ICS 35.040.01
CCS I 6599

T/EI 7491-2025

团　　体　　标　　准

T/EI 7491-2025

生成式人工智能的道德设计规范

Ethical design specifications for generative artificial intelligence

2025-06-29 发布　　2025-06-29 实施

广州市从化区青年创新创业协会　发 布

前　言

本文件按照 GB/T 1.1-2020 给出的规则起草。

本文件由国家工业设计研究院（生态设计领域）提出并归口。

本文件起草单位：浙大城市学院、浙江大学、浙江工商职业技术学院、浙江商业技师学院、杭州慧慧科技有限公司。

本文件主要起草人：应卫强、吴冬俊、沈小丽、张玲燕、姚琤、张旭生、徐雯洁、刘镇宇、庄楠、黄岚清、应放天。

生成式人工智能的道德设计规范

1 范围

本文件规定了生成式人工智能在设计、开发、部署和使用过程中应遵循的道德设计原则、要求、方法及评估准则，旨在确保生成式人工智能系统的开发和应用符合人类的道德价值观和社会伦理规范，促进生成式人工智能技术的健康、可持续发展。

本文件适用于各类生成式人工智能系统的研发机构、企业、组织以及使用生成式人工智能服务的相关方，涵盖设计、开发、部署和使用全过程。

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 38647.1-2020 信息技术 安全技术 匿名数字签名 第1部分：总则

GB/T 38648-2020 信息安全技术 蓝牙安全指南

GB/T 45288.1-2025 人工智能大模型 第1部分：通用要求

T/TMAC 120-2024 人工智能产品伦理风险管理指南

ISO/IEC TR 24368:2022 信息技术 人工智能 伦理和社会问题概述

ISO/IEC 23053:2022 使用机器学习（ML）的人工智能（AI）系统框架

ISO/IEC 22989:2022 信息技术—人工智能概念和术语

ISO/IEC 42001:2023 人工智能管理体系

ISO/IEC 23894:2023 信息技术 人工智能 风险管理指南

3 术语和定义

3.1

生成式人工智能 Generative Artificial Intelligence

基于一定的算法和模型，能够自动生成文本、图像、音频、视频等各种形式内容的人工智能技术。

3.2

道德设计 Ethical Design

在生成式人工智能系统的设计过程中，充分考虑道德和伦理因素，将道德原则融入到系统的架构、

算法、数据和交互设计等各个方面，以确保系统的行为和输出符合道德规范。

3.3

偏见　Bias

由于数据、算法或其他因素导致生成式人工智能系统产生不公平、歧视性或片面的结果。

3.4

可解释性　Explainability

生成式人工智能系统能够以人类可理解的方式解释其决策过程和输出结果的能力。

3.5

透明度 Transparency

生成式人工智能系统的开发、运行和决策过程对相关利益方（如用户、监管机构等）保持公开和可访问的程度。

4 道德设计原则

4.1 尊重人类自主性

生成式人工智能系统应尊重人类的自主决策和选择权利，不得通过操纵或误导等方式侵犯人类的自主性。系统应提供清晰、准确的信息，让用户能够自主决定是否使用系统以及如何使用系统的输出结果。

4.2 公平与非歧视

生成式人工智能系统应避免因种族、性别、年龄、宗教、残疾等因素产生偏见和歧视。在数据收集、算法设计和模型训练过程中，应确保数据的多样性和代表性，避免对特定群体的不公平对待。

4.3 责任与可追溯性

生成式人工智能系统的开发者和使用者应承担相应的责任，确保系统的行为和输出符合道德和法律要求。系统应具备可追溯性，能够记录和追踪其决策过程和输出结果，以便在出现问题时能够进行责任认定和追溯。

4.4 隐私保护

生成式人工智能系统应严格保护用户的隐私信息，遵循相关的隐私保护法律法规。在数据收集、存储和使用过程中，应采取必要的安全措施，确保用户的个人信息不被泄露、滥用或非法获取。

4.5 有益性与无害性

生成式人工智能系统应致力于为人类社会带来积极的影响和价值，避免产生危害人类安全、健康、尊严和社会稳定的结果。系统的开发和应用应符合社会公共利益和道德准则。

4.6 透明度与可解释性

生成式人工智能系统应保持透明度，向用户和相关利益方公开其基本原理、算法架构、数据来源等

信息。同时，系统应具备可解释性，能够以通俗易懂的方式解释其决策过程和输出结果，以便用户和相关利益方能够理解和评估系统的行为。

5 道德设计要求

5.1 数据层面

5.1.1 数据收集

a）数据收集应遵循合法、正当、必要的原则，获得用户的明确授权。
b）收集的数据应具有多样性和代表性，避免因数据偏差导致系统产生偏见。
c）数据收集过程应确保数据的质量和准确性，避免收集虚假、错误或不完整的数据。

5.1.2 数据标注

a）数据标注应遵循统一的标准和规范，确保标注的准确性和一致性。
b）标注人员应具备相应的专业知识和技能，避免因标注错误导致系统产生偏见。
c）数据标注过程应进行质量控制和审核，确保标注结果的可靠性。

5.2 算法层面

5.2.1 算法设计

a）算法设计应遵循道德设计原则，避免因算法缺陷导致系统产生偏见和不公平的结果。
b）算法应具备可解释性，能够以人类可理解的方式解释其决策过程和输出结果。
c）算法设计应考虑系统的鲁棒性和可靠性，避免因输入数据的微小变化导致系统产生不稳定或错误的输出结果。

5.2.2 算法更新

a）算法应定期进行更新和优化，以适应不断变化的需求和环境。
b）算法更新应遵循道德设计原则和要求，确保更新后的算法不会产生新的偏见和问题。
c）算法更新过程应进行充分的测试和验证，确保更新后的算法的稳定性和可靠性。

5.3 系统层面

5.3.1 系统架构设计

a）系统架构设计应考虑道德设计原则和要求，确保系统的各个组件和模块之间的交互和协作符合道德规范。
b）系统架构应具备可扩展性和灵活性，以便在系统运行过程中能够及时调整和优化系统的功能和性能。
c）系统架构设计应考虑系统的安全性和可靠性，采取必要的安全措施防止系统被攻击和破坏。

5.3.2 系统交互设计

a）系统交互设计应遵循用户友好原则，提供清晰、简洁、易懂的界面和操作指南。
b）系统交互设计应充分考虑用户的需求和体验，避免因交互设计不合理导致用户产生误解或困惑。

c）系统交互设计应提供必要的反馈和提示，让用户能够及时了解系统的运行状态和输出结果。

5.4 应用层面

5.4.1 应用场景评估

a）在将生成式人工智能系统应用于具体场景之前，应对应用场景进行全面的评估，分析应用场景可能带来的道德和伦理风险。

b）应用场景评估应考虑应用场景的特点、目标和需求，以及系统的性能和能力，确保系统的应用符合道德设计原则和要求。

c）应用场景评估结果应作为系统设计和开发的重要依据，指导系统的功能和性能设计。

5.4.2 应用过程管理

a）在生成式人工智能系统的应用过程中，应建立完善的管理机制，确保系统的使用符合道德和法律要求。

b）应用过程管理应包括用户培训、使用规范制定、监督检查等环节，确保用户能够正确、合理地使用系统。

c）应用过程管理应及时处理用户反馈和投诉，对系统的使用过程进行持续改进和优化。

5.4.3 应用效果评估

a）应对生成式人工智能系统的应用效果进行定期评估，评估系统的应用是否达到了预期的目标和效果。

b）应用效果评估应考虑系统的道德性和伦理性，评估系统的应用是否符合道德设计原则和要求。

c）应用效果评估结果应作为系统优化和改进的重要依据，推动系统的不断发展和完善。

6 道德设计方法

6.1 多学科协作

生成式人工智能的道德设计需要计算机科学、伦理学、法学、社会学等多学科的专业知识和技能。在系统的设计和开发过程中，应组建跨学科的团队，充分发挥各学科的优势，共同参与系统的道德设计。

6.2 利益相关者参与

利益相关者包括用户、开发者、监管机构、社会公众等。在系统的设计和开发过程中，应充分考虑利益相关者的需求和意见，通过问卷调查、访谈、研讨会等方式，广泛征求利益相关者的意见和建议，确保系统的设计和开发符合利益相关者的利益和期望。

6.3 道德风险评估

在系统的设计和开发过程中，应定期进行道德风险评估，识别系统可能存在的道德风险和问题。道德风险评估应采用科学、客观的评估方法和指标，对系统的各个方面进行全面的评估。评估结果应作为系统设计和开发的重要依据，及时采取措施降低和消除道德风险。

6.4 道德准入嵌入

将道德准则和原则嵌入到系统的设计和开发过程中，通过技术手段实现道德要求的自动化执行。例如，在算法设计中引入公平性约束，在系统交互设计中提供道德提示等。

6.5 持续学习与改进

生成式人工智能技术不断发展和变化，道德设计也需要不断学习和改进。在系统的运行过程中，应建立持续学习和改进机制，及时收集和分析系统的运行数据和用户反馈，总结经验教训，不断优化系统的道德设计。

7.评估与认证

7.1 评估指标体系

建立科学、客观、全面的评估指标体系，对生成式人工智能系统的道德设计进行评估。评估指标体系应包括数据层面、算法层面、系统层面和应用层面等多个方面的指标，具体指标可根据系统的特点和应用场景进行调整和确定。

7.2 评估方法

采用定性和定量相结合的评估方法，对生成式人工智能系统的道德设计进行评估。定性评估方法包括专家评审、案例分析等，定量评估方法包括指标打分、模型评估等。评估过程应严格按照评估指标体系和评估方法进行，确保评估结果的客观、公正、准确。

7.3 认证机制

建立生成式人工智能道德设计认证机制，对符合道德设计标准的系统颁发认证证书。认证机构应具备相应的资质和能力，认证过程应严格按照认证标准和程序进行。获得认证的系统可以在市场上获得更好的信誉和竞争力，同时也为用户提供了一种选择可靠系统的参考依据。

8 实施与监督

8.1 组织实施

生成式人工智能系统的研发机构、企业和组织应建立健全道德设计管理体系，明确各部门和人员的职责和权限，制定详细的实施计划和方案，确保道德设计标准的有效实施。

8.2 监督检查

监管机构应加强对生成式人工智能系统的监督检查，定期对系统的道德设计进行评估和审查。对于不符合道德设计标准的系统，应责令其限期整改；对于严重违反道德和法律规定的行为，应依法予以处

罚。

8.3 社会监督

鼓励社会公众对生成式人工智能系统的道德设计进行监督，通过举报、投诉等方式反映系统存在的问题。研发机构、企业和组织应及时处理社会公众的反馈和投诉，积极改进系统的道德设计。