福祉、防止危害社会、避免侵犯公众利益和个人合法权益的总原则，并能够重点满足《人工智能行业自律公约》中可靠可控、透明可释、保护隐私、明确责任、多元包容等基本要求。

当前，本文件中所指的“可信 AI”聚焦于人工智能系统，即“可信的人工智能系统”。

4 使用说明

本文件在第 5 节《可信要求》中，对所有 AI 系统提出了通用的基本要求和实践指引。

建议 AI 系统提供方可根据不同的应用场景和技术领域，参考第 5 节《可信要求》中的内容做出具有针对性的实践操作。

5 可信要求

可信 AI 应满足可靠可控、透明可释、保护隐私、明确责任、多元包容等五项基本要求。依据现阶段人工智能技术发展水平和产业界的实际情况，本文件规定了三个级别的要求等级，代表了不同的重要性、通用性和可操作性，由强到弱分别为一级要求、二级要求和三级要求，分别对应文中“应”、“宜”、“可”的文字描述，其中一级要求共 30 条，二级要求共 25 条，三级要求共 5 条，具体如表 1 所示。

表 1 可信要求等级对照表			
基本要求	具体要求	要求内容	要求等级（一级：★，二级：★★，三级：★★★）
5.1 可靠可控	5.1.1 系统安全性	5.1.1 1）	★★
		5.1.1 2）	★★
		5.1.1 3）	★★
		5.1.1 4）	★★★
		5.1.1 5）	★
		5.1.1 6）	★★
		5.1.1 7）	★★
	5.1.2 系统成熟度	5.1.2 1）	★
		5.1.2 2）	★★
		5.1.2 3）	★★
	5.1.3 系统稳健性	5.1.3 1）	★★
		5.1.3 2）	★
		5.1.3 3）	★★
	5.1.4 人类对系统的监督和接管能力	5.1.4 1）	★
		5.1.4 2）	★
5.2 透明可释	5.2.1 系统决策过程描述	5.2.1 1）	★
		5.2.1 2）	★★
		5.2.1 3）	★
		5.2.1 4）	★

		5.2.1 5）	★
	5.2.2 系统技术意图描述	5.2.2 1）	★
		5.2.2 2）	★
	5.2.3 系统可复现性描述	5.2.4 1）	★★
		5.2.4 2）	★★★
	5.2.4 外部监督和审查渠道	5.2.5 1）	★
		5.2.5 2）	★★★
5.3 保护隐私	5.3.1 数据收集使用合法合规	5.3.1 1）	★
		5.3.1 2）	★
		5.3.1 3）	★
		5.3.1 4）	★
		5.3.1 5）	★
	5.3.2 数据主体隐私保护	5.3.2 1）	★
		5.3.2 2）	★
		5.3.2 3）	★
		5.3.2 4）	★
		5.3.2 5）	★
	5.3.3 未成年人隐私保护	5.3.3 1）	★
		5.3.3 2）	★
	5.3.4 确保数据安全	5.3.4 1）	★★
		5.3.4 2）	★★
		5.3.4 3）	★★
	5.3.5 防范数据泄露	5.3.5 1）	★★
		5.3.5 2）	★★

		5.3.5 3）	★★
		5.3.5 4）	★
		5.3.5 5）	★★★
5.4 明确责任	5.4.1 明确权利义务	5.4.1 1）	★★
		5.4.1 2）	★★★
	5.4.2 确定责任主体	5.4.2 1）	★★
		5.4.2 2）	★
		5.4.2 3）	★★
	5.4.3 探索 AI 创新保险机制	5.4.3 1）	★★
		5.4.3 2）	★★
		5.4.3 3）	★★
5.5 多元包容	5.5.1 产品需求多样化	5.5.1 1）	★
		5.5.2 2）	★★
	5.5.2 训练数据全面化	5.5.2 1）	★
		5.5.2 2）	★
	5.5.3 算法公平性测试验证	5.5.3 1）	★★
		5.5.3 2）	★

具体可信要求内容如下：

5.1 可靠可控

确保人工智能系统在其整个生命周期内安全、可靠、可控地运行。评估系统自身安全和潜在风险，不断提高系统的成熟度、稳健性和抗干扰能力。确保系统可被人类监督和及时接管，避免系统失控的负面影响。

5.1.1 系统安全性

1）对于 AI 系统输入环节，影响系统安全性的因素有：常规信道攻击(重放攻击)、传输信道攻击和侧信道攻击等传感器欺骗手段。针对以上问题，AI 系统提供方宜采取传感器增强(忽略相应的攻击频段)、输入滤波等措施来检测恶意破坏系统构造的攻击信息,实现对系统输入环节的安全增强。

2）对于 AI 系统数据预处理环节，影响系统安全性的因素有：插值算法逆向、轮询推测还原等重采样攻击手段。针对以上问题，AI 系统提供方宜采取对输入预处理引入随机化或重采样等质量监测等方法来增大攻击难度。

3）对于 AI 系统机器学习模型训练环节，影响系统安全性的因素有：数据投毒、攻击模型后门、对抗样本、逆向、萃取等攻击手段。针对以上问题，AI 系统提供方宜采取鲁棒性机器学习、数据清洗等方法，使污染数据和正常数据产生分布差异；采取检测还原后门、输入过滤、神经元剪裁等策略去除后门；采取直接对抗训练、梯度掩模、输入变化、模型集成、模型正则化、可验证性防御、引入随机波动、对抗样本检测等方法进行对抗样本防御；采取定期删除无关训练细节、差分隐私模型、联邦学习等方法进行逆向防御；采用近似处理、模型水印等方法进行萃取防御。

4）对于 AI 系统输出环节的劫持、篡改、逆向、萃取等攻击手段，AI 系统提供方可采取输出值近似处理、引入随机

波动、限制近似数据频繁访问等方法来提高攻击难度。

5）AI 系统提供方应对系统进行安全测试，以预防潜在的风险。对于代码漏洞风险，AI 系统提供方可利用模糊测试等传统漏洞检测方法进行预防；对于学习偏差、过拟合等问题，AI 系统提供方可根据系统实际情况进行测试样本触发潜在异常的检测，还可以通过输入预估网络，进行是否触犯安全限定检查。

6）AI 系统提供方宜对系统部署安全监控，建议在安全事故前测压演练，事故中实时监控，事故后故障分析，并且定期维护，及时调优，保证能够及时发现、解决或避免系统的安全问题。

7）AI 系统提供方宜建立物理安全隔离区域，并且只允许专业人员进行维护，有效避免恶意攻击以及潜在竞争对手的破解。

5.1.2 系统成熟度

1）AI 系统提供方应根据应用场景，为用户提供相适应的解决方案，精准化选择准确率指标、测试方法以及测试数据集等。

2）AI 系统提供方宜针对系统成熟度开展自评估测试，综合考虑准确率以及其他可能对系统成熟度造成影响的因素，如时间间隔[1]、业务并发量[2]和可用性[3]等，并依据测试结

[1] 时间间隔是指 AI 系统在收到用户的请求到响应请求之间的时间间隔。
[2] 业务并发量是指 AI 系统最大能承受的业务并发量。

果不断优化算法、模型、数据集质量等。

3）AI 系统提供方宜委托第三方检测机构[4]对 AI 系统成熟度的相关指标进行测试，如准确率、时间间隔、业务并发量、可用性等，并公开测试结果和相关测试证书。

5.1.3 系统稳健性

1）AI 系统提供方宜定期对训练数据进行清洗，以防范攻击者通过修改训练数据内容和分布，来影响模型的训练结果。

2）AI 系统提供方应根据系统的应用场景，给出系统稳健性的最低要求。即根据对所有已知的对抗样本攻击或未知攻击的防御效果得出被测试系统模型的稳健性下界，当系统受到外部干扰或处在恶劣环境条件等情况下依然能维持其性能水平的能力。

3）AI 系统提供方宜建立系统置信度的记录，以应对通过训练得到的深度神经网络中的隐藏后门，即模型后门攻击。对于模型后门相对应的标签，很小的输入扰动会引起该标签对应置信度明显的变化，利用这一特性可以通过对置信度的记录发现模型后门，从而保障系统的稳健性。

5.1.4 人类对系统的监督和接管能力

1）AI 系统提供方应针对系统设置后备计划，确保在突发情况下，AI 系统可自动调节恢复或者被专业人员快速接管。

[3] 可用性是指 AI 系统在事件发生后迅速恢复运行状态的能力。
[4] 第三方检测机构是指独立于当事双方的另一方负责公平公正的检测，并可以得出令双方信服的检测结果的机构。

2）AI 系统提供方应针对系统设置“一键关停”能力，确保在突发情况下，能够人为终止 AI 系统服务的能力。

5.2 透明可释

不断提高人工智能系统透明度，促进对人工智能系统的普遍理解。对于系统决策过程、数据构成、系统开发者和技术实施者意图，能够在适当场景下予以描述、监督和重现，积极回应遭受人工智能系统不利影响者的质疑和意见。

5.2.1 系统决策过程描述

1）AI 系统提供方应根据用户需求和应用场景，建立向不同背景的利益相关者解释 AI 系统决策逻辑的能力。例如，对于专业人员，解释 AI 系统中的模型和算法逻辑；对于普通用户，解释 AI 系统的功能逻辑；对于内部监管决策者，解释在相应领域使用 AI 系统的原因。

2）AI 系统提供方宜以 AI 系统可解释为目标，在保证系统性能能够满足任务需求的前提下，尝试使用可解释性较强的模型替代复杂的黑盒模型。如使用传统机器学习模型替代复杂的深度学习模型，或尝试使用集成学习模型、贝叶斯深度学习模型以融合传统机器学习模型较强的可解释性和复杂深度学习算法模型的高性能，或使用其他模型或相关工具在算法逻辑层面提高 AI 系统决策的可解释性。

3）AI 系统提供方应根据不同的应用场景，如自动驾驶应用、日常生活娱乐应用等，评估用户是否可以介入 AI 系

统、设定相关参数并监督其运行情况、以及修改决策结果，并根据用户参与度的评估情况，合理选择相应的决策模型。

4）AI 系统运行过程中，如果有来自于用户的数据，或通过用户的数据加工衍生的数据，需要临时存储，应对用户可知，并根据实际情况，可选择向用户说明具体存储位置。

5）AI 系统提供方应披露 AI 系统决策结果对用户的影响方式和潜在风险，以及用户是否被赋有“一键关停”等终止 AI 系统服务的能力。

5.2.2 系统技术意图描述

1）AI 系统提供方应建立适当的交流机制，告知用户与 AI 系统交互的情况。如：设置一个功能模块，通过一种通俗易懂的表达方式，如文字、图形标识、语音提示等，明确地告知用户当前是否正在与人工智能系统进行交互。

2）AI 系统提供方应根据系统的复杂度、应用场景和用户的实际需求，披露 AI 系统的基本功能、性能表现、使用要求、面向对象、以及 AI 系统在决策流程中扮演的角色。

5.2.3 系统可复现性描述

1）AI 系统提供方宜建立完善的管理机制对系统的训练过程进行记录，提高系统的可复现性，主要包括：建立完善的数据集管理机制，结合现有数据管理策略和工具，详细记录系统各版本训练过程中训练集、测试集的来源和构成情况，以及训练过程所采用的数据预处理操作；建立完善的模型训

练管理机制，详细记录训练模型时所用的硬件平台、系统配置、软件框架、模型版本、模型初始化、超参数、优化算法、分布式运行策略、网络速率、指标、测试结果、以及所采用的其他技巧和工程技术手段等。

2）AI 系统提供方可在研发阶段根据真实部署环境中数据分布的变化，在用户的反馈和配合下，不断更新和调整数据集和算法模型，保证 AI 系统研发的时效性，从而在上线部署之后能够提供与研发阶段相近的性能表现。

5.2.4 外部监督和审查渠道

1）AI 系统提供方应与用户建立使用反馈沟通渠道，用户可通过该渠道反馈使用 AI 系统过程中遇到问题。

2）AI 系统提供方可与用户建立决策审查渠道，根据系统中数据标注和模型训练的实际情况，处理解释决策结果的相关问题。当用户对 AI 系统的决策结果产生质疑，并对其造成实质性影响时，可通过该渠道要求对决策流程及结果进行审查，并请求获取相应的解释报告。

5.3 保护隐私

坚持以合法、正当、必要的原则收集和使用个人信息，尊重和保护个人隐私，特别加强对于未成年人等特殊数据主体的隐私保护，强化技术手段，确保数据安全，防范数据泄露及滥用等风险。

5.3.1 数据集收集使用合法合规

1）AI 系统提供方应从合法渠道收集法律法规允许收集的个人信息，向用户公开产品或服务所具有的个人信息收集功能，不得欺诈、诱骗、误导、强迫用户提供其个人信息。

2）AI 系统提供方在直接收集个人信息前应告知用户收集、使用用户信息的目的、方式和范围，并通过明确的用户操作征得用户的授权同意。

3）AI 系统提供方间接获取个人信息时应确认来源的合法性，明确获得的同意授权范围，是否授权同意转让、共享、公开披露等等。

4）AI 系统提供方收集个人敏感信息时应获得用户对其个人敏感信息进行特定处理行为的明示同意，确保明示同意是用户在完全知情的基础上自愿给出的具体的、清晰明确的愿望表示。

5）AI 系统提供方作为个人信息控制者应满足本节内容，AI 系统提供方作为个人信息处理者应根据相关法律法规和与个人信息控制者签订的合同保护用户隐私。

5.3.2 数据主体隐私保护

1）AI 系统提供方应按照法律法规的要求对用户信息进行去标识化处理，使其在不借助额外信息的情况下，无法识别个人信息主体。

2）AI 系统提供方应按照法律法规的要求对用户信息进行匿名化处理，使个人信息主体无法被识别或者关联，处理后的信息不再能合理识别信息主体。

3）AI 系统提供方应在使用个人信息时不得超出用户授权的范围。

4）AI 系统提供方应按照法律法规的要求保证数据主体享有对个人信息处理活动的知情权、同意权、查询权、更正权、拒绝权、删除权及行使途径。

5）AI 系统提供方停止运营产品或服务时，应按照法律法规的要求彻底删除或匿名处理持有的个人信息。

5.3.3 未成年人隐私保护

1）AI 系统提供方收集、处理年满 14 周岁未成年人的个人信息前，应征得未成年人或其监护人的明示同意；不满 14 周岁的，应征得其监护人的明示同意。

2）未成年人或其监护人要求 AI 系统提供方删除、屏蔽其个人信息时，AI 系统提供方应及时采取删除、屏蔽、断开链接等必要措施。

5.3.4 确保数据安全

1）AI 系统提供方宜建立用户隐私安全保护机制，明确责任部门与人员及安全职责、处罚机制，对负责人员签署保密协议，对用户信息的重要操作设置内部审批流程，并对安全管理人员、数据操作人员、审计人员的角色进行分离设置。

2）AI 系统提供方宜建立安全审计机制，记录用户信息处理活动，防止非授权访问、篡改或删除审计记录，及时处理用户信息违规使用，审计记录保存时间符合法律法规要求。

3）AI 系统提供方宜制定并更新用户隐私安全事件应急预案，定期组织内部相关人员进行应急响应培训和应急演练。

5.3.5 防范数据泄露

1）AI 系统提供方宜分析用户隐私数据的分布和流转，识别并根据数据的暴露节点、暴露对象、暴露途径等建立完整的数据泄露防护机制。

2）AI 系统提供方宜启用防火墙、数据泄露防护系统、数据库加密系统等工具预警、阻断、追溯外部攻击，避免数据泄露。

3）AI 系统提供方宜启用终端安全管理工具、终端数据泄露防护、网关数据泄露防护、虚拟桌面、数据加密、数据脱敏等工具规范内部人员对数据的操作管理。

4）AI 系统提供方传输和存储用户敏感信息时应采用加密、脱敏等安全措施。

5）AI 系统提供方可建立利用联邦学习[5]方法来进行多方联合学习模型[6]的训练，以应对模型过拟合等问题引发的隐私泄露等问题，以保证系统的隐私安全。

[5] 联邦学习(federated learning)

[6] 多方联合学习模型：在该模型下，训练数据并不会离开本地.各方建立一个虚拟共有模型，通过加噪机制交换参数，对共有模型进行共同训练。

5.4 明确责任

不将人工智能系统用于非法或违反伦理的目的。明确人工智能研发、设计、制造、运营和服务等各环节主体的权利义务，在损害发生时，能够及时确定责任主体。倡导相关企业和组织在现有法律框架下创新保险机制，分担人工智能产业发展带来的社会风险。

5.4.1 明确权利义务

1）AI 系统提供方宜开展内部 AI 治理培训，引导和规范与 AI 系统相关的人员遵守基本的法律、伦理准则、以及可信 AI 基本原则，明确内部不同岗位的人员职责，并定期进行内部考核。

2）AI 系统提供方可在一定范围内开展试错性质的测试活动，但应告知受测人员的保密义务，并及时回应受测人员提出的意见和建议。

5.4.2 确定责任主体

1）AI 系统提供方宜成立人员结构合理的内部审查委员会，需要包含技术、运营、法务、伦理等不同专业人士，制定内部政策和审查机制，应对相关的特定问题，以确保 AI 系统符合其提供方的核心价值观和原则。

2）AI 系统提供方应建立完善的系统日志功能，覆盖 AI 系统全生命周期的各个环节，确保能够全面、准确地回溯 AI 系统。

3）AI 系统提供方宜建立适当的知识转移机制，在内部组织结构发生重大变化或 AI 系统涉及的关键人员发生流动时，降低潜在的操作差异化风险。

5.4.3 探索 AI 创新保险机制

1）AI 系统提供方宜鼓励 AI 研发人员进行技术创新，同时通过与现有 AI 技术发展水平相匹配的责任审查机制，帮助研发人员理解监管要求，降低技术创新风险，确保技术应用符合现有的法律法规、AI 系统提供方的核心价值和原则、以及公认的道德伦理标准。

2）AI 系统提供方宜制定伦理风险应对机制和相应的救济机制，对因 AI 系统出现的错误造成人身、财产损失等情况，积极开展救助活动，并协商赔付。

3）AI 系统提供方宜建立针对 AI 系统全生命周期的人员管理、保障管理机制，对模型窃取、用户数据盗取等情况，根据风险及危害等级制定相应的处罚措施。

5.5 多元包容

促进人工智能系统的包容性、多样性和普惠性。加强跨领域、跨学科、跨国界的合作交流，凝聚人工智能治理共识。力争实现人工智能系统参与人员多元化，训练数据全面化。持续测试和验证算法，不因人种、性别、国籍、年龄和宗教信仰等歧视用户。

5.5.1 产品需求多样化

1）AI 系统提供方在保护用户个人信息隐私的前提下，应明确基本和潜在服务用户范围，对不同类别的用户进行实际需求调研，整理分析不同用户的交互方式和操作习惯，保证产品原型设计阶段满足用户多样化需求，并尽量考虑潜在产品服务对象的隐性需求。

2）AI 系统提供方在符合产品投放地区法律法规的前提下，宜实施产品用户体验测试，在产品开发和生产过程中，根据 AI 系统的应用场景，有阶段性地选择面向不同地域、国籍、性别、年龄等用户进行实际产品试用，通过用户反馈和数据分析形成需求链条跟踪机制，进而适时调整产品结构。

5.5.2 训练数据全面化

1）AI 系统提供方应保证训练数据集的数据的多样性，并可针对数据集中不同的类别群体进行分析测试，同时依据测试结果相应地调整数据集的结构。

2）AI 系统提供方应依据任务需求完成训练数据集设计方案，通过统计学的方式或相关工具集，检查模型训练数据集中样本与方案的符合程度，保证训练数据的准确性和完整性。

5.5.3 算法公平性测试验证

1）AI 系统提供方宜将公平性度量纳入算法评价内容，兼顾各类群体特征信息，以防算法存在偏见。若算法上有设

计缺陷，则改进算法，减轻算法对特定变量对应的权重的依赖，尽量避免对某些特定群体做出带有歧视和偏见的决策。

2）AI 系统提供方应保证算法决策判断的鲁棒性，在无特殊要求的前提下，充分考虑适用场景下可能出现的特殊情况，保证算法输出结果不会由于某些环境指标改变而发生分歧。因此，AI 系统提供方在合法合规的前提下，应构建包含通用场景和特殊场景的测试数据集，经过对算法充分测试验证，保证算法对面向对象决策结果一致。

6 结语

因人工智能技术正在快速发展，从可操作、可实现的角度出发，本文件仅是本操作指引的一个阶段性版本，供政府部门、科研机构、企业参考使用，后续还将持续修订完善。