《人工智能设计的伦理准则》(第 2 版) 概要

作者：IEEE 自主与智能系统伦理全球倡议项目

简介

随着自主和智能系统的应用和影响无处不在，我们需要建立社会与政策方面的指南，从而确保这些系统以人为本，并服务于人类价值和伦理准则。为了能够以积极的、非教条的方式推动自主与智能系统的发展，我们科技界需要加强自我反思，需要围绕我们的想象、我们明确或隐含的价值观、我们的机构、符号和表征开展公开和诚实的讨论。

自主与智能系统不应只是实现功能性目标和解决技术问题，而且应造福人类。只有如此，才能使人与技术之间建立更高层级的信任，这是在人类日常生活中卓有成效地普遍使用这些系统的前提。

正如亚里士多德所阐述的那样，“幸福（Eudaimonia）”是一种将人类福祉定义为社会最高美德的实践。幸福大致可以被理解为“繁荣”，它始于有意识的沉思，而伦理思考有助于我们确立理想的生活方式。

无论我们的伦理实践是西方的（亚里士多德主义的，康德主义的），东方的（神道教的，儒家的），非洲的（乌班图式的），或者其他不同传统的，只要自主与智能系统的开发能尊重不可剥夺的人权，尊重用户有益的价值，我们就能将人类福祉的提升作为算法时代衡量进步的优先指标。衡量并表彰经济全方位繁荣应当变得比追求生产效率提升和 GDP 增长这样的单方面目标更为重要。

关于自主与智能系统伦理的 IEEE 全球倡议使命

通过教育、培训和授权，确保从事自主与智能系统设计开发的利益相关方优先考虑伦理问题，只有这样，技术进步才能增进人类的福祉。

所谓“利益相关方”是指任何参与自主与智能系统的研究、设计、制造或信息交流的个人或团体，包括实现这些技术的大学、机构、政府和企业。

我们的目标是：《人工智能设计的伦理准则》将提供一些洞察和建议，为未来从事相关科技领域的技术专家的工作提供重要参考。为了实现这一目标，在当前版本的《人工智能设计的伦理准则》（第 2 版）中，我们提出了一些相关的“议题”和“建议”，希望能够促进符合这些原则的国家政策和全球政策的制定。

IEEE 全球倡议汇聚了来自六大洲的数百名参与者，他们具有相关的技术与人文学科的学术背景，是来自于学术界、

产业界、社会研究领域、政策研究领域以及政府部门的思想领袖。这些思想领袖齐聚一堂，共同发现一些迫在眉睫的现实问题，并达成共识。

IEEE 全球倡议的第二个目标是根据《人工智能设计的伦理准则》，为制定 IEEE 标准提供建议。正是《人工智能设计的伦理准则》(第 1 版和第 2 版)和 IEEE 全球倡议的成员启发和推动了 IEEE P7000™标准工作组。该工作组对于任何人的加入都是自由开放的。

查询更多信息或加入某个工作组，请点击下列链接：

IEEE P7000™ -解决系统设计中的伦理问题的建模过程

IEEE P7001™ -自主系统的透明性

IEEE P7002™ -数据隐私的处理

IEEE P7003™ -算法偏见的处理

IEEE P7004™ -儿童与学生数据治理标准

IEEE P7005™ -雇主数据治理标准

IEEE P7006™ -个人数据的 AI 代理标准

IEEE P7007™ -伦理驱动的机器人和自动化系统的本体标准

IEEE P7008™ -机器人、智能与自主系统中伦理驱动的助推标准

IEEE P7009™ -自主和半自主系统的失效安全设计标准

IEEE P7010™ -合乎伦理的人工智能与自主系统的福祉度量标准

我们是谁

电气电子工程师学会（IEEE）关于自主与智能系统的伦理考虑的全球倡议（“IEEE 全球倡议”）是 IEEE 的一个项目。IEEE 是世界上最大的专业技术组织，致力于推进技术发展、造福人类，该组织在 160 多个国家拥有超过 42 万会员。

IEEE 全球倡议汇聚了相关技术和科学群体的多种声音，以便大家及时发现问题并达成共识。

IEEE 将根据《知识共享署名-非商业性使用 3.0 美国许可》发布《人工智能设计的伦理准则》（EAD）的各个版本。

根据该许可的条款，组织或个人可以随时自行采纳本文件的各部分内容。我们预计选取《人工智能设计的伦理准则》中的部分内容和主题提交正式的 IEEE 流程，其中包括标准的制定。

IEEE 全球倡议和《人工智能设计的伦理准则》是 IEEE 的一个更大计划的一部分，该计划被称为 IEEETechEthics™，致力于在技术伦理领域开展开放的、广泛的和包容的对话。

《人工智能设计的伦理准则》第 2 版（概览）

I. 宗旨

智能和自主的技术系统的设计，旨在减少日常生活中的人工活动。 正因为如此，这些新领域对个人和社会的影响已引起人们的关注。目前的讨论涉及对积极影响的倡导，也涉及关于隐私侵害、歧视、技能丧失、负面经济影响、关键基础设施的安全风险以及社会福祉之长期影响的警告。正是由于系统的这些性质，只有它们能够符合人类的道德价值和伦理原则，这些系统才能充分实现其益处。因此，我们必须建立框架，指导我们认识这些技术可能造成的技术以外的影响，并就此进行对话和讨论。

II. 目标

合乎伦理地设计、开发和应用这些技术，应遵循以下一般原则：

- **人权**：确保它们不侵犯国际公认的人权
- **福祉**：在它们的设计和使用中优先考虑人类福祉的指标
- **问责**：确保它们的设计者和操作者负责任且可问责
- **透明**：确保它们以透明的方式运行
- **慎用**：将滥用的风险降到最低

III. 目的

个人数据权利和个人访问控制

人们有权决定其个人数据的访问权限，有权利用知情同意控制其个人数据的使用，这是人类的基本需要。个人需要各种机制来帮助建立、维护其独特的身份和个人数据，还需要其他政策和做法，使他们能明确知晓融合或转售其个人信息将产生的后果。

通过经济效应增进福祉

通过价格合理的通信网和互联网的普遍接入，智能与自主技术系统可以为任何地方的人群所用并使其受益。它们可以显著改变制度和制度性关系，使其朝着更加以人为本的结构发展；它们还能促进人道主义和发展问题的解决，从而增加个人和社会的福祉。

问责的法律框架

智能系统与机器人技术的融合带动了系统的发展，这类系统模仿人类，具有部分自主性，有完成特定智力任务的能力，甚至还可能拥有人类的外貌。因此，复杂的智能和自主技术系统的法律地位问题与更广泛的法律问题交织在了一起，这些法律问题涉及如何确保问责制，以及当这类系统造成损害时如何分配责任。以下是需要考虑到的通用框架的一些例子：

- 智能与自主技术系统应适用相关的财产法
- 政府和行业利益相关者应该确定哪些决策和操作决不能委托给这些系统，并制定规则和标准，以确保人类能够有效地控制这些决策，以及能够有效地为造成的损害分配法律责任

透明和个人权利

虽然自我完善的算法和数据分析可以使影响公民的决策自动化，但法律应该强制要求透明性、参与性和准确性，包括以下目标：

- 必须允许当事人、其律师和法院可以合理地获取政府和其他国家机关采用这些系统所产生和使用的所有数据和信息
- 如果可能的话，系统中嵌入的逻辑和规则必须对监管人员开放，并接受风险评估和严格测试
- 系统应当生成用于决策的事实和法律的审计数据，并服从第三方核查
- 公众有权了解是谁通过投资来制定或支持关于这类系统的伦理决策

教育和知悉的政策

有效的政策应当保护和促进安全、隐私、知识产权、人权和网络安全，以及公众对智能与自主技术系统对社会的

潜在影响的认识。为确保政策最符合大众利益，这些政策应当：

- 支持、推广和实施国际公认的法律规范
- 提升劳动力在相关技术方面的专业知识
- 产生对研究和开发的引领作用
- 制定规则以确保公共安全和问责
- 教育公众知悉相关技术的社会影响

IV. 理论基础

经典伦理学

IEEE 全球倡议通过汲取具有两千多年历史的经典伦理学的精华来探讨数字时代人类道德。它探究了包括世俗哲学传统的既有的伦理学体系，讨论了科学的与宗教的两种路径。通过考察用以界定自主论和本体论的哲学基础，IEEE 全球倡议探讨了智能技术系统在自主能力方面声称的潜能，以及非道德系统的道德问题，并追问由非道德系统做出的决策是否会产生道德影响。

福祉指标

检验延展智能和自动化是否增进人类福祉，必须有衡量福祉的明确指标。通用的正向指标包括利润、职业安全和财政健康。这些指标尽管很重要，但未能涵盖个人或社会福祉的全部内容。心理的、社会的和环境的要素非常重要。涵盖这些要素的福祉指标将使利益评估由单纯的技术进步变得更为全面，并提供了机会来检验损害人类福祉的意外负面后果。此外，这些指标还有助于我们辨别智能技术系统会在哪些地方提高人类福祉，进而为社会创新和技术创新提供新的路径。

将价值嵌入自主系统

如果机器作为准自主的主体参与到人类社区之中，那么这些主体就应当遵守社区的社会规范与道德规范。将规范嵌入这些系统，需要对系统所在的社区有一个清晰的描述。另外，即使在一个特定的社区中，不同类型的技术也需要嵌入不同类型的规范。首先要做的就是识别系统所在的特定社区的规范，特别是与预先设计的特定任务相关的规范。

指导合乎伦理的研究和设计的方法

为了开发增强和提升人类福祉与自由的智能技术系统，基于价值的设计方法将人类进步置于技术系统开发的核心地位，这与公认的“机器应该服务于人类而不是相反”是一致的。系统的开发者应采用基于价值的设计方法，开发可持续的系统。这种系统可以依据社会成本和能够提升组织经济价值的优势这两个维度来进行评估。

V. 未来技术关切

重塑自主武器

设计造成人身伤害的自主系统与传统武器或设计不造成伤害的自主系统相比，需要考虑更多的伦理维度。这些伦理维度至少包括以下几个方面：

- 确保武器系统在人类有效的控制中
- 自动化武器的设计应包含供审计的追踪数据，以便确保可问责和可控
- 包含自适应和可学习的系统，以透明和可理解的方式向操作人员解释其推理和决策
- 培训自主系统的操作人员，其身份应可清晰识别
- 自动功能行为的实现对操作人员而言是可预测的
- 确保技术开发人员能够理解其工作的后果
- 制定职业伦理守则，妥善处理有意造成伤害的自主系统的开发

所谓的通用人工智能（AGI）和超人工智能的安全性和有益性

与其他强大的技术一样，智能和自我改善的技术系统的开发和使用涉及相当大的风险。这些风险可能来源于滥用或不良设计。然而，根据某些理论，当系统接近并超过 AGI 时，无法意料的或无意的系统行为将变得越来越危险且难以纠正。 并不是所有的 AGI 级别的系统都能够与人类利益保持一致，因此，当这些系统的能力越来越强大时，应当谨慎并确定不同系统的运行机制。

情感计算

情感是智能的核心之一。愤怒、恐惧和喜悦等情绪与动力往往是人类行为的基础。为确保智能技术系统在各种情况下最大可能地服务于人类，以及参与或服务人类社会的人造物不能通过放大或抑制人类的情感体验来造成伤害。即使是在一些系统中设计的初步的人工情绪，也会影响决策者和公众对它们的理解。

混合现实

随着混合现实技术在我们的工作、教育、社会生活和商业事务中的应用越来越普遍，混合现实技术可能会改变我们对身份和现实等概念的理解。尤其是随着技术从耳机转向更加精细、更整合的感官增强设备，混合现实世界的实时个性化能力引发了有关个人权利和个人多重身份管理的伦理问题。

（免责声明：EAD2 并不代表 IEEE 的立场或观点，而是代表 IEEE 自主与智能系统伦理全球倡议项目委员会成员的资深意见，该意见旨在为专家提供关于人工智能领域的方向性指导。）

诚挚感谢：

IEEE 自主与智能系统伦理全球倡议项目中国委员会发起组织了翻译工作，在此特别致谢其翻译工作组的成员对此文做出的特别贡献。对翻译工作做出贡献的人员为：

王亮迪： IEEE 自主与智能系统伦理全球倡议项目中国委员会主席，翻译项目的发起，策划和审定。

张文瀚： 项目经理，负责协调，统筹翻译工作的全部过程。

万岩：教授，北京邮电大学经济与管理学院，翻译工作组总协调人，负责总体协调各位参与翻译工作的专家，参与第一轮翻译工作，并撰写第一轮翻译初稿。

陈鹏： 副教授，北京语言大学信息科学学院，翻译工作组总协调人，负责总体协调各位参与翻译工作的专家，参与第一轮翻译工作，并撰写第一轮翻译初稿；参与校阅翻译终稿，并翻译校阅标准内容介绍。

刘战雄： 博士，南京农业大学政治学院，参与第一轮翻译工作，并撰写第一轮翻译初稿。

洪延青：博士，四川大学网络空间安全研究院，参与第一轮翻译工作，并撰写第一轮翻译初稿。

李伦： 所长，湖南师范大学人工智能道德决策研究所，审阅，修正并补充翻译初稿，撰写翻译第二轮翻译稿。

廖备水： 教授，浙江大学语言与认知中心，审阅，修正并补充翻译初稿，撰写翻译第二轮翻译稿；校对翻译终稿。

曹建峰： 研究员，腾讯研究院，审阅，修正并补充翻译初稿，撰写翻译第二轮翻译稿。

高旖蔚： 博士研究生，中国科学院， 参与最后翻译终稿的校对。

另外，感谢中国社科院**段伟文**、中国科学院**李真真**与**吴焦苏**、腾讯研究院**徐思彦**、浙江大学**王志坚**、今日头条**贺**

佳与**成浩**、中国科学技术发展战略研究院**廖苗**、中国人民大学**张吉豫**、北京大学**刘先华**、**郭耀**以及**和鸿鹏**对翻译工作的支持。

对翻译若有反馈和疑问，或者感兴趣参加中国委员会的工作，包括未来的翻译工作，请联系 IEEE 标准协会中国工作组 IEEE-SA@qq.com 。IEEE 诚挚欢迎广大专家学者阅览《人工智能设计的伦理准则》（第 2 版）全文。