Click here to go back to main China Committee Page

## 《以伦理为基准的设计》

--在人工智能及自主系统中将人类福祉摆在优先地位的愿景（第一版）

为了从人工智能及自主系统（AI / AS）的潜力中充分获益，我们需要超越臆想，不局限于只是寻找更多计算的能力或解决问题的能力。我们需要确保这些技术与人类在道德价值观和伦理原则方面保持一致。人工智能及自主系统远不止于是实现功能性的目标和解决技术问题，而必须以一种造福于人们的方式行事，才能使人类与技术之间达成崇高的信任，让人工智能及自主系统能够在日常生活中被人们建设性地普遍使用。

正如亚里士多德所阐述的“幸福（Eudaimonia）”，它是一种将人类福祉定义为社会最高美德的实践。幸福大致理解为“繁荣”，它始于有意识的沉思，对伦理的思考有助于定义我们希望怎样生活。通过将人工智能及自主系统的创造与其用户和社会的价值观保持一致，我们优先将人类福祉的增加作为算法时代进步的指标。

## 我们是谁？

电气电子工程师学会（IEEE）关于人工智能及自主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 40 万会员。

IEEE 全球倡议为人工智能及自主系统社群提供了一个汇集多种声音的机会，以便大家及时发现问题并形成共识。

IEEE 将根据《知识共享 署名-非商业性使用-相同方式共享 3.0 美国许可证》制定《伦理准则设计》（EAD）。

根据该许可证的条款，组织或个人可以随时自行采纳本文件的各方面内容。还拟选取《伦理准则设计》的内容和主题提交正式的 IEEE 流程，其中包括标准制定等。

IEEE 全球倡议和《伦理准则设计》使得 IEEE 在更广泛的范围推出新项目，致力于在技术伦理领域开展开放、宽泛和包容的对话，这个项目被称为 IEEETechEthics™。

## IEEE 全球倡议的使命

确保每位技术专家得到教育、培训和赋权，并在自主和智能系统的设计和开发中会优先考虑伦理问题。

“技术专家”，在这指的是任何参与人工智能及自主系统的研究、设计、制造或信息交流的人，包括实现这些技术的大学、组织和公司。
本文件体现了 100 多名全球思想领袖的集体智慧，他们分别来自学术界、科学界、政府和企业界，从事人工智能、法律和伦理、哲学、政策领域的相关工作。我们的目标是希望《伦理准则设计》中同仁们提供的见解和建议能为人工智能及自主系统技术专家未来的工作提供关键参考。为了实现这一目标，当前版本《伦理准则设计 V1》（EAD v1）确定了人工智能及自主系统的领域中的问题并提出了可选建议。

IEEE 全球倡议的第二个目标是基于《伦理准则设计》为开发 IEEE 标准提供建议。 IEEE P7000™ - 解决系统设计时伦理问题的流程模型，是第一个受到该倡议启发的 IEEE 标准项目（已批准并且正在开发中）。 另外两个标准项目，IEEE P7001™ - 自主系统的透明度和 IEEE P7002™ - 数据隐私流程也已被批准。由此可见，该倡议对人工智能及自主系统的伦理问题产生了实质性影响。

## 结构和内容

伦理准则设计包括八个部分，每个部分围绕着一个与人工智能及自主系统有关的特定话题进行阐述，分别由 IEEE 全球倡议相应的委员会组织探讨。每个委员会都列出了涉及的议题及其可选建议。以下是各委员会的摘要以及各个部分所涉及的问题：

### 一、一般原则

一般原则委员会已清晰地說明适用于所有类型人工智能及自主系统的一般性的伦理关注，即：
1. 体现人权的最高理想；
2. 排列出对人类和自然环境最大利益的优先顺序；
3. 随着人工智能及自主系统逐渐演化成为社会技术系统，缓解其带来的风险和负面影响。
委员会的目标是将其确定的原则、议题和建议选项，最终用于巩固和支撑人工智能及自主系统新伦理治理框架下的未来规范和标准。

议题：

- 我们如何确保人工智能及自主系统不侵犯人权？（制定人权原则）
- 我们如何保证人工智能及自主系统是可以问责的？（制定责任原则）
- 我们如何确保人工智能及自主系统是公开透明的？（制定公开透明原则）
- 我们如何扩大人工智能及自主系统技术所带来的利益，最小化其被滥用的风险？（制定教育和认知原则）

**二、将价值观嵌入自主智能系统**

为了开发成功而有益于社会的自主智能系统，对技术社群至关重要的是，需要理解和能够将相关的人类规范或价值观嵌入到他们的系统中。“价值观嵌入自主智能系统委员会” 针对嵌入价值观于自主智能系统已采取了较宽广的目标，以三管齐下的方式协助设计者：

1. 识别受自主智能系统影响的特定社群的规范和价值观；
2. 将这些特定社群的规范和价值观建置于自主智能系统中；以及
3. 评估社群中的这些规范和价值观，在人类与自主智能系统间的一致性与相容性。

议题：

1. 拟嵌入自主智能系统的价值观并非放诸四海皆准，而是在很大程度上针对特定的用户社群和任务。
2. 道德超载：自主智能系统通常会受到多重而可能相互矛盾的规范和价值观的约束。
3. 自主智能系统可能会有内置数据或算法的偏颇，从而不利于某些群体成员。
4. 一旦建立了相关的规范集(自主智能系统在特定社群中的特定角色)，尚未明确如何将这些规范嵌入到计算机架构中。
5. 建置于自主智能系统的规范必须要与相关社群的规范相互兼容。
6. 让人类和自主智能系统之间达到正确的信任水平。
7. 对自主智能系统价值观的一致性进行第三方评估。

**三、指导伦理研究和设计的方法论**

现代的人工智能及自主系统组织应该确保人类福祉、赋权和自由作为其发展的核心。为了创造有助于实现这些宏伟目标的机器，“指导伦理研究和设计的方法论委员会”提出了若干议题和建议选项，以保障相关组织所采用的系统设计方法可以创造人类价值，如《世界人权宣言》中所定义的人权。与人类价值相一致的设计方法应该成为现代人工智能和自主系统组织关注的焦点，以推动人类基于伦理准则的向前发展。机器应该服务于人类，而不是相反。 系统设计方法符合伦理，才能确保人工智能在追求商业上的经济效益和社会中的社会效益之间达到均衡。

议题：

1. 伦理学不只是一门专业学位课。
2. 我们需要跨学科和跨文化的教育模式来应对人工智能及自主系统中的独特问题。
3. 区分嵌入在人工智能设计中不同文化中的价值观。
4. 行业内缺乏基于价值观的伦理文化和实践。
5. 缺乏具有价值观意识的领导层。
6. 缺乏赋权，以致于没有引起人们对价值观的关注。
7. 技术社群缺乏主人翁意识和责任意识。
8. 吸纳不同利益相关者进而为人工智能及自主系统发展创造良好环境。
9. 文件匮乏阻碍伦理设计。
10. 算法不一致或对算法缺乏监督。
11. 缺乏独立的审查组织。
12. 使用黑箱组件。

四、通用人工智能(AGI)和超级人工智能(ASI)的安全与福祉

未来具高能力的人工智能系统（有时被称为通用人工智能；AGI），可能对世界产生像农业或工业革命一般规模的变革性影响，带来前所未有的世界繁荣。“通用人工智能和超级人工智能安全福祉委员会” 提出了若干议题和建议选项，通过人工智能社群一致的努力，协助确保这种转变向着正面的方向发展。

议题：

1. 随着人工智能系统能力的提升 - 即评估其在横跨更多样化领域之际，具较高度自主性优化更复杂目标函数的能力 - 非预期或非刻意的行为变得越来越危险。
2. 改良未来具更广泛能力的人工智能的安全性可能会有困难。
3. 研究人员和开发人员在开发和部署愈加自主和有能力的人工智能系统时，将面临一系列逐渐更复杂的伦理和技术安全议题。
4. 未来的人工智能系统对世界可能具有像农业或工业革命般规模的影响。

**五、个人数据与个人访问控制**

数据不对称是个人信息领域的一个主要的伦理困境。

为了解决这种不对称性，“个人数据和个人访问控制委员会”提出了若干议题和可选建议，以表达人们的一项基本需求：作为唯一身份的守护者，人们能够定义、访问和管理其个人数据。 委员会认识到，尽管没有完美的解决方案，任何数字工具都可能被黑客攻击，但是，委员会建议启用一个人们可以控制他们的自身感觉的数据环境，并提供了工具和发展实践的案例，以消除数据不对称，推动未来往积极的方向发展。

议题：

1. 个体如何在算法时代定义和组织他/她的个人数据？
2. 个人身份信息的定义和范畴是什么？
3. 如何定义”对个人数据的控制”？
4. 如何重新定义”数据访问“以尊重个人？
5. 如何重新定义“对个人数据的同意”，以使其尊重个人？
6. 那些看起来微不足道分享的数据会被误判为个人不希望共享。
7. 数据处理程序如何确保访问和收集数据的结果（正面和负面）对个人是明确的，以便用户在真正知情的情况下选择“同意”？
8. 一个人是否能有个性化的人工智能或算法守护者？

六、重构自主武器系统

相对于传统武器和不以造成伤害为目的的自主系统，以造成物理伤害为目的的自主系统会产生额外的伦理后果。有关这种自主系统的专业伦理可以而且应该有更高的标准，得到更广泛的关注。概括地说，“重构自主武器系统委员会”建议技术组织接受如下内容：对武器系统进行有意义的人类控制对社会有益；用审计追踪机制确保问责制的有效运行，从而确保实现这种控制；创造这些技术的人了解其工作的可能后果；专业伦理准则能恰当地应对那些旨在造成伤害的武器系统。

议题：

1. 专业组织行为准则通常存在重大的漏洞，它们据此会忽视持有者的工作以及他们所创造的工件和介质，这些持有者及其作品都要在同等程度上遵循相同的价值观和标准。
2. 人工智能、自主系统和自主武器系统等重要概念的定义混乱不清，这阻碍了对关键问题进行更具实质性的讨论。
3. 自主武器系统天然地可用于隐蔽目的，且不可溯源。
4. 有多种方式可以用来减轻对自主武器系统之行动的责任。
5. 自主武器系统可能具有不可预测性（视其设计和战术应用而定），（机器）学习系统使预测其应用的问题更加复杂。
6. 使自主武器系统开发合法化的努力创立了一些先例，这些先例在中期内在地缘政治上具有危险性。
7. 将人类监督排除在战场外，很容易导致无意中侵犯人权，以及无意中升级紧张局势。
8. 自主武器系统有各种各样的直接和间接客户，这将造成一种复杂且令人担忧的扩散和滥用的局面。
9. 自主武器系统中的自动化类型天然地鼓励冲突快速升级。
10. 尚无有关自主武器系统的设计保证验证方面的标准。
11. 很难理解自主武器系统和半自主武器系统相关工作的伦理边界。

七、经济/人道主义议题

旨在减少我们日常生活中的人类干预的技术、方法论和系统正在快速发展，并准备以多种方式改变个人的生活。 “经济/人道主义议题委员会”的目的是找到构成“人类与技术”全球生态系统的关键驱动因素，阐述其带来的经济和人道主义后果，并指出可以实施解决方案的关键机会，以化解关

键节点的紧张局势。委员会的目的是：就人类及其机构和新兴的信息驱动技术之间的关系中存在的主要担忧提出一个务实的方向，促进跨学科跨部门的对话，从而更充分地听取专家及同仁就这些问题方向性的思考。

问题：

1. 媒体对人工智能及自主系统的误解会迷惑公众。
2. 人们通常没有从除了市场以外的角度来看“自动化”。
3. 在机器人/人工智能时代，就业的复杂性被忽视。
4. 技术变革太快，现有的劳动力培训/再培训方法跟不上。
5. 任何人工智能政策都可能会减缓创新。
6. 世界各地的人工智能和自主技术的发展水平不均衡。
7. 缺乏对个人信息的访问和理解。
8. 需要增加发展中国家在 IEEE 全球倡议中的积极代表
9. 人工智能及自主系统的出现可能加剧发达国家和发展中国家之间及其内部的经济和权力结构差异。

八、法律

人工智能及自主系统的初期发展引发了很多复杂的伦理问题。这些伦理问题几乎总是直接转化为具体的法律挑战，或者引发出一些困难的附带法律问题。“法律委员会”认为，这方面有很多工作需要律师来参与，但迄今为止，尽管是在迫切需要的领域，却很少有从业者和学者参与进来。 律师需要参与这些领域中与监管、治理、国内和国际立法相关的讨论。人工智能及自主系统为人类和我们的地球带来了巨大利益，值得律师们深思熟虑地为未来付出努力。

议题：

1. 我们如何提高自主和智能系统的问责制和可验证性？
2. 我们如何确保人工智能是公开透明的，且尊重个人权利？
3. 如何设计人工智能系统，以确保受这些系统的危害时可施以法律问责？
4. 如何确保以尊重个人数据完整性的方式来设计和部署自主和智能系统？

**这份文件是如何准备的**

本文件采用开放、协作和共识的建立方法，遵循 IEEE 标准协会的“行业连接”项目程序开展。随着行业不断磨练和改进其对新兴技术问题的思考，“行业连接”项目旨在促进组织和个人之间的合作，帮助孵化潜在的新标准活动和标准相关的产品及服务。

## 如何引用《伦理准则设计》

请按照以下方式引用《伦理准则设计》第 1 版：

IEEE 关于人工智能及自主系统的伦理考虑的全球倡议。《伦理准则设计：人工智能及自主系统以人类福祉为先的愿景（第一版）》 IEEE，2016.

http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html。