doi:10.3772/j.issn.2095-915x.2021.04.009

IEEE 人工智能社会伦理规范实施机制研究

开放科学（资源服务）标识码（OSID）

黄国彬 黄恋 陈丽

北京师范大学政府管理学院 北京 100875

摘要：[目的 / 意义] 人工智能社会伦理规范是合理运用人工智能技术的有力保障，通过成文的社会伦理规范，实践者在开展与人工智能有关的项目时能够遵循明确的可行依据，而研究其在具体应用中是如何被实施的则具有必然的现实意义。[方法 / 过程] 本文通过对 IEEE 人工智能社会伦理规范的相关政策进行文本分析，从实施基础、目标定位、实施流程、监督约束、程度衡量等方面梳理出一套可推广的实施机制，并将其归纳为准备、实施、检验三个阶段。[结果 / 结论] 该 IEEE 人工智能社会伦理规范实施机制不仅切合项目流程，也重视内部组织管理和外部利益协调，能够为人工智能技术社会伦理规范在不同领域发挥作用提供参考。

关键词：人工智能；社会伦理规范；实施机制

中图分类号：G31，G35

Research on the Implementation Mechanism of IEEE Social Ethical Standard for Artificial Intelligence

HUANG Guobin HUANG Lian CHEN Li

School of Government, Beijing Normal University, Beijing 100875, China

Abstract: [Objective/ Significance] Social ethical standard of artificial intelligence(AI) is a powerful guarantee for rational use of AI technology. Through the written standard, practitioners can follow a clear and feasible basis when carrying out projects and it is of certain practical significance to study how it is implemented in specific applications. [Methods/Process] Based on the text analysis of the relevant documents of IEEE social ethical standard for artificial intelligence, this paper combs out a set of promotable implementation mechanism from the aspects of implementation basis, target positioning, implementation process, supervi-

作者简介 黄国彬（1979-），博士，副教授，研究方向为信息法学、信息分析；黄恋（1995-），硕士研究生，研究方向为信息分析、信息资源管理，E-mail：578387176@qq.com；陈丽（1996-），硕士研究生，研究方向为信息分析、信息资源管理。

引用格式 黄国彬 , 黄恋 , 陈丽 . IEEE 人工智能社会伦理规范实施机制研究 [J]. 情报工程 , 2021, 7(4): 105-116.

doi:10.3772/j.issn.2095-915x.2021.04.009

sion, and degree measurement, and summarizes it into three stages: preparation, implementation, and test.[Results /Conclusions] The implementation mechanism of IEEE social ethical standard for artificial intelligence not only conforms to the project process, but also attaches importance to internal organization management and external interest coordination, which can provide reference for AI social ethics and its standard to play a role in different fields.
Keywords: Artificial intelligence; social ethical standard; implementation mechanism

引言

人工智能技术在各行业领域都获得了研发与应用的探索[1-4]。随着其相关实践力度的加大，为了维护人类利益，使人工智能技术在合乎社会伦理道德的前提下更好地服务人类社会，很多机构纷纷建立了其各自的人工智能社会伦理规范。不同机构制定的人工智能社会伦理规范各有其合理性和现实性，而就其如何具体落实到人类社会，能够在真正意义上处理人与智能技术的关系则需要建立有效的实施机制。

针对 IEEE 标准协会 (IEEE Standards Association) 在 2019 年发布的自动与智能系统领域伦理规范设计（Ethically Aligned Design，EAD）进行剖析，并就其可适用于各行业的实施机制进行研究和提炼，将能够为不同领域人工智能技术社会伦理规范的落实提供参考，有助于积极推动人工智能社会伦理规范发挥现实作用。

1 研究现状

1.1 国内研究现状

对CNKI以“人工智能”并“伦理道德”或“社会伦理”为主题词进行检索，一共获得目标文献共计 198 篇。从学科分布来看，除自动化技术相关的学科领域外，有关人工智能社会伦理在现实应用方面的研究讨论主要集中在法律、商业、教育三大领域，这些具体问题的提出和应对建议主要基于教学经验与文献基础。教育领域以刘斌[5]、雷晓维[6]等人的研究为代表，前者重点探讨教育领域的主体在智能化进程中所需要形成的智能教育素养，并从主体自身与外部政策、系统和各方职能部门四方面探讨如何具体提升该素养；后者则主要针对整个教育过程，从评价、管理、保障等方面探讨人工智能在该领域的落实方法。商业领域以张庆龙[7]为代表的学者主要探讨在智能财务方面的技术应用所带来的系列变革，并提出具体实施时应重点考虑隐私安全、数据保护与责任划分问题。而法律领域以罗洪洋等[8]为代表，主要探讨了在整个司法过程中，人工智能技术的定位与风险应对。对目标文献进行分析时发现，国内学者们在结合社会伦理对现实问题作探讨时，常着眼于某一行业领域人工智能技术的融合路径，基于此，再对技术应用过程中的潜在伦理问题进行挖掘。

就其中的伦理规范问题而言，由于国内如今权威机构出台的有关人工智能社会伦理规范的政策标准文本较少，主要集中在生物领域，专门针对国内某一政策标准如何实施的研究文

章还基本空白，一些学者如李伦等[9]、刘伟等[10]将人工智能社会伦理本身看作研究点，认为社会伦理只是人工智能伦理研究的一部分，社会伦理的实施应落脚到人机关系优化上，最终为维护社会稳定、公平、正义服务。而如张世龙[11]等学者则在综合考察科技创新与社会伦理道德间的冲突问题后，提出从信息沟通、教育宣传、法律、责任、约束等五个方面建立协调机制。更有闫坤如等[12]学者针对人工智能的责任伦理、风险伦理等一系列问题提出笼统的解决方案。这类研究从宏观层面探讨了伦理实施路径，但其主体并未涉及到社会伦理规范。国内学者们关注的话题实则还处于人工智能社会伦理规范构建层面的探讨，内容方面主要以宏观政策的方向性建议为主要研究点。如段伟文[13]的研究认为，应该立足人机关系，开展对人工智能的价值校准和伦理调试。以医学领域郭晓斐等[14]的研究为例，其就人工智能在医疗领域可能面临的系列法律问题为切入点，认为行业应联合多领域专家共同成立人工智能伦理委员会，对智能技术从伦理性和安全性等方面进行评估。这些研究中对于衡量评估指标并未涉及，但已经在建设层面有所探讨。

1.2 国外研究现状

对 web of science 以“AI ethics”与“standard”或“principle”为主题词进行检索，获得目标文献共计 209 篇，其中，大部分目标文献都处于人工智能伦理规范的制定层面。与国内存在差异的是，除自动化技术相关领域外，有关人工智能社会伦理在现实应用方面的研究讨论的学科分布主要集中在医学和法律领域。法律领域中以 Choong-Kee[15]为代表的学者，研究某一行业如汽车行业中自动驾驶的人工智能法律责任明确问题；而还有以 Luetge 和 Poszler[16]为代表的学者则对某行业整体已发布的完整的人工智能社会伦理规范进行注释和进一步说明，试图通过已有的准则规范来帮助应用主体制定行业政策。医学领域中以 Cecilia 等[17]为代表的学者主要关注当新的技术手段被引入治疗应用时所应遵循的道德规范，其也探讨了在实施过程中相关道德规范的局限性和挑战；一些学者如 Felicia 和 Amitabha[18]在内则重点探讨具体的医疗环节如何遵循社会伦理道德规范对人工智能和人类的职能进行分工。存在一些文献处于两领域交叉地带，如以 Jean 和 Emily[19]为代表的学者就基于人工智能在应用到医疗卫生行业时所涉及的数据治理相关法律责任问题进行了研究。

可以看出国外学者除了在探讨人工智能在不同行业领域的融合路径时进一步挖掘其具体落实的问题，也有就某一已发布的规范进行实践层面的探讨。然而一般涉及实践层面时，学者们主要是对现实中出现的问题提出应对建议或作应对经验总结，并未能从已发布的人工智能社会伦理规范中挖掘出一套与实际应用相适应的实施机制。如 Philip[20]在研究建筑行业应对人工智能的方法时，就结合了机器道德指南，首先对问题进行了明确，再从不同主体的角度分别总结其潜在的具体有效方法，该方法是描述性的，未能形成一套机制；而 Luciano 和 Josh 等[21]在 Atomium 的 EISMD 计划中则提出了五项人工智能伦理原则，为了支持该原则的使用，其还有针对地设计了 20 条具体实施建议，

doi:10.3772/j.issn.2095-915x.2021.04.009

并在国家政府层面对实施主体进行了限定，但该程度依然没能达到实施机制的地步。因此，目前从各种人工智能社会伦理规范中挖掘出有效的实施机制将弥补该缺憾，并从现实意义角度推动学界和业界对规范的研究和制定。

2 实施机制

从时间阶段来看，实施机制的内容主要可分为准备阶段、实施阶段、检验阶段三个模块。其中，准备阶段主要凸显实施工作开展前所需要的铺垫，包括从社会文化氛围角度奠定实施基础，并提前设计好目标定位；而在实施过程中，工作主要分两方面开展，一是垂直将人工智能伦理融入项目流程的各个环节中，二是对各个环节进行监督约束。最后，实施情况还需要检验，因此在检验阶段，给出了验证其社会伦理规范实施程度的参照标准。

结合规范内容来说，IEEE 对人工智能所应该遵循的社会伦理规范进行了描述[22]，主要从八个角度对人工智能从创造到使用进行了限制，其内容如表 1 所示。

表 1 IEEE 人工智能社会伦理规范准则

编号	关键词	内容
1	人权	人工智能的建立和运作应尊重、促进和保护国际公认的人权
2	利益	人工智能创造者应将增进人类利益作为发展的主要成功标准
3	数据力	人工智能创造者应赋予个人访问和安全分享其数据的能力，以保障人们掌控自身身份的能力。
4	有效性	人工智能的创建者和操作员应提供其有效性和适用性的证据
5	透明度	人工智能决策的基础应始终是可查证的
6	问责	人工智能的创建和运行应为所有决策提供明确的理由
7	滥用意识	人工智能创建者应防范人工智能运行中的所有潜在误用和风险
8	能力	人工智能创建者应明确说明，操作员应遵守安全有效运行所需的知识和技能

在具体的实施中，IEEE 将尊重人权作为贯彻始终的核心立足点，将为人类发展谋求利益作为实施人工智能社会伦理规范的目标，将能力作为执行社会伦理规范的必要条件。而为了确保人工智能有关项目在其项目进程中确实落实了社会伦理，主要从数据力、有效性、透明度、问责、滥用意识等方面作进一步细化。

2.1 准备阶段

在具体项目中落实人工智能伦理规范并非一蹴而就的，首先要进行的工作是确保社会能够对实施工作的开展提供支持，这就要求在现实社会中嵌入人工智能伦理规范的概念，让人们有足够的意识在开展各类人工智能技术相关的工作项目时能秉承社会伦理规范，并以之为指导工作目标的重要思想。

2.2.1 实施基础

确保社会能够为具体项目中伦理规范实施工作的开展提供支持，需要在整个社会营造人工智能社会伦理规范的文化氛围，主要分两步进行[23]。

（1）向人们介绍关于人工智能社会伦理的

新概念

介绍这一概念的有效方法需要对某一项目上不同工作层次的工作人员提出能力要求。对于基础层而言，一般要求其能力可提供直接支持，最好是具备开展意识宣传的能力。对于领导层而言，一般要求其积极推动、参与或负责以伦理为导向的人工智能发展战略的末端整合，以便在经济利益与技术为社会带来的变革之间取得平衡。

（2）帮助人们在各自的工作环境中理解人工智能社会伦理

这要求在项目的不同工作领域中，找到能够帮助推进人工智能社会伦理规范的关键人物或宣传者，通过让这些特定人员战略性地分散从而扩散信息。

在没有明确战略要求的情况下，可能有些人员在从事人工智能实践时已经自觉践行了符合其个人道德标准的社会伦理规范。因此，在项目的创建或进行人力整合时，提升这些人员的话语权是至关重要的。

2.1.2 目标定位

要将人工智能社会伦理规范纳入具体项目工作中，需要在工作的指导思想中嵌入相关内容，在此基础上去考虑诸多方面的公众利益，并以之为伦理规范的实施目标。具体说来，从宏观层面来看，在实施过程中应该要解决以下方面的问题[24]。

（1）确保人工智能规范能支持、促进国际公认的法律政策。

涉及人工智能社会伦理的国家政策和法规应以权利导向为基础，应该从一定程度上为人权标准提供国际公认的法律框架，既能考虑到技术对个人的影响，又能解决公平和歧视方面的问题。

（2）发展人工智能及伦理方面的专业知识

促进在技术和其他方面的技能发展，进一步提高决策者、监管者等就这些人工智能技术的各个方面提出明智建议和决策的能力。有效管理人工智能及相关技术需要特定人员来了解并能够分析技术、政策目标和整体社会价值观之间的相互作用。足够深度和广度的知识背景将有助于确保社会伦理规范在成功支持实践创新的同时、还能坚持社会基本原则，并能够保护公众安全。

（3）确保伦理规范是人工智能项目的核心部分

支持人工智能研究与开发工作，需要重点关注人工智能的伦理影响。为了从这些新技术中获益，同时确保它们满足社会需求和价值观，各界应积极参与支持相关的研发工作。人工智可通过一系列技术来改善社会，如通过在物流、交通、医疗、法律和环境等方面的决策支持来提高社会生活质量，或减少社会事故概率等。为了确保对个人、社会和项目产生积极影响，应该在研发、采购、使用、评估等各个环节特别注重人工智能的伦理发展和部署。在这些环节中，无论国家、企业、个人都应该衡量的重要标准包括，从人工智能技术获得的社会利益，这些技术涉及的伦理因素，以及使用人工智能增加或减少的风险等。

（4）确保人工智能技术的安全性和责任性

各行业在确保人工智能社会伦理规范适应其工作的同时，还应该解决人工智能算法的透明度、可解释性、可预测性、偏差和责任性，

doi:10.3772/j.issn.2095-915x.2021.04.009

以及风险管理、隐私安全、数据保护措施等问题。合适的解决方法取决于具体情况，应通过以人权为基础、以人类福祉为主要目标的策略来制定。对此，IEEE 建议开展人工智能项目时采用一套包括了学界、业界以及政府相关部门等人工智能利益相关方的知情程序，用以协调其人工智能社会伦理规范的具体工作，以解决与人工智能的治理和安全部署有关的问题。

2.2 实施阶段

2.2.1 实施流程

（1）树立需求意识

在人工智能相关项目工作中的为工作人员树立对社会伦理规范的需求意识是至关重要的第一步。当项目工作人员具备对伦理规范的认识和需求以后，项目团队的基础层和领导层就可以着手将伦理规范意识反映并嵌入到项目的工作流程中。IEEE 就树立工作人员对社会伦理规范的需求意识这一步提出如下可行的做法，大致说来主要是通过人员交流来落实，如：利用人工智能社会伦理规范开展网络研讨会，并在会后组织人员进行问答；把人工智能社会伦理规范交给项目领导层或者关键技术的设计负责人；确保引用符合人工智能社会伦理规范的数据和研究等。

（2）建立核心团队

确定一个跨学科的多样化团队，并向他们提供项目组织内的额外资源和权限。这要求确定关键的利益相关者和天生为特定团队宣传的人，以企业为例，这样的角色通常由产品经理担任。此外，还需要确定关键决策者和团队领导，避免在伦理规范的具体实施过程中出现将之排除在外的可能。对于技术核心人员而言，需要从一开始就明确社会伦理规范的价值，并将其作为自身角色的基本要求，使得其从意识阶段进入规范工作的行动阶段。

（3）修正项目进程

基于目前正开展的项目所要依据的核心价值观，对能够产生较大影响的社会伦理规范所产生的具体执行细则进行评估。而在项目阶段之初，不仅应该确保利益相关者在人工智能的潜在风险和利益上保持一致，还应该创建一个风险分类清单，以便适当地优先考虑和处理已经确定的风险。为了更好的项目成果，在涉及到人工智能相关的整个项目阶段需要不断寻求主题领域的专家建议以不断修正项目进程中的大小决策，这些建议将基于人工智能社会伦理规范，落脚到人工智能的开发、采购、使用等诸个环节。此外，还应该确保对受众有足够了解，并能在符合社会伦理规范的背景下向受众展现人工智能的价值，例如预先确定基于公平原则的人工智能优先级，防止代表性不足的受众（包括残疾人等）受到难以衡量的影响；或确保项目目标受众的感知或行为是具备可操作性的，对于其可能出现的潜在问题，项目方应该设计好应对框架，帮助规避意外的公平、歧视等问题。

（4）沟通协调各方

与目前正在为项目设计人工智能相关技术的人员沟通，并联系项目咨询方、拥簇方、宣传方、领导层等，在各方沟通渠道顺畅的情况下就项目管理问题和涉及的伦理问题进行讨论，以避免意外的负面影响，促进这一人工智能相关的项目建立信任度和透明度。

（5）记录工作进度

IEEE 建议在项目工作进度上应该记录的内容包括：①项目目标、受众和预期用途。②项目团队成员，包括过去和现在的。③做出的关键决策和达到的里程碑。④模型。⑤数据集。⑥记录非目标即谁不应使用项目成果，以及确定其不应被用于某些场合的用例。⑦拖延计划预期的意外影响。⑧按分类法评级分类的风险。⑨伦理审查或讨论的注释。⑩受众的反馈。⑪项目进程中出现的大小事件。

（6）组织伦理规范

就伦理规范在实施过程中的组织问题而言，IEEE 认为首先应从推广角度切入。其建议开展宣传活动，提高人工智能在社会伦理方面的透明度，鼓励围绕社会伦理规范进行对话，分享经验教训。从资源配置角度看，应在人员技能方面制定并实施伦理方面的技能培养计划；与领导层合作，确保道德规范成为管理和领导议程的一部分，以便其能够配置专门用于人工智能社会伦理规范实践的资源（资金、人员、技术）。从项目管理角度看，应首先定义可能增进或破坏人工智能伦理规范实践的文化和行为，并制定变革管理计划，以改变文化并弥合行为差距；其次，调整奖励和认可结构，以激励道德决策和发展实践；创建（或调整）管理结构，以尽量避免社会伦理规范影响失效；定义各类指标和关键绩效；调整灵活度和以人为本的项目流程，以确保伦理规范贯穿于整个开发、设计，和交付链；突出和庆祝项目进程中有关伦理规范发挥了成果影响的具体事宜。

2.2.2 监督约束

人工智能的开发、采购、使用等诸个环节必须基于适当且细致的监督约束框架，以证明其部署和运作符合社会伦理，从而促进人的尊严、权利和福祉。其所涉及到的内部或外部监督约束机构主要包括项目方和政府，具体说来，大致涵盖以下方面的内容 [25]。

（1）效力衡量

①为衡量工作提供支持。由项目资本方或政府来资助和支持不断开展的衡量工作，以便为在社会伦理规范制度中部署或可能部署的措施提供有效、公开的衡量标准。这种支持可以采取多种形式，包括投资建立一系列监管测量实验室、工作组提供直接支持，或通过雇佣认可可信的第三方来间接支持。

②促进数据集的创建。项目方或政府应促进数据集的创建，这些数据集可用于评估在社会伦理规范下人工智能相关实践应用的有效性。在建立或协助建立这类数据集时，项目方或政府必须考虑到潜在的社会价值观（如个人数据应受到保护等），确保数据集既能维护社会价值观，又可用于现实世界。IEEE 认为在这一方面所行举措都应透明，并接受公众监督。

③描述测量和验证。项目方应采取有效措施衡量其人工智能社会伦理规范的有效性，一般而言可采取两类手段，包括多系统比较评估与单系统验证。单系统验证的结论通常更具体，能够回答在特定实例中是否有效的问题。多系统比较评估的结论通常更具有普遍性，可以回答在一般的现实环境中是否有效的问题。同样，IEEE 认为应在不披露知识产权的情况下，以公众都能理解的清晰语言描述测试程序和结果，且这些描述还应向利益相关者发布以供其监督审查，有必要或适当情况下还应该面向公众作

doi:10.3772/j.issn.2095-915x.2021.04.009

公开发布。

④定义有效指标。项目方应设法定义有意义的指标作为衡量效力的标准。在选择和定义指标时，应寻求所有利益相关者的意见，以确保在社会伦理规范体系下涵盖开发、采购、使用等涉及人工智能的诸多环节。与测量验证一致，IEEE 明确指出这些指标应该很容易被公众理解。此外，还应确保项目所定义的有效性指标具有易获得性和可使用性（可供所有利益相关者使用）。最后，应该提供有关解释和响应指标的指导。

⑤开展指导工作。从政府和行业协会来说，其应开展思想层面的指导工作，向人工智能相关项目方从事社会伦理规范部署及运行的人员和受其运行结果影响的人员通报显著的有效性措施，以及有关社会伦理规范的能力和局限性。

⑥遵循指导。从项目方来说，将人工智能社会伦理规范应用于具体任务的工作人员应遵循其内部领导层或外部政府部门制定的有效性测量指导。在遵循层面来讲，主要包括在关于获得哪些指标、如何和何时获得这些指标、如何对给定结果作出反应、何时采用衡量有效性的替代方法等方面。

⑦解释和评估测量。从项目方来说，在解释和回应适用于社会伦理规范的人工智能有效性测量时，解释结果的人员应考虑到给定人工智能部署的具体目标和情况的变化。定量评估还应辅以对给定结果的实际意义及其是否需要补救的定性评估。这项工作应该由具备正确判断能力的人员来完成，一般要求其具备所需专业背景知识和实际经验。

⑧制定有效标准。项目方及外部的行业协会、政府等应合作制定衡量和报告人工智能社会伦理规范实施有效性的标准。此外，这些标准的制定还应得到学界和法律界的投入。

（2）能力规范

①提供技能指导。对项目方而言，人工智能社会伦理规范的实施如果想达到预期的效力水平，则项目内外有关领域的专家应为人工智能的开发、使用者提供所需知识、技能、经验的的指导。该指导应该是清晰易用的、可扩展的，包括对不符合这些要求时所涉风险的描述，并且应以公众都能理解的形式加以记录。

②制定管理运行政策。项目方的创建者和开发者应制定书面政策，以管理人工智能的开发运作方式。在制定这些政策时，基本思路应该基于社会伦理规范，内容应包括：人工智能的实际应用规范；有效使用的前提条件；开发、操作、评估等环节的人员所需的培训和技能；评估人工智能有效性的程序；解释人工智能结果时要考虑的因素；人工智能正常运行时，各个受影响方的预期结果；不当使用所带来的具体风险等。此外，管理政策还应明确可能涉及到人工智能的情况。基于透明度原则，这些政策也都应该是公开的。

③为系统运行提供综合保障。就项目方而言，开发者应该为人工智能的运行提供综合保障。保障措施可以包括在某些情况下向使用者发出通知和警告；根据使用者的专业水平限制其对人工智能功能的访问；在潜在的高风险条件下关闭系统等，这些保障措施应是灵活的。除此之外，还应在政府（如司法部门）的监督下制定对环境敏感的政策。

④保障受影响者权益。对于项目外部的政

府或其他机构而言，应规定向任何因为项目程序中适配了人工智能而受到影响的利益方告知人工智能在该过程中所起的作用。此外，受影响的一方还应求助于领域专家（具备知识背景者）的判断。

⑤技能应用于职责。无论是项目方还是外部政府或其他机构，从事人工智能实践、解释和执行的人员都应认识到如何基于伦理，有效地将技能应用于其专业职责。这些从业人员所属的行业协会（如果有），应通过教育计划和职业伦理守则，努力确保其成员充分了解技术能力要求，以便有效和可信地履行职责。人工智能的开发、使用、评估者应适应社会伦理规范，无论是人工智能技术人员还是非技术人员，或努力提升自身适应能力，或找出具有相应能力的个人以支持其履行职责。

（3）责任分配

①明确各方责任。这需要由项目方创建者阐明并记录所有参与人工智能开发运营的人员在成果方面的明确责任。确定法律层面的责任时，应考虑到整个设计、开发、采购、部署、运行和有效性验证环节的各个人员，相应地分配责任。以商业往来为例，在与外界沟通谈判提供使用人工智能产品、服务的合同时，人工智能的提供者和使用者、经营者应在合同条款上明确规定对所获得产品使用结果的责任范围，包括明确其对应用结果的具体责任，人工智能产生的结果与预期不同时的潜在责任，以及人工智能的使用者和经营者应了解其可能在多大程度上对不良结果承担责任。

②适应监督审查。为了给人工智能实践结果分配责任，人工智能的创建者、经营者、使用者都适用于同一套政策法律体系，应服从其监督机制和调查审核。IEEE 建议如果项目方涉及到的人工智能是由有直接公众互动的组织（如执法机构等）使用和部署的，也可以由项目外部的第三方专门审查机构进行监督调查。这也要求项目方保存其在具体进展中所遵循的各类明确的程序或决策文件。

③建立激励机制。从事人工智能开发和运营的组织应考虑建立个人和集体的激励机制，以确保人工智能的结果符合伦理标准，维持其应尽的责任。

（4）权限限定

①协调利益相关者。项目方和政府应促进利益相关者之间的对话，包括参与技术设计、开发、采购、部署、运行和验证的利益相关者；可能直接受到成果影响的利益相关者；可能间接受到成果影响的利益相关者，包括一般公众；在伦理、政治和法律方面具有专门知识的人；在透明度与其他优先事项（例如安全、隐私、所有权等）间取得平衡的问题方面具有专门知识的人。在具体实现各方利益平衡时，项目方和政府应考虑如何协调利益相互竞争的情况。

②制定透明度原则。一方面，具体政策的制定者在依据社会伦理规范为人工智能实践制定执行细则时，应要求其制定的原则既要对可能披露的信息类型间的区别足够敏感，也要对开发、使用等诸多环节足够敏感，更要对给定成果有足够敏感。另一方面，也应考虑适当保护知识产权，即使基于透明度原则，需要公开如成果效力性能等在内的某类信息，也应该对知识产权进行保护。这就要求制定者在一定程度上考虑到，其所保证的公开级别将取决于特

doi:10.3772/j.issn.2095-915x.2021.04.009

定情况下的利害关系。

③设置公共利益管理者。政策制定者应考虑为某个特别指定的“公共利益管理者”或“可信赖的第三方”设置职能，赋予其一般公众所无法获得的信息权限。这样的公共利益管理者将负责评估信息以回答公共利益问题，同时也有义务不披露在得出该回答时获取信息的具体细节。

④信息授权。对于包含合作或利益来往的项目而言，涉及到人工智能成果（产品、服务等）的谈判协商时，双方应在文件条款中，明确规定处于不同类别、层次的人具有不同的信息查看权限，即规定不同权限的人群所能获得的关于项目开发、使用、评估的信息属于不同层级。

⑤建立错误共享机制。外界政府或其他机构应在适当的情况下与项目方的人工智能开发者及其他利益相关者合作，促进建立错误共享机制，以便能够更有效地识别和纠正其在社会伦理规范下开发或使用人工智能存在的不足，例如安全应用或风险评估算法中的系统性识别错误等。

⑥提供举报人保护。在人工智能的开发或运行不符合预期，或结果没有得到正确解释的情况下，无论外部还是内部的监督机构，都应向主动提供信息的个体提供举报人保护。例如，如果某一利益方采用面部识别技术是基于非法或不符合伦理规范的目的，或其使用方式处于预期用途之外，则接收举报的机构应该向提供信息的个人给予安全保护使其免遭报复。

2.3 检验阶段

人工智能社会伦理规范对于不同行业领域而言需要依据其行业特色有针对性地实施，但总的说来，可以从各行业领域的项目进程角度提炼一套泛化的实施效果检验框架。该框架主要从项目基础层的支持、领导层的认同、指标绩效、组织影响等方面进行衡量举措，通过落后、基本、高级、卓越四个层次的表现具体验证其社会伦理规范实施程度，如表 2 所示。

表 2 人工智能伦理实施程度参照表

	落后	基本	高级	卓越
内部培训支持和人力资源	•基础层人员自觉寻找合适的人工智能伦理资源 •可能会受到鼓励，但没有官方支持 •更加注重法规遵从性	•团队成员所需的研讨会 •获得宣传者的支持 •专家审查组	•信息咨询组 •每个成果/解决方案的关键利益相关者和宣传者 •将特定用例添加至现有流程	•人工智能伦理嵌入决策实践各个环节中，而不是作为某一模块插入 •融入所有角色的职业培训
领导层认同	•领导层认可人工智能伦理，但不优先考虑	•完成入门级培训 •合规知识	•在新项目中纳入人工智能社会伦理 •集体协议中包含人工智能社会伦理 •领导层更新/了解团队努力	•伦理实践和观点融入到战略中 •激励符合伦理的行为并为不符何伦理的行为作出惩罚 •支持符合伦理的努力措施
指标和绩效	•一般说来，除了质量标准外，没有明确的人工智能伦理原则	•确定了基本的质量指标（人权评估、社会福利指标） •用户研究过程中实施的一些指标	•进一步制定和维护可被公众信任和理解的指标 •价值与成果符合伦理相关 •风险分类，优先考虑缓解措施 •有差异的隐私策略	•基于研究达到可信赖目标 •使用方不断反馈调整 •与工作者收入挂钩 •列出常见漏洞和问题清单
组织影响	•不会明显改变组织结构。	•团队实践与组织原则、问责制度相关联	•团队间密切合作以改进流程 •由上至下的理解和问责 •创造并支持讨论和建设性评论的文化氛围	•改变工作方向 •改变与使用者的关系 •改变工作者

3 建议

人类社会的科技水平与其依靠代代承继才拥有的文化背景存在发展不协调的问题，因此，需要通过在人工智能相关技术或项目中实施伦理规范，找到二者间的融合路径，实现为人类发展谋求福祉的根本诉求。根据目前从 IEEE 提炼的实施机制可总结出以下几点建议。

（1）构建文化氛围

人工智能技术从设计研发到推广使用需要社会各界的投入与支持。因此，需要不断在社会的科技创新发展进程中向民众构建和强化社会伦理的价值认同。有效的方法通常有赖于通识教育和社会宣传，这将会使得伦理规范始终贯穿各个学科领域、被各个项目环节自然遵循，为各行业从业者普遍认可和理解，也能避免社会伦理被作为单独的模块生硬嵌入人工智能技术相关的项目中。

（2）区分实施主体

在人工智能技术相关的项目中实施社会伦理规范对处于不同环节不同利益方的人员而言，是必然存在差异的。从技术设计者、开发者到推广者、使用者等，各有侧重倾向。因而在具体某一项目的伦理规范实施过程中，各方应提前明确其自身角色属性，划分好各自实施工作的范围和重点，并找到不同利益方之间合作互助的沟通方式。

（3）建立监督框架

对于人工智能技术相关项目的各个环节而言，应该始终处于给定的监督约束框架中开展实施工作。一般来说，监督方主要分项目内部的自我监督和第三方监督机构的外部监督，而监督内容应该包括实施规范的效力、实施规范者的能力、各利益方责任的分配、权限限定等。此外，有效的监督除了在国家层面或整个行业层面依靠对监督框架本身标准作明确定义外，还应将执行人员的专业素养纳入考量，以此保障监督制度的针对性和执行度。

总的来说，遵循人工智能社会伦理规范是现实社会运用人工智能技术的重要前提。人工智能技术为现实社会的发展带来了系列变革，从便捷性、安全性、公平性等诸多因素对其进行考量，利弊皆存，这就要求明确的社会伦理规范来约束各行业领域人工智能技术的应用。为此，需要明确人工智能社会伦理规范的具体实施机制。而以 IEEE 为代表，其人工智能社会伦理规范在一般项目开展中的实施机制，也以此为目标导向，因地制宜，因时制宜，在切合实际项目流程的基础上，强化了项目内部的组织管理与项目外部各方利益的协调，有力地服务现实社会各行业领域的项目开展与当代社会的智能化建设。

参 考 文 献

[1] Bostrom N. Superintelligence: Paths, dangers, strategies[M]. Oxford: Oxford University Press, 2014.

[2] Alzou'bi S, Alshibly H, Al-Ma'aitah M. Artificial intelligence in law enforcement, a review[J]. International Journal of Advanced Information Technology, 2014, 4(4):1-9.

[3] Helbing D. Big data society:Age of reputation or age of discrimination?[A]. Helbing D. Thinking ahead-essays on big data,digital revolution, and participatory market society[M]. Cham: Springer, 2015.

[4] Lindsay R K. Complexities of the mind at work.(Book Reviews:What computers can't do. A critique of arti-

doi:10.3772/j.issn.2095-915x.2021.04.009

ficial reason)[J]. Science, 1972, 176(4035):630-631.
[5] 刘斌 . 人工智能时代教师的智能教育素养探究 [J]. 现代教育技术 , 2020, 30(11):12-18.
[6] 雷晓维 . 我国人工智能与教育深度融合路径探析 [J]. 软件导刊 , 2020, 19(10):84-87.
[7] 张庆龙 . 智能财务七大理论问题论 [J]. 财会月刊 , 2021(1):23-29.
[8] 罗洪洋 , 李相龙 . 智能司法中的伦理问题及其应对 [J]. 政法论丛 ,2021(1):148-160.
[9] 李伦 , 孙保学 . 给人工智能一颗“良芯(良心)”——人工智能伦理研究的四个维度 [J]. 教学与研究 , 2018(8):72-79.
[10] 刘伟 , 赵路 . 对人工智能若干伦理问题的思考 [J]. 科学与社会 , 2018, 8(1):40-48.
[11] 张世龙 , 缪军翔 . 科技革命时代科技创新与伦理的冲突及其协调机制研究 [C]. 第十四届（2019）中国管理学年会论文集 . 中国管理现代化研究会、复旦管理学奖励基金会 : 中国管理现代化研究会 ,2019:277-282.
[12] 闫坤如 , 马少卿 . 人工智能伦理问题及其规约之径 [J]. 东北大学学报 (社会科学版), 2018, 20(4):331-336.
[13] 段伟文 . 人工智能时代的价值审度与伦理调适 [J]. 中国人民大学学报 , 2017, 31(6):98-108.
[14] 郭晓斐 , 赵平 , 高翠巧 . 医疗人工智能发展面临的法律与伦理挑战及对策研究 [J]. 中国肿瘤 , 2019, 28(7):509-512.
[15] Lee C K. How the technology of autonomous driving affects the scope and level of driver's duty of care and the necessity for embedding ethical ability in autonomous vehicles[J]. Journal of hongik law review, 2016, 17(4):443-472.
[16] Christoph L, Poszler F, Acosta A J, et al. AI4People:Ethical Guidelines for the Automotive Sector – Fundamental Requirements and Practical Recommendations[J]. International Journal of Technoethics, 2021, 12(1):101-125.
[17] Canales C, Lee C, Cannesson M. Science Without Conscience Is but the Ruin of the Soul:The Ethics of Big Data and Artificial Intelligence in Perioperative Medicine[J]. Anesthesia and analgesia, 2020, 130(5):1234-1243.
[18] Stokes F, Palmer A . Artificial Intelligence and Robotics in Nursing:Ethics of Caring as a Guide to Dividing Tasks Between AI and Humans[J]. Nursing Philosophy, 2020, 21(4):e12306.
[19] Baric-Parker J, Anderson E E. Patient Data Sharing for AI:Ethical Challenges, Catholic Solutions[J]. The Linacre quarterly, 2020, 87(4):471-481.
[20] Mcaleenan P . Moral Responsibility and Action in the Use of Artificial Intelligence in Construction[J]. Management Procurement and Law, 2020, 173(4):1-8.
[21] Floridi L, Cowls J, Beltrametti M, et al. AI4People - An Ethical Framework for a Good AI Society:Opportunities, Risks, Principles, and Recommendations[J]. Social Science Electronic Publishing. 2018,28(4):689-707.
[22] IEEE standards. General principles. [EB/OL]. [2020-12-19]. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_general_principles.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined.
[23] IEEE standards. Ead for business. [EB/OL]. [2020-12-19]. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead/ead-for-business.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined.
[24] IEEE standards. Policy of ead1e. [EB/OL]. [2020-12-19]. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_policy.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined.
[25] IEEE Standards. Law of ead1e. [EB/OL]. [2020-12-19]. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_law.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined.