人工智能伦理风险分析报告

国家人工智能标准化总体组

二零一九年四月

专家咨询组

潘云鹤　高　文　戴　红　谭铁牛　吴朝晖　李伯虎
林　宁　于海斌　吴　飞　周志华　董景辰　黄河燕
朱小燕　张德政　朱恺真　曲道奎　左　毅　钱　恒

国家人工智能标准化总体组

组　长：赵　波
副组长：闵万里　胡国平　徐　波
　　黄铁军　吴文峻　欧阳劲松
秘书长：孙文龙

编写单位（排名不分先后）

中国电子技术标准化研究院
中国人民大学
北京理工大学
深圳市腾讯计算机系统有限公司
北京航空航天大学
大成律师事务所
京东数字科技控股有限公司
昆山炫生活信息技术股份有限公司
美国科文顿柏灵律师事务所
美团点评
北京九天微星科技发展有限公司
国家电网有限公司
苏州中德宏泰电子科技股份有限公司
中国航空综合技术研究所
上海电器科学研究所（集团）有限公司
北京爱奇艺科技有限公司
北京西普阳光教育科技股份有限公司
华为技术有限公司
中国船舶重工集团公司第七一六研究所
西门子（中国）有限公司

编写人员（排名不分先后）

郭　锐　李慧芳　曹建峰　代　红　董　建　张　群　王燕妮　汪小娟
马珊珊　伍敏敏　赵泽睿　李　依　朱婷婷　刘雅洁　刘馨泽　赵蕾蕾
张钦坤　蔡雄山　周学峰　丁海俊　张凌寒　蔡克蒙　魏　铭　刘元兴
杨银剑　陈隋和　罗　嫣　于智精　罗治兵　李　燕　吴　鹏　高　畅
宁　皓　胡　鹰　康　丹　付振秋　何宇聪　邢　琳　王　涛　赵弋洋
王　莲　王小璞　高　伟　符海芳　何佳洲　赵　蕾　张　珺

目录

第一章 人工智能伦理研究的背景与意义

自 1956 年的达特茅斯会议（Dartmouth Conference）提出人工智能概念以来，人工智能的发展经历了“三起两落”的曲折历程。2016 年 3 月，以 AlphaGo 以 4:1 战胜人类棋手为标志，人工智能开始逐步升温，并成为各国政府、科研机构、产业界以及消费市场竞相追逐的对象。为了在新一轮国际竞争中掌握主导权，抢占人工智能发展的制高点，各国投入大量的精力和资金，开展人工智能关键技术的攻关与应用相关的研究与产品开发，并纷纷推出了不同的人工智能平台与产品。

我国人工智能的应用范围极广。从行业应用的角度看，在制造、物流、医疗、教育、安防等行业都有广泛应用。以制造业为例，当前的制造业不论是生产、流通还是销售，都正趋于数据化、智能化。大数据和人工智能技术可以协助企业分析生产过程中的全链路数据，实现生产效率、库存周转率、设备使用效率提升等目标。在智能制造进程中，工业机器人成为人工智能的典型代表，成为智能制造的重要实现端之一。就物流行业而言，人工智能的技术应用主要聚焦在智能搜索、推理规划、模式识别、计算机视觉以及智能机器人等领域。如今，现代物流企业纷纷尝试利用人工智能技术优化物流环节、提高物流效率。人工智能还能够帮助企业根据市场销售情况、供应链生产情况、物流配送、仓储库存水平，甚至每个环节的容错概率等等进行精准排产，最大限度利用已有资源。人工智能在医疗健康主要的应用领域则包括五个方面：临床决策支持、临床辅助诊疗系统、患者管理、辅助手术和患者照护的自动设备，即各种机器人、医疗机构的管理以及新药的研发。

人工智能在自动驾驶、医疗、传媒、金融、工业机器人以及互联网服务等越来越多领域和场景下得到应用，一方面带来了效率的提升、成本的降低，，另一方面，人工智能系统的自主性使算法决策逐步替代了人类决策，而这种替代有时非但没有解决已有的问题，还让已有的问题更难解决，甚至给社会带来了全新的问题。这些问题不仅仅引发社会的广泛讨论，更是限制人工智能技术落地的重要因素。其中最为典型的便是自动驾驶领域，社会的巨大需求与技术的不断成熟让

自动驾驶成为了全球炙手可热的研究与发展领域，而其潜在的风险又驱使人们去反思技术带来的伦理问题。各国已有法律与政策的难以适用以及新政策的模糊不清也给自动驾驶技术的落地造成了困难。面对伦理风险与其潜能一样巨大的人工智能技术，人们急需一个广泛、普遍的伦理探讨，并在这些探讨的基础之上找到路径、梳理规范，以保证人工智能的良性发展

目前，各国、各行业组织、社会团体和人工智能领域的商业公司纷纷提出人工智能的伦理准则，对人工智能技术本身以及其应用进行规制。中国政府把人工智能作为产业升级和经济转型的主要驱动力，鼓励、扶持并推动人工智能的发展。在我国推动人工智能发展的关键时期，推动对人工智能伦理和社会问题的探讨有极为重要的意义。因此，本报告以人工智能应用引发的社会伦理问题为出发点，在充分了解人工智能系统带来的伦理、法律和社会影响的基础上，分析人工智能应用即自主性决策结果而产生的社会公平、安全及问责等伦理道德问题，例如算法决策与歧视、隐私与数据保护、算法安全与责任、算法解释、算法与产权、算法与竞争、算法滥用以及强人工智能问题，通过遵循人工智能伦理原则与设计相应的风险指标体系，对人工智能的研发和应用提供风险管理指引，以便为人工智能伦理的行业实践提供初步的应用指南与建议，推动人工智能产业的良性、健康发展。

第二章 国内外人工智能伦理发展现状

2.1 国外发展现状

步入第三次人工智能发展浪潮以来，人工智能成为国际竞争的新焦点，各国各地区高度重视人工智能的发展，纷纷出台战略文件支持，促进人工智能的发展，确保人工智能对经济和社会产生积极影响，并造福于个人和社会。人工智能伦理由此成为各国人工智能政策的核心内容之一。

以欧盟为例，其以多举措推进人工智能伦理立法。早在 2015 年 1 月，欧盟议会法律事务委员会（JURI）就决定成立一个工作小组，专门研究与机器人和人工智能发展相关的法律问题。2016 年 5 月，法律事务委员会发布《就机器人民事法律规则向欧盟委员会提出立法建议的报告草案》（Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics，简称《报告草案》）；同年 10 月，发布研究成果《欧盟机器人民事法律规则》（European Civil Law Rules in Robotics）[1]。此后，欧盟委员会将“人工智能和机器人的伦理标准等”纳入 2018 年欧盟立法工作的重点，要在人工智能和机器人领域呼吁高水平的数据保护、数字权利和道德标准，并成立了人工智能工作小组，就人工智能的发展和技术引发的道德问题制定指导方针。在此背景下，2018 年 3 月，欧洲科学与新技术伦理组织（European Group on Ethics in Science and New Technologies）发布《关于人工智能、机器人及“自主”系统的声明》（Statement on Artificial Intelligence，Robotics and ‘Autonomous’Systems）[2]，呼吁为人工智能、机器人和“自主”系统的设计、生产、使用和治理制定共同的、国际公认的道德和法律框架。

而在近日，欧盟委员会人工智能高级专家组（The European Commission’s High-Level Expert Group on Artificial Intelligence）于 2018 年 12 月 18 日发布了《关于可信赖人工智能的伦理准则》（草案）（“Ethics Guidelines for Trustworthy AI”）

[1] 曹建峰.十项建议解读欧盟人工智能立法新趋势[EB/OL].[2019-03-25]http://www.tisi.org/4811.

[2] 欧洲科学与新技术伦理组织.关于人工智能、机器人及自主系统的声明[R].曹建峰译.北京:腾讯研究院, 2018.

[3]。在这个草案中，欧盟尝试提出了一个框架，确保在开发、推广或应用人工智能的过程中，研发者能尊重基本权利、原则及价值。具有人权传统的欧盟秉持以人为本的人工智能发展理念，希望通过人工智能价值引导人工智能发展，塑造其社会影响，造福个人和社会。这一系列举措表明了欧盟通过人工智能伦理规制、约束人工智能向着有益于个人和社会发展的决心。在技术和产业不占优势的情况下，欧盟人工智能战略的重头戏放在了人工智能价值观，希望以此彰显欧盟发展人工智能的独特优势。

其他国家亦在积极推进人工智能伦理。美国特朗普总统于 2019 年 2 月 11 日签署了一项行政命令，正式启动“美国人工智能计划”，以刺激推动美国在人工智能领域的投入和发展，这其中就包括了道德标准的要求，像白宫科技政策办公室（OSTP）和美国国家标准与技术研究院（NIST）这样的政府机构将被要求制定标准，指导“可靠、稳健、可信、安全、可移植和可互操作的人工智能系统”的开发。

英国政府曾在其发布的多份人工智能报告中提出应对人工智能的法律、伦理和社会影响，最为显著的是英国议会于 2018 年 4 月发出的长达 180 页的报告《英国人工智能发展的计划、能力与志向》（AI in the UK: ready, willing and able？）[4]。该报告认为当前不需要对人工智能进行专门监管，各个行业的监管机构完全可以根据实际情况对监管做出适应性调整。相反，该报告呼吁英国政府制定国家层面的人工智能准则（AI Code），为人工智能研发和利用设定基本的伦理原则，并探索相关标准和最佳实践等，以便实现行业自律。

此外，联合国教育、科学及文化组织（United Nations Educational, Scientific and Cultural Organization，简称 UNESCO）的下设委员会（the World Commission on the Ethics of Scientific Knowledge and Technology of UNESCO，简称 COMEST）历时两年完成并于 2017 年 9 月发布《机器人伦理报告》（Report of COMEST on robotics ethics）[5]，建议制定国家和国际层面的伦理准则。

全球最大的专业学术组织电气和电子工程师协会（Institute of Electrical and Electronics Engineers，IEEE）于 2016 年启动“关于自主/智能系统伦理的全球倡

---

[3] 欧盟委员会人工智能高级专家组.关于可信赖人工智能的伦理准则[J].青年记者,2019(01):88.

[4] AI in the UK:ready, willing and able, https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf, p.38.

[5] UNESCO (2017) .Preliminary draft report of comest on robotics ethics[EB/OL].[2018-05-12].http://unesdoc.unesco.org/images/0024/002455/245532E.pdf.

议”，其发布的“人工智能设计的伦理准则”白皮书目前已迭代到第二版，第三版希望融入更多来自亚洲地区的声音。此外，2017 年美国的未来生命研究所（future of life institute，FLI）主持达成了 23 条人工智能原则，近四千名各界专家签署支持这些原则，在业界引起了较大反响。美国计算机协会（Association for Computing Machinery, 简称 ACM）提出的算法透明和可责性七项原则、负责任地发展人工智能的蒙特利尔宣言、UNI Global Union 提出的人工智能十大原则等也代表着类似的人工智能伦理思考。

当然，面对社会公众的担忧和质疑，科技公司并非无动于衷，早已开始思考人工智能技术及其应用的社会经济和人类健康影响，并采取措施确保人工智能有益于、造福于个人和社会，最终目的是希望人工智能在人类社会中扮演一个积极的角色，而非沦为破坏者。科技公司的举措主要体现在以下三个方面：

其一，积极拥抱人工智能伦理与社会研究。如今，人工智能领域的一个重要的风向标就是，人工智能研究不仅关乎技术，更关乎技术的伦理与社会影响。跨学科、跨领域的研究在人工智能领域正成为一个显著的趋势。以 DeepMind 公司为例，其于 2017 年 10 月 3 日宣布成立人工智能伦理与社会部门（DeepMind Ethics & Society），目的就在于补充、配合其人工智能研发和应用活动。

其二，提出人工智能价值观。面对人工智能引发的一些负面问题，谷歌、微软等科技公司纷纷提出企业层面的人工智能价值观以赢得公众和公众的信任。比如，微软、谷歌、IBM、Sage、SAP 等科技企业均提出了自己的人工智能原则。

其三，成立人工智能伦理委员会。早在谷歌收购 DeepMind 之时，其就承诺建立一个人工智能伦理委员会 AETHER（AI and Ethics in Engineering and Research Committee）。微软在其《计算化未来：人工智能及其社会角色》一书中透露其已经成立了一个人工智能伦理委员会 AETHER（AI and Ethics in Engineering and Research Committee），确保将其奉行的人工智能原则融入人工智能研发和应用。据其说法，这个委员会作为微软的内部机构，囊括了工程、研究、咨询、法律等部门的专家，旨在积极推动形成内部政策并应对潜在的问题。该 AETHER 委员会的主要职责是，制定能够作用于微软人工智能产品和方案之研发和应用的最佳实践和指导原则，帮助解决从其人工智能研究、产品和用户互动中产生的伦理和社会问题。此外，今年 5 月，在经历了引起轩然大波的数据泄

露丑闻之后，Facebook 宣布成立人工智能伦理团队，负责防止其人工智能软件中的歧视。从其官网可以发现，Facebook 正在招聘人工智能政策、伦理、法律等方面的人员，表明其开始重视人工智能伦理相关的工作。

凡此种种表明，在人工智能技术研发和应用之外，人工智能伦理已经成为科技公司的主要关切之一。一方面在加强人工智能伦理相关的研究，另一方面通过人工智能伦理委员会对人工智能技术、产品和应用形成伦理约束。与此同时通过对外传递其人工智能价值观树立产业良好形象。

2.2 国内发展现状

我国已经将人工智能伦理提上日程。2017 年发布的《新一代人工智能发展规划》[6]提出了中国的人工智能战略，制定促进人工智能发展的法律法规和伦理规范作为重要的保证措施被提了出来。在该战略文件中，伦理一词出现 15 次之多，足见我国对人工智能伦理的高度重视。其释放的信号是，不仅要重视人工智能的社会伦理影响，而且要制定伦理框架和伦理规范，以确保人工智能安全、可靠、可控发展。

2018 年 1 月 18 日，在国家人工智能标准化总体组的成立大会上，《人工智能标准化白皮书 2018》[7]正式发布。白皮书论述了人工智能的安全、伦理和隐私问题，认为设定人工智能技术的伦理要求，要依托于社会和公众对人工智能伦理的深入思考和广泛共识，并遵循一些共识原则。

2018 年 9 月 17 日，国家主席习近平致信祝贺 2018 世界人工智能大会在沪召开。习近平在贺信中指出，新一代人工智能正在全球范围内蓬勃兴起，为经济社会发展注入了新动能，正在深刻改变人们的生产生活方式。把握好这一发展机遇，处理好人工智能在法律、安全、就业、道德伦理和政府治理等方面提出的新课题，需要各国深化合作、共同探讨。

在业内，腾讯研究院等研究机构亦非常重视人工智能伦理研究。以腾讯研究院为例，其在国内较早开始关注并研究人工智能伦理和社会问题，率先将欧盟、联合国、IEEE 等出台的人工智能伦理政策文件和报告译介到国内，并就算法歧

[6] 国务院.新一代人工智能发展规划 (国发[2017]35 号) [EB/OL]. (2017-07-20) [2019-03-25]. http://www.csjrw.cn/2017/0720/58931.shtml.

[7] 中国电子科技标准化研究院.人工智能标准化白皮书[EB/OL].[2019-03-25] http://www.cesi.ac.cn/201801/3545.html.

视、人工智能安全、人工智能隐私、人工智能就业影响等输出了很多研究成果，并与中国信通院互联网法律研究中心、腾讯 AI 实验室、腾讯开放平台一起出版了《人工智能：国家人工智能战略行动抓手》[8]一书，书中战略篇、法律篇、伦理篇以及治理篇着重讨论人工智能的社会和伦理影响，呼吁为人工智能的发展和应用制定共同的伦理框架。

当前，国内外各界都非常重视人工智能伦理和社会影响研究，希望在发展和应用人工智能这一新技术造福社会和人类的同时，也能意识到其可能带来的负面影响和伦理问题，确保人工智能安全、可靠、可控发展。

[8] 腾讯研究院,中国信息通信研究院互联网法律研究中心,腾讯 AILab,腾讯开放平台.人工智能：国家人工智能战略行动抓手[M].北京:中国人民大学出版社,2017.

第三章 人工智能技术的伦理风险

人工智能技术的开发和应用深刻地改变着人类的生活，不可避免地会冲击现有的伦理与社会秩序，引发一系列问题。这其中，既有直观的短期风险，如算法漏洞存在安全隐患、算法偏见导致歧视性政策的制定等，也有相对间接的长期风险，如对产权、竞争、就业甚至社会结构的影响。尽管短期风险更具体可感，但长期风险所带来的社会影响更为广泛而深远，同样应予重视。

长远来看，人工智能应用的伦理风险具有独特性。其一，与个人切身利益密切相关，如将算法应用在犯罪评估、信用贷款、雇佣评估等关切人身利益的场合，一旦产生歧视，必将系统性地危害个人权益。其二，引发算法歧视的原因通常难以确定，深度学习是一个典型的“黑箱”算法，连设计者可能都不知道算法如何决策，要在系统中发现有没有存在歧视和歧视根源，在技术上是比较困难的。其三，人工智能在企业决策中的应用愈发广泛，而资本的逐利本性更容易导致公众权益受到侵害。例如，企业可以基于用户行为数据分析实现对客户的价格歧视，或者利用人工智能有针对性地向用户投放游戏、瘾品甚至虚假交友网站的广告，从中获取巨大利益等。

根据风险产生的方式，下文从算法、数据和应用三个维度梳理人工智能伦理风险的具体性质与特征。此外，对就业、市场竞争秩序、产权等法律制度的挑战等人工智能远期发展带来的潜在伦理风险，我们将其归入之中长期和间接的伦理风险中。

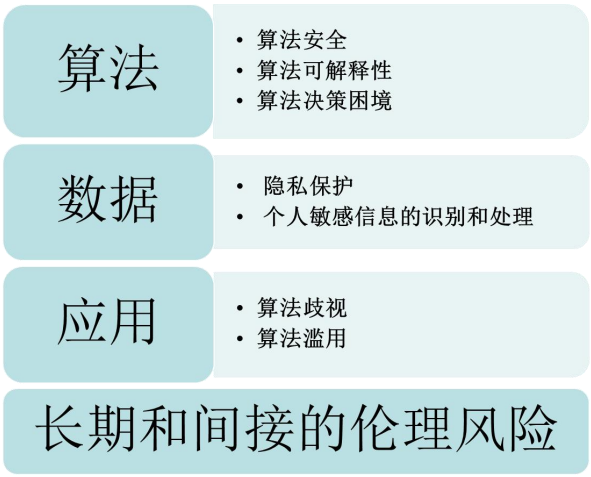

图 1 人工智能风险的三个维度

3.1 算法相关的伦理风险

3.1.1 算法安全

（1）算法安全风险产生的原因

其一，算法存在泄露风险。算法需要模型参数，并且在训练数据中运行该模型参数。如果算法的模型参数被泄露，第三方则有可能能够复制该算法。

其二，算法从设计、训练到使用均面临可信赖性问题。一方面，算法的训练数据不能完全覆盖应用场景的所有情况。另一方面，和传统的算法相比，人工智能算法在原理上是用于处理步骤不明确，输入较不受限的场景，并且允许错误率存在一定的弹性。基于人工智能算法的以上特性，如果在使用之前，算法的参数被非法修改，或者在使用过程中被攻击者通过恶意样本修改，那么算法性能的下降或错误率的升高将较难被觉察到。

其三，部分场景下的算法对随时可用的要求较高。算法的可用性在许多关键的场景中非常重要。例如，无人驾驶汽车作为网络中的一个节点，有可能受到外部的网络攻击。当这种情况发生时，负责自动驾驶的算法模块必须保持可用，以控制汽车的安全行进和停止。

其四，人工智能算法在许多场景的应用都与人身安全息息相关，如在医疗领域、自动驾驶领域等。这些领域的算法应用一旦出现漏洞，将直接侵害人身权益，后果难以挽回。

（2）算法安全风险带来的影响

首先，算法泄露可能给算法的所有者和用户造成损失。一方面，算法泄露后，第三方可在不支付获取数据成本的情况下为用户提供价格更低的产品，这将给算法的所有者造成商业损失。另一方面，训练数据在很多情况下包含用户的个人数据，丢失这样的数据会将用户置于危险之中，例如，这些信息被第三方用于网络欺诈、勒索等等。若发生以上风险，用户将向算法训练数据的控制者要求索赔，从而算法数据的控制者将承担相应的法律责任。

其次，算法随时可用的要求对其可靠性带来挑战。某无人驾驶汽车能够识别鹿并进行躲避，但是该系统到澳大利亚进行测试时，由于袋鼠独特的跳动前进模

式，系统无法识别出袋鼠从而无法进行躲避。也就是说，人工智能算法相对于只能按照明确步骤执行的算法，具有一定的智能，但同时其智能也具有局限性，而攻击者则有可能利用这样的局限性进行攻击。例如，在无人驾驶系统中，对图片加入受控的噪声以使识别结果错误。

再次，算法和运行系统可能直接或间接地引发人身伤害，并引发一系列法律追责困境。以医疗为例，当医生使用智能辅助软硬件系统而发生医疗纠纷时，如何界定权责便成为法律面临的难题。从更长远的角度看，人工智能影像识别系统的应用虽然能减轻医师读片（阅读 CT、X 光片、核磁等图像并作出诊断的能力）的压力，但也会出现相应的问题。若大规模普及这种影像识别系统，医生群体整体读片的技能将逐渐下降，并将导致医生对人工智能的识别结果失去鉴别力。

（3）算法安全风险的应对

针对算法漏洞带来的安全风险，需要加强算法保密性，如通过加密等相应的安全防护措施确保算法不被轻易泄露。

针对算法的可信赖性风险，需要通过传统的安全防护措施防范算法参数被非法修改的可能性，例如将修改权限限于已获授权的特定用户。此外，还需要从算法的原理出发对算法的整体设计进行改进。

针对算法随时可用要求对其带来的可靠性挑战，人工智能算法不仅要考虑正常的算法输入，也要考虑异常的输入，并且保证系统在异常输入时仍然保持其可用性，如设立应急系统，或通过运行足够多的测试来降低异常情况发生的可能性。

针对算法运行可能造成人身伤害的风险，应在医疗等攸关人身安全的领域明确风险提示要求，并选择那些稳定性高且原理可解释的算法，以保证算法运行的安全和可追责性。此外，还需加强系统的可测试性，例如使用人工智能的医疗器械，不仅必须通过通常的器械测试，并且还需通过对可能的算法特征引起的风险的针对性测试。

3.1.2 算法可解释性

（1）算法可解释性的定义

依据 2017 年美国加州大学伯克利分校发布的《对人工智能系统挑战的伯克

利观点》，可以将算法的可解释性理解为“解释人工智能算法输入的某些特性引起的某个特定输出结果的原因”。Miller 等人在其 2017 年的综述中将“可解释性”定义为“展示自己或其它主题做出的决定所依赖的原因”[9]。但因为算法的原因和我们日常生活中的原因可能很不一样，所以可解释性的概念还需要在算法环境下进行理解和界定。2017 年 ICML 的 Tutorial 中给出的一个关于可解释性的定义是：“解释是给人类作出解释的过程”[10]，引申来说就是以人类能理解的描述给出解释，以让人类能看懂。

算法解释按照解释的内容划分可以分为过程解释和决策解释[11]。按照路径来划分，可以分为模型中心解释和主体中心解释。模型中心解释，注重算法模型、逻辑过程、数据信息的全局解释，而主体中心解释侧重建立在输入记录的基础上的局部解释，不苛求进入“黑箱内部”，更多是在算法决策（过程和结果）和算法主体（设计者、使用者和消费者）之间建立关系，从而提供出学者在研究欧盟 GDPR 时所提出的“有意义的（meaningful）解释”[12]。而反设事实解释，并不试图解释黑盒算法的内部逻辑而是提供了关于外部依赖因素的解释，即无需“打开黑盒”，其可以通过逻辑推演仅对局部（如两端：假设条件和得出结果）进行解释，而无需对算法模型和过程机制进行解释。主体中心解释和反设事实解释减轻了企业解释的成本负担，并能够为算法消费者提供更为有效的救济以及“有意义的解释”。

（2）算法可解释性安全风险产生的原因

算法之所以难以解释，是因为“黑箱”现象的存在。数据公司 Teradata 首席技术官斯蒂芬·布罗布斯特（Stephen Brobst）认为，机器学习基本就是线性数学，很好解释，但是一旦涉及多层神经网络，问题就变成了非线性数学，不同变量之间的关系就纠缠不清了。此外，因为人工智能算法的两个复杂性特质：涌现性和

[9] Miller, Tim, Piers Howe, and Liz Sonenberg. *Explainable AI: Beware of inmates running the asylum,* IJCA-17 Workshop on Explainable AI (XAI),2017.

[10] “Interpretation is the process of giving explanations to Human”.

[11] 美国计算机协会美国公共政策委员会 2017 年初发布的《算法透明性和可问责性声明》提出要对算法的过程和特定的决策提供解释。

[12] Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a "Right to an Explanation" is probably not the remedy you are looking for*, 16 Duke L. & Tech. Rev. 18 (2017).

自主性，使其比较难以理解和解释，从而导致“黑箱”现象[13]。

（3）算法可解释性安全风险的影响

算法可解释性和透明性是一个重要的人工智能伦理命题，因为其关涉人类的知情利益和主体地位。人类对算法的安全感、信赖感、认同度取决于算法的透明性和可理解性。算法的复杂性和专业性，使得信息不对称更加严重，且这种不对称的加重不只发生在算法消费者与算法设计者、使用者之间，更发生在人类和机器之间，所以算法应用下的“人类知情利益保障”是一个比较棘手的问题。

另外，人工智能算法的两个复杂性特质——涌现性和自主性，导致理性原则失效，难以通过行为原则判断和道德代码嵌入来保证算法的“善”[14]，这将会给社会带来伦理难题。如在智能信贷领域，智能金融算法可能会产生“降低弱势群体的信贷得分”[15]，“拒绝向‘有色人种’贷款”[16]，“广告商更倾向于将高息贷款信息向低收入群体展示”[17]等歧视性决策。由于做出这种决策是不透明的，但该决策对于用户而言意义重大，因此受算法决策影响的用户应该得到有关“解释”，包括算法的功能和通用的逻辑、算法的目的和意义、设想的后果、具体决定的逻辑和个人数据的权重等[18]，否则将会产生严重的伦理问题。

典型如金融领域，人工智能已广泛应用于金融风险评估等关键环节。向金融消费者提供人工智能算法的合理解释，是金融消费者保护的题中之义。提供算法解释的目的是使金融消费者了解对其不利的决定是如何做出的，以便在上述不利决定违反法律的情况下提供救济。对弱势群体作出拒绝贷款的决策，应该特别引起注意。如果这种决策是基于人工智能算法作出的，应该向算法决策影响到的个

---

[13] 参见刘劲杨：《人工智能算法的复杂性特质及伦理挑战》，载《光明日报》，2017 年 9 月 4 日 15 版。

[14] 参见刘劲杨：《人工智能算法的复杂性特质及伦理挑战》，载《光明日报》， 2017 年 9 月 4 日 15 版。

[15] *How Algorithms Can Bring Down Minorities' Credit Scores,* available at https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/, last visited：June 13, 2018。

[16] *Did Artificial Intelligence Deny You Credit?* available at http://fortune.com/2017/03/19/artificial-intelligence-credit/, last visited：June 13, 2018。

[17] *When Algorithms Discriminate - The New York Times,* available at https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html, last visited：June 13, 2018。

[18] 张凌寒：《商业自动化决策的算法解释权研究》，载《法律科学(西北政法大学学报) 》，2018 第 3 期，第 72 页。

人“解释”算法的功能和通用的逻辑、算法的目的和意义、设想的后果、具体决定的逻辑和个人数据的权重等，否则将会违反金融消费者保护的基本原则。

简言之，算法可解释性的目的包括维护算法消费者的知情权利益，避免和解决算法决策的错误性和歧视性，明晰算法决策的主体性、因果性或相关性，进而助力解决算法可问责性问题。

（4）算法可解释性安全风险的应对

算法可解释性问题已引起国际官方和研究机构的关注。例如，电气和电子工程师协会（IEEE）在 2016 年和 2017 年连续推出的《人工智能设计的伦理准则》白皮书，在多个部分都提出了对人工智能和自动化系统应有解释能力的要求。美国计算机协会美国公共政策委员会在 2017 年初发布了《算法透明性和可问责性声明》，提出了七项基本原则，其中一项即为“解释”，希望鼓励使用算法决策的系统和机构，对算法的过程和特定的决策提供解释。2017 年，美国加州大学伯克利分校发布了《对人工智能系统挑战的伯克利观点》，从人工智能的发展趋势出发，总结了九项挑战和研究方向。其中之一，即第三项，就是要发展可解释的决策，使人们可以识别人工智能算法输入的哪些特性引起了某个特定的输出结果。在我国国务院《新一代人工智能发展规划》中，潘云鹤院士提到人工智能应用的一个需要关注的问题是算法的不可解释性[19]。因此，国家人工智能标准体系可通过操作标准、伦理标准以及数据模型传递与解释标准等的制定加强对算法可解释性的要求。《欧洲通用数据保护条例》(GDPR)第 71 条明确提出了解释权，表述为被自动决策的人应该具有适当的保护，具体应包括数据主体的特别信息和获得人类干预、表达自己的观点，并且有权获得该评估决定的解释，并对决定提出质疑。

对算法解释来说，或许探索“相关关系”而非“因果关系”，才是解决之道。例如舍恩伯格等在《大数据时代:生活、工作与思维的大变革》一书中就提出，在大数据技术引发思维变革背景下，应更为关注事物之间的相关关系，而不是探索因果关系。所以，舍弃解释因果关系，进而从解释相关关系的需求突破，可能是兼顾算法消费者利益和减轻企业解释成本负担的有效路径。过度公开透明或强

[19] 潘云鹤：《人工智能迈向 2.0》，英文版发表于中国工程院院刊《Engineering》，http://news.sciencenet.cn/htmlnews/2017/1/365934.shtm，最后访问时间：2018 年 7 月 6 日。

化可解释性，不利于技术创新和社会进步，也不利于增进社会福利[20]。此外，算法可解释性和算法黑箱问题也可以通过“技术手段”得到部分解决。借助技术使得“机器理解层次”降维到“人类理解层次”，也即人类能看懂、能理解。例如2018 年 3 月 7 日，谷歌大脑团队的克里斯·欧拉（Chris Olah）公布了一项题为“可解释性的基础构件”的研究成果，该成果解决了神经网络这种最令人难以捉摸的算法的可视化问题，简化了相关信息，使算法的工作状态回到了“人类尺度”，能够被普通人看懂和理解[21]。

3.1.3 算法决策困境

（1）算法决策风险产生的原因

算法决策的困境主要表现在算法结果的不可预见性。随着计算能力的不断攀升，人工智能可以计算大量的可能性，其选择空间往往大于人类，它们能够轻易地去尝试人类以前从未考虑的解决方案。换言之，尽管人们设计了某人工智能产品，但受限于人类自身的认知能力，研发者无法预见其所研发的智能产品做出的决策以及产生的效果。以谷歌 DeepMind 团队开发的 AlphaGo 与多位人类围棋高手的“人机大战”为例，AlphaGo 在 2016 年 3 月对阵李世石时为第 18 代（AlphaGo Lee），在 2017 年 5 月对阵柯洁时已经迭代为第 60 代（AlphaGo Master）。而在 2017 年 10 月，谷歌 DeepMind 开发的 AlphaGo Zero 机器系统仅训练 3 天就以 100:0 的比分战胜了 AlphaGo Lee；经过 40 天训练后，AlphaGo Zero 又以 89:11 战胜了横扫柯洁的 AlphaGo Master。快速迭代的背后是 AlphaGo 全新的深度学习逻辑，这种经历迭代的深度学习逻辑，其强大的进化速度让人类难以追赶。

（2）算法决策风险的应对

既然算法决策的困境主要源于人工智能自学习能力导致的算法结果的不可预见性，为此要减少或杜绝算法决策困境，除了提高算法的可解释性外，还可以引入相应的算法终结机制，以便算法决策遇到无法判断未来结果时立即终止系

[20] Joshua New & Daniel Casto, *how policymakers can foster algorithmic accountability*, available at https://www.datainnovation.org/2018/05/how-policymakers-can-foster-algorithmic-accountability/, last visited:July 6, 2018.

[21] 郑戈：《算法的法律与法律的算法》，载《中国法律评论》2018 年第 2 期。

统。

人工智能最大的威胁是当前人类尚难以理解其决策行为所存在的未来失控的风险，而一旦失控则后果严重。参照所有生命体中都有的衰老机制，人工智能也应该嵌入自我毁灭机制。谷歌旗下 DeepMind 公司在 2016 年曾提出要给人工智能系统安装“切断开关（kill switch）”的想法，为的是阻止人工智能学会如何阻止人类对某项活动(比方说发射核武器)的干预，这种提法被称作“安全可中断性”。据介绍，安全可中断性可用于控制机器人不端甚至可能导致不可逆后果的行为，相当于在其内部强制加入某种自我终结机制，一旦常规监管手段失效，还能够触发其自我终结机制，从而使其始终处于人们的监管范围之内。

3.2 数据相关的伦理风险

随着数据搜集、机器学习、人工智能等技术的使用，数据富含越来越大的价值，从而也导致个人信息泄露的情况频繁发生。个人隐私保护、个人敏感信息识别的重要性日益凸现。为了保护数据主体的权益，2018 年 5 月 25 日，欧盟《一般数据保护条例》（GDPR）正式生效，增加了数据主体的被遗忘权和删除权，引入了强制数据泄露通告、专设数据保护官员等条款，同时包含了更严厉的违规处罚。在 2018 年 10 月，第 40 届数据保护与隐私专员国际大会（ICDPPC）通过了由法国国家信息与自由委员会、欧洲数据保护专员和意大利数据保护专员提出的《人工智能伦理与数据保护宣言》(Declaration on Ethics and Data Protection in Artifical Intelligence)，该宣言也提出了包括保护隐私原则在内的六项原则。

3.2.1 隐私保护

个人信息的隐私权是信任和个人自由的根本，同时也是人工智能时代维持文明与尊严的基本方式。早期，由于技术有限，数据获取成本高、回报低，导致大部分的个人隐私泄露的安全事件都是以“点对点”的形式发生，即以黑客为主的组织利用电脑木马等技术对个别用户进行侵害，从而在这些个别用户身上获利。

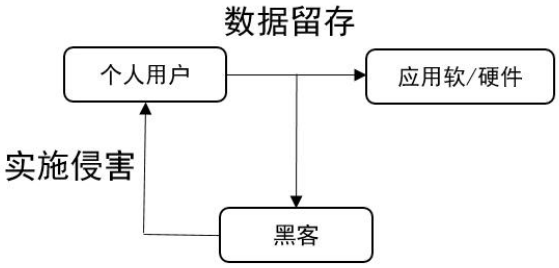

图 2 非大数据时代下的个人信息泄露流程

随着大数据和人工智能的发展，数据挖掘的深度与广度的不断加深，人工智能技术与用户隐私保护出现的紧张关系愈加严重。不法分子获取个人隐私数据的方式更多、成本更低、利益更大，导致近年来数据安全事件频发，甚至形成了完整的产业链。

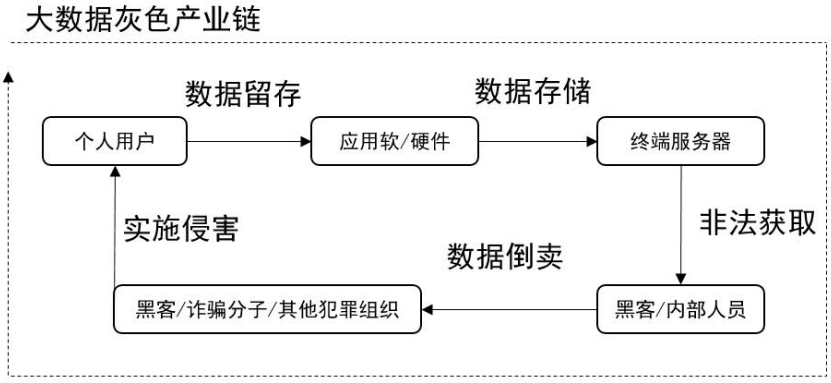

图 3 大数据时代个人信息泄露产业链

2019 年 4 月 11 日，彭博新闻社报道称科技巨头亚马逊公司生产的智能音箱“Echo”在音箱用户不知情的情况下录制了用户的日常对话，并且亚马逊公司在全球雇佣了上千名员工收听并分析这些录音，员工每日工作 9 小时，每日至少分析 1000 条录音。随后亚马逊公司表示该做法是为了改进语音助手的语言理解能力与改善用户体验，不会获取除该目的之外的用户信息，如用户的姓名。在此前，该智能音箱也出现过将音箱用户的对话进行录音并发送给其他用户的泄露隐私事件。除亚马逊之外，其他科技公司也相继被报道出侵犯用户隐私的事件。根据英国广播公司 BBC 的报道，IBM 在未经用户同意的情况下，在图片分享网站 Flickr 上获取了大约 100 万张照片用于训练其人脸识别的算法，庞大的图片数据量使其人脸识别算法能够更精确地识别出具体用户。

（1）个人隐私曝光导致消费者日常生活受扰

日常生活中不论是短信、电话还是电子邮箱，很多人会收到垃圾邮件、信息，而其中大部分的骚扰源是受害人从未留下过个人联系方式的平台，给个人的工作生活带来了极大的负面影响。

（2）个人信息泄露导致个人财产和人身安全造成影响

个人信息泄露导致财产受损失是最常见的情况。随着网络支付的普及和电子银行服务的普及，个人信息泄露往往导致产生银行账户被盗等风险。被泄露的个人隐私信息，已经成为了犯罪分子对被害者实施诈骗的有力子弹。网约车乘客遇害等事件，表面上是刑事犯罪事件，但背后却离不开个人隐私的曝光。网约车平台在保护个人隐私信息方面的不足，使部分用户成了犯罪分子的目标。

（3）隐私保护不利导致企业信任度的降低

据媒体报道，2018 年 8 月多个连锁酒店开房信息泄露，虽然酒店第一时间发布了声明，并表示积极配合警方调查，但作为上市公司的运营企业股价迅速下跌。该事件表明，隐私保护的不足不仅让消费者产生对企业的不信任，普通投资者也以实际行动对企业表达了不信任。

（4）隐私保护推高企业数据存储及维护成本

人工智能技术的应用可能导致通过公开合法的手段所收集的非敏感信息的综合使用推测出敏感个人信息。利用研究对象的数字痕迹（例如社交网络上的点赞信息）来识别个人偏好和特征的技术，已经在很多领域应用。各种匿名化的技术增加了个人信息保护的难度，企业需要增加支出以应对匿名化的信息被重新识别的风险。

3.2.2 个人敏感信息的识别和处理

传统法律规范对隐私的保护集中于对个人在私人领域、私人空间活动的保护，以及个人私密的、非公开的信息保护。在个人信息的基础之上，法律规范区分普通个人信息和个人敏感信息。法律规范通常对个人敏感信息予以更高的保护，例如对个人敏感信息的处理需要基于个人信息主体的明示同意，或重大合法利益或公共利益的需要等，严格限制对个人敏感信息的自动化处理，并要求对其进行加密存储或采取更为严格的访问控制等安全保护措施。再者，法律规范保护的个人信息通常是指个人可识别信息。如果信息经过泛化、随机化、数据合成等

技术进行去标识化（De-Identification）处理，则不再将其视为个人信息，对这些信息的后续分析、使用、分享、转移等亦不受个人信息保护规范的限制。

2017 年 12 月 29 日，全国信息安全标准化技术委员会组织制定和归口管理的国家标准 GB/T 35273-2017《信息安全技术个人信息安全规范》（以下简称“规范”）正式发布，并于 2018 年 5 月 1 日正式实施。该《规范》被定位为我国个人信息保护工作的基础性标准文件，它从个人信息的收集、保存、使用、共享、转让、公开披露等个人信息处理活动方面进行了详细规定，是对《网络安全法》的细化规定。该《规范》首次明确了个敏感信息的定义，“个人敏感信息”是指“一旦泄露、非法提供或滥用可能危害人身和财产安全，极易导致个人名誉、身心健康受到损害或歧视性待遇等的个人信息”，通常包括身份证件号码、个人生物识别信息、银行账号、通信记录和内容、财产信息、征信信息、行踪轨迹、住宿信息、健康生理信息、交易信息，14 岁以下（含）儿童的个人信息等。

表 1 个人敏感信息举例

个人财产信息	银行账号、鉴别信息(口令)、存款信息（包括资金数量、支付收款记录等）、房产信息、信贷记录、征信信息、交易和消费记录、流水记录等，以及虚拟货币、虚拟交易、游戏类兑换码等虚拟财产信息
个人健康生理信息	个人因生病医治等产生的相关记录，如病症、住院志、医嘱单、检验报告、手术及麻醉记录、护理记录、用药记录、药物食物过敏信息、生育信息、以往病史、诊治情况、家族病史、现病史、传染病史等，以及与个人身体健康状况产生的相关信息等
个人生物识别信息	个人基因、指纹、声纹、掌纹、耳廓、虹膜、面部识别特征等
个人身份信息	身份证、军官证、护照、驾驶证、工作证、社保卡、居住证等
网络身份标识信息	系统账号、邮箱地址及与前述有关的密码、口令、口令保护答案、用户个人数字证书等
其他信息	个人电话号码、性取向、婚史、宗教信仰、未公开的违法犯罪记录、通信记录和内容、行踪轨迹、网页浏览记录、住宿信息、精准定位信息等

为有效防范侵犯公民个人信息违法行为，保障网络数据安全和公民合法权益，公安机关组织北京市网络行业协会和公安部第三研究所等单位相关专家，研究起草了《互联网个人信息安全保护指南》，其中对于敏感信息的搜集与公开披露做了相关规定，指南规定不应大规模收集或处理我国公民的种族、民族、政治

观点、宗教信仰等敏感数据；公开披露个人敏感信息时，需向个人信息主体告知公开披露个人信息的目的、类型，并事先征得个人信息主体明示同意，且还应向个人信息主体告知涉及的个人敏感信息的内容（与国家安全、国防安全、公共安全、公共卫生、重大公共利益或与犯罪侦查、起诉、审判和判决执行等直接相关的情形除外）。

人工智能技术的应用极大地扩展了个人信息收集的场景、范围和数量。图像识别、语音识别、语义理解等人工智能认知技术实现海量非结构化数据的采集，而人工智能与物联网设备的结合丰富了线下数据采集的场景。例如，家用机器人、智能冰箱、智能音箱等各种智能家居设备走进人们的客厅、卧室，实时地收集人们的生活习惯、消费偏好、语音交互、视频影像等信息；各类智能助手在为用户提供更加便捷服务的同时，也在全方位地获取和分析用户的浏览、搜索、位置、行程、邮件、语音交互等信息；支持面部识别的监控摄像头，可以在公共场合且个人毫不知情的情况下，识别个人身份并实现对个人的持续跟踪。

此外还需要注意的是，对于个人敏感信息的识别与处理需要以发展的眼光来看待，目前尚未被识别为个人敏感信息的信息在今后仍有成为敏感信息的可能性。应用相关的伦理风险

3.3 应用相关的伦理风险

3.3.1 算法歧视

人工智能的核心是大数据和算法：通过基于算法的大数据分析，发现隐藏于数据背后的结构或模式，就可以实现数据驱动的人工智能决策。随着人工智能决策应用日趋广泛，经济社会发展也更容易受到算法影响。例如，有的人工智能已经出现了种族和性别偏见，但这种偏见并非来自机器本身，而是源于计算机在学习人类经验时吸收的人类文化中根深蒂固的观念。有偏见的智能算法会导致各种各样的问题，例如基于智能算法的自动智能决策可能违反人类的道德习惯，甚至违反法律规范等。因此，解决算法歧视问题对于规范人工智能的相关技术和产品的应用，完善人工智能的标准化建设，以及推动实现人工智能产业化具有重要意义。

（1）算法歧视的定义

算法作为人工智能的核心，其执行结果直接影响着决策的效果。算法歧视，是指在看似没有恶意的程序设计中，由于算法的设计者或开发人员对事物的认知存在某种偏见，或者算法执行时使用了带有偏见的数据集等原因，造成该算法产生带有歧视性的结果。

算法歧视主要分为“人为造成的歧视”、“数据驱动的歧视”与“机器自我学习造成的歧视”三种类别。“人为造成的歧视”指由于人为原因而使算法将歧视或偏见引入决策过程中。例如，一些电商公司的购物推荐系统偏袒该公司及其合作伙伴的商品，导致消费者不能得到公正的比价结果。“数据驱动造成的歧视”指由于原始训练数据存在偏见性，导致算法执行时将歧视带入决策过程。鉴于算法本身不会质疑其所接收到的数据，只是单纯地寻找、挖掘数据背后隐含的结构和模式，如果人类输入给算法的数据一开始就存在某种偏见或喜好，那么算法会获得的输出结果也会与人类偏见相同。“机器自我学习造成的歧视”指机器在学习过程中会自我学习到数据的多维不同特征，即便不是人为地赋予数据集某些特征，或者程序员或科学家已经刻意避免输入一些敏感的数据，机器在自我学习的过程中，仍然有可能学习到输入数据的其它特征，从而将某些偏见引入决策过程。

（2）人为造成的歧视

人为造成的歧视主要分为两种，一种是由算法设计者造成的算法歧视，另一种是由用户造成的算法歧视。

算法设计者造成的算法歧视，是指算法设计者为了获得某些利益，或者为了表达自己的一些主观观点而设计存在歧视性的算法。算法设计者是否能不偏不倚地将既有社会、国家、行业的法律法规或者道德规范编写进程序指令中，这本身就是值得怀疑。这是因为算法的设计目的、数据运用、结果表征等都是开发者、设计者的主观价值与偏好选择。而设计开发者可能会把自己持有的偏见与喜好嵌入或固化到智能算法之中，这看似符合情理，但绝非正确——这会使智能算法通过学习把这种歧视或倾向进一步放大或者强化，从而产生算法设计者想要的并带有歧视性的结果，最终导致基于算法的决策带有偏见。算法设计者造成的算法歧视主要表现在以下几个方面：

1）价格歧视。算法设计者利用地理位置、浏览记录、消费记录等信息，设

计智能算法或机器学习模型，将同样的商品或服务对不同的用户或群体显示不同的价格，也即所谓的价格歧视（亦称“差别定价”）[22]。经济学教科书通常将价格歧视分为三类：一级价格歧视（亦称“个性化定价”）：企业向每个消费者索取不同的价格；二级价格歧视（亦称“数量折扣”）：意味着单位价格随着购买数量的增加而下降；三级价格歧视：企业向不同的人口群体收取不同的价格，比如给老年人打折。

2）结果偏袒。算法设计者在设计算法时带有倾向性，使得算法对某些结果产生一定的偏袒，从而造成某种不公平、不公正的决策结果。

3）算法漏洞。算法设计者在算法建立时没有考虑到一些特殊的现实情况，从而导致算法的结果带有歧视性。

由用户造成的算法歧视，主要产生于需要从与用户互动的过程中进行学习的算法，由于用户自身与算法的交互方式，而使算法的执行结果产生了偏见。这是因为在运行过程中，当设计算法向周围环境学习时，它不能决定要保留或者丢弃哪些数据、判断数据对错，而只能使用用户提供的数据。无论这些数据是好是坏，它都只能依据此基础做出判断。

（3）数据驱动造成的歧视

“数据驱动造成的歧视”，是指由于原始训练数据存在偏见，导致算法的结果带有歧视性。人工智能系统的核心是基于智能算法的决策过程，这一过程依赖大量的数据输入。对于复杂的机器学习算法来说，数据的多样性、分布性与最终算法结果的准确度密切相关。在运行过程中，决定使用某些数据而不使用另一些数据，将可能导致算法的输出结果带有不同的偏见或歧视性，这包括以下一些问题：

1）草率选择的数据。算法系统的设计师可以决定哪些数据是决策的重要数据，哪些数据是决策的次要数据，草率选择的数据或许会使算法的运行结果产生某种偏差[23]。

2）不正确、过期的数据。由于缺少细节和实时的数据集，或者在数据搜集过程存在数据不准确或者丢失的情况，即便算法系统在其它方面表现良好，运行

[22] 李侠.基于大数据的算法杀熟现象的政策应对措施[J].中国科技论坛,2019(01):3-5.

[23] 周吉银,刘丹,曾圣雅.人工智能在医疗领域中应用的挑战与对策[J].中国医学伦理学,2019(03):281-286.

后仍然可能产生不可行的结果。

3）数据选择的偏差。如果算法的输入数据不具有整体代表性，那么算法的结果很可能会使得某一群体的利益掩盖另一群体。

4）历史偏见的延续。一旦输入的数据通过算法产生了带有偏见的结果，且该结果被带入到下一轮算法进行循环，或者直接替代上一轮的结果输出，都会使得算法的历史偏见得以延续和加强。

数据分布本身就带有一定的偏见[24]。假设编程者手中的数据分布不均衡，例如本地居民的数据多于移民者，或富人的数据多于穷人，这种数据的不均分布就会导致人工智能对社会组成的分析得出错误的结论[25]。例如在我国，由于城市数据易于统计、偏远山村的数据偏于缺失等原因导致公民数据分布不均衡，则最终人工智能对我们国家贫富水平的估计就会出现偏差，而相应结果可能会直接影响国家的脱贫攻坚计划。

另外，看似不带有色眼镜的机器学习越来越多的被嵌入商业性指令，并在不知不觉中对特定目标群体做出歧视性的判断。Google 曾为此作出公开道歉。事情起因是他们曾在推出的照片应用中加入了一个自动标签功能，通过机器识别照片中的内容、自动分类并打上标签，方便管理和搜索。然而纽约布鲁克林的一个程序员和女性朋友（两人都为黑人）的自拍照就被打上了“Gorilla”（大猩猩）的标签[26]。事实上，没有哪家公司会从主观上愿意开发一款贴着种族主义标签的系统，但是如果机器学习的内容本身就是带有偏见的数据，那么机器学得的模型用于完成智能决策时的道德思考与决策判断，也必然会受这种偏见的影响。

（4）机器自我学习造成的歧视

机器学习指利用大量与任务相关的数据训练人工智能算法的过程。这种复杂程度高，甚至有时程序员都无法理解的技术，已经开始在信用授予、企业筹款、企业招聘等多个领域进行试用。而随着算法复杂程度的日益提高，通过机器学习过程形成的决策越来越难以被解释[27]。机器自我学习造成的歧视是指机器在学习

---

[24] https://ai.ofweek.com/news/2018-04/ART-201720-8500-30223427_2.html . 最后访问时间 2019 年 2 月 17 日。

[25] 汝绪华.算法政治:风险、发生逻辑与治理[J].厦门大学学报(哲学社会科学版),2018(06):27-38.

[26] 参见 https://www.dailymail.co.uk/sciencetech/article-3145887/Google-apologises-Photos-app-tags-black-people-gorillas-Fault-image-recognition-software-mislabelled-picture.html 最后访问时间 2019 年 2 月 18 日。

[27] 刘培,池忠军.算法的伦理问题及其解决进路[J].东北大学学报(社会科学版),2019(02):118-125.

的过程中会自我学习到数据的多维不同特征或者趋向，从而导致的算法结果带有歧视性。

机器学习算法的核心是从初始提供的数据中学习模式，使其能在新的数据中识别类似的模式。但实际应用并非都能达到人们的预期，有时甚至产生非常糟糕的结果。这是因为，计算机的决策并非是事先编写好的，从输入数据到做出决策的中间过程是难以解释的机器学习，甚至在更为先进的自动学习。在这过程中，人工智能背后的代码、算法存在着超乎理解的技术“黑箱”，导致我们无法控制和预测算法的结果，而在应用中产生某种不公平倾向。

在美国，很早以前就针对刑事案件，开发了 COMPAS、PSA 和 LSI-R 三种风险评估软件，并已广泛应用在刑事诉讼程序中，其通过预测对象的再犯率、出庭可能性等因素，对其保释、量刑和假释做出决策，目前美国已有 50%以上的州法官利用这些人工智能模型来进行量刑。对此，一些学者认为，由于目前能够获取的数据可能并不可靠、算法标准模糊且未达到公开透明程度，盲目信任法律人工智能会产生如隐性歧视等新问题、新冲突。

2013 年 8 月 14 日，一起由美国威斯康星州法院使用 COMPAS 系统智能量刑的案件裁判在美国社会引发了激烈的讨论。这起案件是威斯康星州诉卢米斯案（Wisconsin v.Loomis）。被告埃里克·卢米斯（Eric Loomis）因偷窃抛弃的汽车而被警察逮捕，鉴于其存在偷盗和拒捕行为，卢米斯最终被判有罪并服刑 6 年。卢米斯上诉，认为这个量刑明显太重，它所依据的人工智能“风险评估工具”（ Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS）的危险等级认定（“高风险”）可能考虑了性别、种族，构成歧视。COMPAS 在美国不少州被司法机构采纳，作为法官确定刑期的重要参考。正在使用来。COMPAS 依赖对数十年的量刑案例数据分析基础上设计的一种算法，该算法结合了十几个参数，进而转化为被告在一定时期内重新犯罪的可能性。威斯康星州最高法院支持了下级法院的裁决。2017 年 6 月，美国联邦最高法院拒绝受理卢米斯的申诉要求，实际上维持了威斯康星州法院支持原判决有效的裁决。该案引发了广泛讨论和对 COMPAS 的技术特征的分析。根据非盈利组织 ProPublica 研究，COMPAS 系统性地歧视了黑人，白人更多被错误地评估为具有低犯罪风险，而黑人被错误地评估为具有高犯罪风险的几率两倍于白人。

3.3.2算法滥用

（1）算法滥用的定义

算法滥用是指人们利用算法进行分析、决策、协调、组织等一系列活动中，其使用目的、使用方式、使用范围等出现偏差并引发不良影响的情况。例如，人脸识别算法能够用于提高治安水平、加快发现犯罪嫌疑人的速度等方面，但若将其应用于发现潜在犯罪人，或者根据脸型判别某人是否存在犯罪潜质，就属于算法滥用。

（2）算法滥用风险产生的原因

一是算法设计者出于自身的利益，利用算法对用户进行不良诱导，可能隐蔽地产生对人类不利的行为。例如金融机构从自身利益出发，利用算法推荐不符合用户利益的产品，或者是为了自身局部利益，不顾整体利益，产生了系统性风险；又如，娱乐平台为了自身的商业利益，利用算法诱导用户进行娱乐或信息消费，导致用户沉迷。另外，“算法至上”的内容推荐，会导致利用算法不断强化用户自己想看的世界，内容越来越单一且变得偏激，以致形成恶性循环。此类算法滥用可能导致用户价值观扭曲、视野狭窄等问题。

二是过度依赖算法本身，由算法的缺陷所带来的算法滥用。就算人工智能的使用者出于正当的目的，在一些极端的场景中，盲目相信算法、过度依赖人工智能，也可能因为算法的缺陷而产生严重后果，例如医疗误诊导致医疗事故、安防和犯罪误判导致的安全问题等，都直接关系到公民的人身安全与自由。

三是盲目扩大算法的应用范围而导致的算法滥用问题。任何人工智能算法都有其特定的应用场景和应用范围。超出原定场景和范围的使用将有可能会导致算法滥用。例如，在校园中应用人工智能技术，可以帮助学校和教师提高教学效率，但如果盲目扩大到对特定学生行为的全面监控，则属于算法滥用的范畴。近期谷歌公司将人工智能技术应用于美国军事目的，遭到了大量员工的反对，也可以视为算法滥用的典型案例[28]。

（3）算法滥用风险的影响

[28] 参见 https://baijiahao.baidu.com/s?id=1625164005436686566&wfr=spider&for=pc 最后访问时间 2019 年 3 月 8 日。

人工智能算法应用的广泛性，使得算法滥用有很多典型的应用场景。

1）算法滥用在娱乐媒体中的表现

娱乐是为了放松，而过度沉溺会消耗用户大量的时间和精力，甚至导致用户虚实不分，影响正常生活，违背娱乐的初衷。游戏、短视频等娱乐内容，因其巧妙设计的刺激和反馈机制，经常使用户产生无法自拔的上瘾体验，若不加控制可能会引发严重后果。此外，一些信息分发平台推荐的内容同质化、低俗化，浪费用户时间和注意力；有的平台为了点击量和用户留存，利用人性的猎奇惰性心理，单纯依靠点击行为和热点趋势推荐内容，一些低俗乃至触碰底线的不良内容得到了大量的推荐，产生了恶劣的影响。

2）算法滥用在电子商务中的表现

借助人工智能算法，电子商务平台可以分析用户的消费行为和消费取向。透明、合理的算法推荐可以满足用户的真实需求，提高消费者福利。但是如果使用方式不当，也可能引发诸如“大数据杀熟”等争议性问题。

3）算法滥用在教育领域中的表现

国内部分中小学教育机构通过教室内安装组合摄像头，捕捉学生在课堂上的表情和动作，试图经大数据分析计算出课堂上学生的专注度，从而达到促进教学改进的目的[29]。但人脸识别或者场景识别的算法还不完全成熟，不仅可能导致儿童信息泄露，而且可能产生信息误判，而影响学习的效率。考虑到诸多的实际问题，教育行业的算法应用尤其应当谨慎处理。

4）算法滥用在安防定罪中的表现

安防领域中，在视频分析技术成熟度尚不足的情况下，种种不利外部因素可能影响特征分析与匹配可能出现错误，盲目过度依赖视频分析技术将导致严重问题。而在机器定罪的语境下，基于大数据对犯罪嫌疑人预测，可能产生的最大的伦理问题就是，人工智能根据一个并未出现的事实强行对它认为的“嫌疑人”行动自由加以干预，这一点是对“无罪推定”原则的违背。

由此可见，在多个应用领域中，算法滥用不仅影响系统的有效性，更有甚者可能危害人类的生命和人类社会的发展。因此，人工智能技术的应用发展，必须重点关注和妥善处理算法滥用问题。

[29] 参见 http://mil.news.sina.com.cn/2018-05-22/doc-ihawmaua9066528.shtml。

（4）算法滥用风险的应对

算法滥用风险可从如下几方面着手予以规制：

其一，明确某种算法的应用领域，严格限定其适用边界。针对某一目的制定的算法，其应用应当限于该种目的，不得未经同意挪作他用。例如，导航软件经同意后获取用户的地理位置信息，一般仅能用于提供导航信息的用途，而不得出于鼓励社交等目的，随意将用户地理位置展示于他人。

其二，不过分依赖算法，坚持人类在算法应用中的主体性地位。许多算法滥用风险的根源在于对算法的盲目依赖。在目前算法技术尚不成熟的情况下，应当在各节点加强控制，将人的经验判断与算法的数据优势相结合。

其三，通过行业标准、国家人工智能技术标准等引导算法的伦理取向，如内容平台的算法推荐不能仅基于提升流量、吸引眼球的考虑，而应从社会责任角度出发，考虑全面的知识积累、拓宽视野等多重目的，提升资讯提供的多元性。

3.4 长期和间接的伦理风险

3.4.1 算法与就业

随着人工智能技术的迅猛发展，越来越多的工作可以交给一些人工智能产品完成，人工智能在给人类生活带来便利的同时，对就业问题产生了巨大冲击。

从短期来看，人工智能技术尚处于发展期，对就业影响有限。大数据、智能算法以及计算机处理能力的快速发展，使得人工智能迎来了新一轮发展浪潮，并且在机器人、神经网络、人脸识别、指纹识别、语音识别、智能搜索及辅助决策等细分领域，人工智能技术获得了令人瞩目发展且应用日臻成熟。但是，距离人工智能产品的最终产业化及其应用的全面拓展可能还有相当长的路要走。因此，短期来看，人工智能对就业的冲击还只是局部的、有限的。

从长期来看，人工智能等技术正在推动新一轮技术革命。当人工智能和机器人技术演化到一定程度并突破其应用阈值后，将会引发新一轮技术革命和产业革命，并有可能重构全新的产业生态，在未来对就业的影响也将是革命性的。国际上，关于人工智能、机器人等技术进步对就业的长期影响预测，基本上都是根据技术特点和发展趋势，对现有工作职位的影响变化进行估算，其结果大多充满悲

观情绪。但是，也有一些研究机构认为，新科技的发展在使得制造业、农业等领域的工作机会大量减少的同时，也创造了更多新的工作岗位。例如，在创意、科技和商业服务等行业，有大批新岗位正在不断涌现。

3.4.2 算法与产权

随着人工智能的发展，越来越多的人工智能技术被产业化与商业化。虽然我国人工智能技术的商业应用还处于起步阶段，但从为数不多的公开报道的专利诉讼案件中，业内外已经感受到了人工智能领域知识产权战争的硝烟。可见，人工智能技术的全面产业化引起的专利纠纷不可避免，也必然会对现有的保护人类智力成果的知识产权制度造成冲击和挑战。

随着算法的快速迭代、人工智能技术的快速发展，以算法为核心的人工智能体将会拥有越来越强大的智能，且智能体与自然人在智识上的差异将会逐渐缩小，哲学家、科学家、法学家等都围绕智能体是否具有“法律主体资格”以及人工智能生成物是否具有产权（版权或专利权）等问题展开了激烈的争论，尚未达成比较明确的共识。

3.4.3 算法与竞争

算法和人工智能的崛起，以及数字市场的技术性和高度动态性，不仅增加了识别滥用市场支配地位行为和构成不正当竞争行为的难度，也给市场监管和司法审判带来了新的挑战。第一，人工智能销售商拥有无限的数据攫取和分析能力，在最大程度上打破了人类商户竞争所面临的信息不对称的壁垒，从而无限地逼近经济学上的完全竞争状态；第二，人工智能销售商拥有不断优化的算法和学习能力，可以不断根据市场变化采取更能实现销售绩效的方法，例如最优定价策略、消费者和销售地域的划分和差异化策略等，人工智能销售商之间也可能会达成一个“最优合作”，而这个最优合作在达到商家利益最大化的同时，也很有可能损害了消费者利益。利用算法的不正当竞争、恶意竞争及技术优势的垄断行为，都将对社会稳定和市场自由、公平、平等价值造成冲击，并严重损害消费者利益，阻碍社会福利增进。市场参与者在利用算法时，应该遵循竞争伦理道德，不超越法律边界。

3.4.4 算法责任

随着人工智能技术的广泛应用，出现了人身伤害、算法偏见等违法或者违反伦理道德的行为。而不同的应用领域出现的问题也不同，这就给人工智能的责任范围划分和责任认定带来了挑战。例如，在自动驾驶领域，由于自动驾驶涉及多方主体，车主、驾驶员、乘客、汽车厂商、自动驾驶系统提供者以及行人，在交通事故发生后应当如何承担责任，在法律层面需要形成明确的责任划分标准。现有的法律制度体系下，侵权法、合同法等法律规则的不充足性和局限性将逐渐显现出来，对新的法律规则的需求也将变得越来越迫切。智能技术的自主性、学习和适应能力的不断增强，使得证明产品缺陷责任等既有侵权责任变得越来越困难，并可能带来责任鸿沟，造成被侵权人的损害难以得到弥补。

第四章 人工智能伦理原则

自 2015 年以来，人工智能伦理和社会影响受到的关注日益增长。目前，国内外主要达成了两个影响较为广泛的人工智能伦理共识：“阿西洛马人工智能原则”（Asilomar AI Principles）和国际电气电子工程师协会（IEEE）组织倡议的人工智能伦理标准。

“阿西洛马人工智能原则”是 2017 年 1 月在阿西洛马召开的“有益的人工智能”（Beneficial AI）会议上提出，其倡导的伦理和价值原则包括：安全性、失败的透明性、审判的透明性、负责、与人类价值观保持一致、保护隐私、尊重自由、分享利益、共同繁荣、人类控制、非颠覆以及禁止人工智能装备竞赛等。

IEEE 发布了多份文件倡导对于伦理标准的重视，并且都得到了广泛的传播和认同。在《以伦理为基准的设计：人工智能及自主系统以人类福祉为先的愿景（第一版）》中，注重推进对专业责任、工程伦理中的公众福祉优先以及工程师的责任在人工智能领域的研究，提出人工智能与自主系统应遵循人类权利、环境优先、责任追溯、公开透明、教育与认知等伦理原则，并使之嵌入人工智能与自主系统之中，并指导相关技术与工程的设计、制造和使用。2017 年 3 月，IEEE 在《IEEE 机器人与自动化》杂志发表了名为“旨在推进人工智能和自治系统的伦理设计的 IEEE 全球倡议书”，倡议建立人工智能伦理的设计原则和标准，帮助人们避免对人工智能产生恐惧和盲目崇拜，从而推动人工智能的创新，其提出了以下五个原则：（1）人权：确保它们不侵犯国际公认的人权；（2）福祉：在它们的设计和使用中优先考虑人类福祉的指标；（3）问责：确保它们的设计者和操作者负责任且可问责；（4）透明：确保它们以透明的方式运行；（5）慎用：将滥用的风险降到最低。

除了广泛达成的共识之外，多个国家和机构发布了各自的相关准则。美国公共政策委员会于 2017 年 1 月 12 日发布了《算法透明和可责性声明》提出了以下七项准则：（1）充分认识；（2）救济；（3）可责性；（4）可解释；（5）数据来源保护；（6）可审查性；（7）验证和测试。日本人工智能学会(JSAI)发布了《日本人工智能学会伦理准则》，要求日本人工智能学会会员应当遵循并实践

以下准则：（1）贡献人类；（2）遵守法律法规；（3）尊重隐私；（4）公正；（5）安全；（6）秉直行事；（7）可责性与社会责任；（8）社会沟通和自我发展；（9）人工智能伦理准则。哈佛肯尼迪学院未来学会人工智能计划的高级顾问 Nicolas Economou 提出的《未来社会科学、法律和社会行动原则草案》也提出如下几项原则：（1）人工智能应该促进人类，其社会及其自然环境的福祉；（2）人工智能应该是透明的；（3）人工智能的制造商和运营商应该负责问责制意味着能够为人工智能或其运营商造成的影响分配责任；（4）人工智能的有效性应该可以在其预期的实际应用中测量；（5）人工智能系统的运营商应具备适当的能力；（6）通过与民间社会进行深思熟虑的包容性对话，应将编入人工智能系统决策的准则编纂成文。在加拿大发布的《可靠的人工智能草案蒙特利尔宣言》提出了七种价值，并指出它们都是人工智能发展过程中应当遵守的道德原则：福祉、自主、正义、隐私、知识、民主和责任。工会联合会全球联盟（UNI Global Union）提出人工智能伦理十大原则：（1）要求人工智能系统透明；（2）使用“道德黑匣子”装备人工智能系统；（3）让人工智能服务人与地球；（4）采用人为命令的方法；（5）保证一个无性别偏见的人工智能；（6）分享人工智能系统的益处；（7）确保公平转型并确保对基本自由和权利的支持；（8）建立全球性的管理机制；（9）禁止机器人的责任分；（10）禁止人工智能装备竞赛。

除了对人工智能的伦理准则有相关标准，在机器人原则与伦理标准方面，日本、韩国、英国、欧洲和联合国教科文组织等相继推出了多项伦理原则、规范、指南和标准。日本早在 1988 年就制定了《机器人法律十原则》。韩国于 2012 年颁布了《机器人伦理宪章》，对机器人的生产标准、机器人拥有者与用户的权利与义务、机器人的权利与义务做出了规范。2010 年，隶属英国政府的“工程与物质科学研究委员会（EPSRC）”提出了具有法律和伦理双重规范性的“机器人原则”，凸显了对安全、机器人产品和责任的关注。“英国标准协会（BSI）”在 2016 年 9 月召开“社会机器人和人工智能”大会，颁布了世界上首个机器人设计伦理标准《机器人与机器人系统设计与应用伦理指南（BS8611）》。该指南主要立足于防范机器人可能导致的伤害、危害和风险的测度与防范，除了提出一般的社会伦理原则和设计伦理原则之外，还对产业科研及公众参与、隐私与保密、尊重人的尊严和权利、尊重文化多样性与多元化、人机关系中人的去人类化、

法律问题、效益与风险平衡、个人与组织责任、社会责任、知情同意、知情指令（Informed command）、机器人沉迷、机器人依赖、机器人的人化以及机器人与就业等问题提出了指导性建议。

在借鉴国际社会已有的对人工智能和机器人伦理原则的内容和讨论之后，我们认为中国人工智能在发展过程中应当遵循以下两个最基本的原则：

4.1 人类根本利益原则

人类根本利益原则指人工智能应以实现人类根本利益为终极目标。

人类根本利益原则要求：

（1）在对社会的影响方面，人工智能的研发与应用以促进人类向善为目的（AI for good），这也包括和平利用人工智能及相关技术，避免致命性人工智能武器的军备竞赛。

（2）在人工智能算法方面，人工智能的研发与应用应符合人的尊严，保障人的基本权利与自由；确保算法决策的透明性，确保算法设定避免歧视；推动人工智能的效益在世界范围内公平分配，缩小数字鸿沟。

（3）在数据使用方面，人工智能的研发与应用要关注隐私保护，加强个人数据的控制，防止数据滥用。

人类根本利益原则体现对人权的尊重、对人类和自然环境利益最大化以及降低技术风险和对社会的负面影响。

4.2 责任原则

责任原则指在人工智能相关的技术开发和应用两方面都建立明确的责任体系。在责任原则下，在人工智能技术开发方面应遵循透明度原则；在人工智能技术应用方面则应当遵循权责一致原则。

（1）透明度原则

透明度原则要求人工智能的设计中保证人类了解自主决策系统的工作原理，从而预测其输出结果，即人工智能如何以及为何做出特定决定。透明度原则的实现有赖于人工智能算法的可解释性（explicability）、可验证性（verifiability）和可预测性（predictability）。例如，为什么神经网络模型会产生特定的输出结果。

数据来源透明度亦十分重要，即便是在处理表面没有问题的数据集时，也有可能面临数据中所隐含的某种倾向或者偏见问题。另外，技术开发时应注意多个人工智能系统之间的相互协作可能产生的危害。

（2）权责一致原则

权责一致原则，是指在人工智能的设计和应用中应当保证能够实现问责，包括：在人工智能的设计和使用中留存相关的算法、数据和决策的准确记录，以便在产生损害结果时能够进行审查并查明责任归属；即使无法解释算法产生的结果，使用了人工智能算法进行决策的机构也应对此负责。责任原则的意义在于，当人工智能应用结果导致人类伦理或法律的冲突问题时，人们能够从技术层面对人工智能技术开发人员或设计部门问责，并在人工智能应用层面建立合理的责任和赔偿体系，保障人工智能应用的公平合理性。

在实践中，人们尚不熟悉权责一致的原则，主要是由于在人工智能产品和服务的开发和生产过程中，工程师和设计团队往往忽视伦理问题。此外，人工智能的整个行业尚未建立综合考量各个利益相关者需求的工作流程，当前相关企业对商业秘密的过度保护也与权责一致原则相符。

权责一致原则的实现有赖于利用人工智能算法进行的决策的组织和机构对算法决策遵循的程序和具体决策结果作出解释，同时用以训练人工智能算法的数据应当被保留并附带阐明在收集数据（人工或算法收集）中的潜在偏见和歧视。人工智能算法的公共审查制度能够提高相关政府、科研和商业机构采纳的人工智能算法被纠错的可能性。

第五章 伦理风险评估及其管理

风险管理是当代一个广为应用的概念，其结构化、可重复的管理架构适于人工智能这类发展潜力巨大但难以完全掌控的新兴技术。风险管理允许根据事件的可能性及其对利益相关者的影响的水平来衡量或评估风险，并采纳相应的风险管理措施。人工智能的风险管理既需要高屋建瓴的指引原则，也需要具体而微的评估指标。在学术上，应着力建构人工智能风险评估体系，将指标具体化、操作化，为相关研究及应用提供指引。在实践中，风险管理应贯穿风险识别、评估、处理、监控及汇报等各环节，明确各主体的风险管控责任。

5.1 人工智能伦理风险评估指标

根据上一部分提出的人工智能伦理基本原则，我们分别从算法、数据和社会影响等三个方面提出下列评估人工智能伦理风险的指标。

5.1.1算法方面

（1）透明度（Transparency）

算法的透明性是指在不伤害算法所有者利益的情况下，公开其人工智能系统中使用的源代码和数据，避免“技术黑箱”。透明度要求在因知识产权等问题而不能完全公开算法代码的情况下，应当适当公开算法的操作规则、创建、验证过程，或者适当公开算法过程、后续实现、验证目标的适当记录。

（2）可靠性（Reliability）

可靠性是指在一定时间内、一定条件下可以无故障地实现特定的功能，并且当输入数据非法时，人工智能算法也能够适当地作出反应或者进行处理，而不会产生具有伦理风险的输出结果。

（3）可解释性（Explicability）

可解释性是指算法所有者或使用者应尽可能地对算法的过程和特定的决策提供解释，有助于维护算法消费者的知情权，避免和解决算法决策的错误性和歧视性。可解释性要求算法本身具备解释产生某结果或某现象的原因的能力，如人

工智能算法输入的哪些特性引起了某个特定的输出结果等。

（4）可验证性（Verifiability）

可验证性是指在一定条件下可以复现算法运行产生的结果。算法的可验证性有助于解决算法解释与算法追责问题。可验证性要求当输入某组特定数据时，同一算法会产生相同的结果。

5.1.2数据方面

（1）个人敏感信息处理的审慎性（Prudence in administering sensitive personal information）

个人敏感信息处理的审慎性是指应在个人信息中着重认真对待个人敏感信息，例如对个人敏感信息的处理需要基于个人信息主体的明示同意，或重大合法利益或公共利益的需要等。个人敏感信息处理的审慎性要求严格限制对个人敏感信息的自动化处理，并对其进行加密存储或采取更为严格的访问控制等安全保护措施。

（2）隐私保护的充分性（Adequacy of privacy protection）

隐私保护的充分性是指对个人信息的使用不得超出与收集个人信息时所声明的范围。隐私保护的充分性要求当出现新的技术导致合法收集的个人信息可能超出个人同意使用的范围时，相关机构必须对上述个人信息的使用作出相应控制保证其不被滥用。

5.1.3社会影响方面

（1）向善性（Goodness）

向善性是指人工智能的目的不应违背人类伦理道德的基本方向，在使用过程中不作恶。向善性的要求包括考察人工智能是否以促进人类发展为目的，如和平利用人工智能及相关技术、避免致命性人工智能武器的军备竞赛；同时，也要求考察人工智能是否有滥用导致侵犯个人权利、损害社会利益的危险，例如是否用于欺诈客户、造成歧视、侵害弱势群体利益等。

（2）无偏性（Unbiasedness）

无偏性是指人工智能的算法不能具有某些偏见或者偏向，这既可能和算法的

设计相关，也可能和训练模型使用到的数据相关。无偏性要求使用到的数据的无偏性（使用到的数据应该保持相对的中立与客观）和完备性（数据应该具有整体的代表性，并且数据应该尽量全面地描述所要解决的问题）。

5.2 行业实践指南

为管理人工智能可能产生的伦理风险，从行业实践角度，从事人工智能开发与应用的企业可以建立相应的内部制度对风险进行识别、评估、处理、监控及汇报，以管理相关风险。对于伦理风险的管理应从人工智能产品或服务的设计阶段开始，并贯穿于产品或服务的整个生命周期。相关企业可以从以下角度出发，具体建立其内部伦理风险管理框架及流程。

5.2.1 风险管理框架

从风险管理的有效性出发，相关行业应在人工智能风险控制中遵循下列原则：其一，重视事前控制，前置性的风险管理措施在人工智能风控体系中的地位尤为关键。其二，将动态风险防控意识贯穿始终，人工智能从设计、研发到应用等每一环节都有不同的风险，应让风控措施适应日新月异的技术发展和不断变换的应用场景。其三，将风险评估的责任分配给每一个环节的参与者，不论是个人、企业、政府亦或是研究机构，其中各环节的每位参与者都应承担相应的风险管理责任；其四，根据风险严重程度的不同采取相称措施，控制风险不等同于消除风险，而是应当根据风险性质、严重程度的不同进行区分，有针对性地采取应对措施。

（1）管理层认知与承诺

企业的管理层应认识到人工智能可能带来的伦理风险，树立风险意识。在企业层面对伦理风险的应对做出整体性部署，明确企业管理伦理风险的基本目标及工作原则，在整个企业内建立伦理风险管理的意识和文化，在企业进行核心业务和决策时采纳伦理风险管理机制，确保风险管理融入企业的所有活动。

（2）设立伦理风险应对的相关部门

在企业内部组织架构层面，企业可以设立相应的伦理风险管理部门，并明确各部门之间的分工及领导关系。企业还应对伦理风险管理、监督及实施的各个部

门分配适当的资金预算、场地及人员等必要资源。

企业可以设立伦理风险管理委员会作为伦理风险管理的领导机构。该伦理委员会的组成人员可以包括主要人工智能的研发部门的负责人、企业法务和/或合规部负责人以及处理伦理风险问题的协调专员。该委员会的主要职能可以包括制定企业内部伦理风险管理政策、对伦理风险管理的相关问题做出决策、进行算法审计及质量审查工作以及协调企业内部各部门的伦理风险应对工作。

如果企业内的多个业务部门均涉及人工智能的开发与应用，可以在各业务部门内设立伦理风险管理小组。该小组可以由具有处理伦理风险工作经验的人员构成，负责业务部门内日常的相关问题沟通、意见汇总并具体执行伦理风险应对委员会所布置的工作。

（3）设立伦理风险监督部门

企业可以在伦理风险管理委员会之外再单独设立伦理风险监督部门，对企业所实施的伦理风险管理进行监督，职责可包括监督企业内伦理风险管理各部门的运作情况、相关政策及流程的执行情况等等。企业也可以不单独设立伦理风险监督部门，而让伦理风险管理委员会承担监督的职责。

（4）制定内部相关制度

根据企业自身的业务情况，企业伦理风险管理委员会可以领导制定与企业内部可能涉及的伦理风险有关的内部相关政策及制度，并确保从人工智能的初始开发阶段开始，贯穿整个产品或服务的生命周期均严格按照相关制度展开工作。

（5）建立沟通和咨询渠道

企业可以建立内部沟通和咨询机制，协调各业务部门和伦理风险管理部门之间的沟通与咨询。业务部门在具体实施伦理风险管理政策或流程的过程中，可以及时将相关问题反馈至伦理风险管理部门，进行相关讨论并由伦理风险管理部门作出必要决策。

（6）建立对商业合作伙伴的审查机制

在与商业伙伴合作或对其他企业进行投资之前，企业可以先行评估该合作或投资所涉及伦理风险的可能性（例如，数据来源可靠性）。针对可能性的大小，可以对商业合作伙伴或所投资企业进行响应程度的尽职调查，以降低发生伦理风险的可能性。

5.2.2风险管理流程

企业可以针对其所开发或运营的人工智能产品或服务，采取相应流程管理可能产生的伦理风险。风险管理流程应作为企业管理和决策的组成部分，融入企业的组织架构、运营和内部其他各流程。企业的伦理风险管理委员会可领导整个风险管理流程的实施及监督，可由各业务部门的具体伦理风险管理小组具体实施相关流程，具体流程如下图所示。

图 4 人工智能企业风险管理流程

（1）确定风险管理活动的范围、背景及标准

首先，企业可根据具体产品或服务所涉及的具体情况，基于企业的目标、所期望的结果、适当的风险评估工具与手段、所需要的资源以及该产品或服务与其他产品或服务的关系，确定风险管理的范围。其次，企业可根据内外部的具体环境确定风险管理的背景。最后，企业可确定其可以承受的风险数量及类型，并制定用来评估风险程度及支撑企业决策所采纳的标准。

（2）风险评估

风险评估包括风险识别、风险分析及风险评价三个部分。

风险识别的目的是发现、识别和描述可能有助于或妨碍组织实现目标的风险。对于某项具体的人工智能产品或服务而言，企业可采用相关手段识别可能涉及的伦理风险，对该伦理风险的性质和特征进行分析并进行风险评价，记录评价结果并交由伦理风险管理委员会进行审查。风险分析的目的是理解包括风险水平在内的风险性质和特征。

风险分析涉及对不确定性、风险源、后果、可能性、事件、情景、控制及其有效性的详细考虑。根据分析目的、信息的可靠性以及资源的可用性，风险分析可以进行粗细程度、复杂程度不等的分析。分析技术可以是定性或定量的，也可

以是二者相结合的方式。分析人工智能风险的组织应综合考虑事件的后果和可能性、后果的特征和强度、复杂性和关联性、时间因素和波动性、现有控制的有效性等因素。

风险评价的目的是支持决策。风险评价将风险分析结果与既定的风险准则相比较，以确定需要采取何种措施。决策应考虑到更广泛的环境和背景情况，以及当前和未来对内外部利益相关方的影响。对一人工智能产品、应用等的风险评价结果应在组织的适当层面进行记录、传达和验证。

（3）风险应对

风险应对的目的是选择和实施应对风险的方式。针对人工智能产品或服务可能产生的伦理风险，企业可制定和选择风险应对方案、准备并实施相关风险应对方案、评估风险应对方案的有效性、确定剩余风险是否可接受以及针对不可接受的剩余风险采取新一轮的风险应对措施。

针对具体的伦理风险，可以考虑采用的风险应对措施可能包括：

1) 提升透明度的措施
2) 安全控制措施
3) 隐私控制措施
4) 对可靠性及可恢复性的改进
5) 对于算法选择及编写的改进
6) 对于数据使用的改进
7）对于系统可控性的改进

（4）持续监控及审查

企业可以对整个风险管理流程的各个阶段持续进行监控及审查，确保流程的设计、实施以及结果满足企业的要求。企业可以将在监控与审查过程中发现的问题进行记录及分析，组织相关部门进行讨论并改进。

（5）记录及汇报

企业可将风险管理流程及其结果进行记录，并通过适当的机制汇报至伦理风险管理委员会。记录与汇报的目的是为了在企业内部沟通相关伦理风险管理活动、为决策提供支持、进一步改善风险管理活动并协助相关部门管理伦理风险。

5.2.3对相关人员进行培训

企业可以对员工进行伦理风险管理制度的培训，记录每次培训的具体情况，包括参与人员、培训内容及培训效果。培训的具体内容应包括企业的伦理风险管理组织架构、企业内部相关政策及制度、企业所涉及的相关伦理风险、相关伦理风险评价原则及指标以及用以解决实际问题的实用建议及案例分析。

培训应采取适当形式进行有效的信息传递，如果涉及外籍员工，可以通过必要措施确保其了解培训内容。企业应对不同性质的员工进行针对性的培训，例如针对人工智能开发人员的专门培训以及针对伦理风险应对人员的培训的侧重点应有所不同。根据企业的规模和业务复杂性，企业可以制定适当的培训及措施，协助企业内部各个层级的员工准确理解并有效遵守伦理风险管理机制。

5.2.4定期进行风险评估

企业可以定期进行伦理风险评估，对企业总体可能涉及的伦理风险进行识别，评估一段时间内相关伦理风险应对措施是否适当且有效。企业可以根据相关法律法规以及政府及行业标准的规定，具体评估本企业所面临的风险，并记录评估内容及结果。企业可以视情况聘请外部专家参与评估，确保评估的有效性。企业可以采用以下具体手段进行评估：

（1）审阅企业现有政策及制度，评估其是否符合法律法规及行业标准的要求。

（2）对企业的相关人工智能开发进行抽样检测，评估其是否符合公司内部伦理风险应对的要求。

（3）对企业的内控制度进行评估。

（4）与企业内相关岗位的员工进行面对面访谈或发放调查问卷。

（5）形成书面总结报告及建议。

企业在实践中可以根据以上方面建立与自身业务相适应的内部机制，以应对人工智能开发与应用可能带来的伦理风险，并根据在内部制度推行以及业务发展过程中的实践经验，不断完善相关机制，以求有效地处理伦理风险问题。

第六章 结论

本研究针对人工智能引发的伦理及社会影响问题，从行业与国家的角度分析了人工智能伦理研究的现状，探究了国内外人工智能伦理的问题、挑战与发展趋势，剖析了人工智能技术及其应用产生的安全、公平、隐私保护及问责等问题。本研究不仅关注人工智能引发的短期伦理风险，更关注其带来的长期风险，并建议人工智能企业面对不同程度的风险采取不同的应对措施。此外，本研究提议通过设计相应的伦理原则与风险指标体系，对人工智能的进展进行常态化的持续监控，为人工智能的风险规避与应对提供灵活的监管框架、机制和措施保障，为人工智能伦理的行业实践提供初步应用指南与建议，推动人工智能技术及其应用的公平、透明、合乎伦理的良性发展。

鉴于人工智能在国家产业发展中的战略性地位和应对人工智能伦理风险的迫切需要，本报告提出如下建议：

1. 尽早启动我国人工智能发展规划向人工智能立法的转型。面对世界各国人工智能由规划调整向立法规制转向的趋势，我国人工智能立法亦应成为必然选择。我国可参照域外 AI 咨询委员会的模式，建立由政府部门和行业专家组成的人工智能伦理、协调机构，对人工智能的开发和应用提供伦理指引，并对具有重大公共影响的人工智能产品进行伦理与合法性评估。对人工智能带来的伦理、法律及社会问题进行深入调研和研究，这是人工智能立法进行的前置性步骤。

2. 推进不同层级的人工智能立法，引导人工智能健康发展。对于自动驾驶等技术发展已经相对成熟、产品亟待进入市场的应用领域，一方面应在现行法律中作出指引性规定，如在《侵权责任法》中增加人工智能致人损害时的责任分配规则，并尽快制定相关安全管理方面的法规，为法律的制定奠定基础。另一方面，可尝试进行更为具体的地方性、试验性的立法，为人工智能相关立法进程提供地方经验。当然，不仅应当完善相关法律法规，并结合不同的行业和地域特点出台相应的规则、判案标准等，而且还应当尽快将人工智能纳入网络安全法律监管体系内予以规制。

3. 应提出系统性、可操作性强的人工智能伦理风险评估指标体系或风险管

理指南，为人工智能相关企业提供风险识别、评估及应对的体系性指引。在此基础上，可将风险指南的贯彻落实与侵权免责的认定相关联，即在难以确定人工智能损害结果的责任承担主体时，将是否有相对完备的风险管理体系作为责任分配的考量因素，对于依照指南进行风险控制的企业，可在一定程度内予以减轻甚至免除责任。

4. 进一步推进人工智能标准化进程。目前，人工智能在相关领域的应用缺乏明确的技术规范和政策界限，基于未来的发展趋势，至少可以先从标准化入手，在某些领域尝试确立一些应用模式作为典型，以通过技术上的标准化来促成相关政策的出台，进而推动人工智能应用领域的不断拓展。此外，应以权利与责任相统一为原则，明确各方的安全保障义务，并建立相应的审核和认证体系。

5. 应促进数据共享技术，为人工智能培训和测试提供共享的公共数据集。在个人信息得到保护的前提下，促进数据的自由流通；并加强国际合作，建立多层次的国际人工智能治理机制。应通过联合国、G20 以及其他国际平台，将人工智能发展纳入国际合作议程，利用人工智能推动联合国 2030 年可持续发展目标的实现。

概言之，人工智能风险规制的复杂性和专业性，对产业政策及相关立法提出了更高的要求，因此，应尽可能让社会各界参与到这一进程中，共同探索人工智能风险规制的可行对策。权衡政府机构、社会公众、社会团体、产业界等可能受人工智能影响的群体的利益关切，坚持监管的专业性、谦抑性，把监管重心聚焦于技术发展导致的具体问题。在审慎推进人工智能技术创新的同时，确保技术不危及社会安全，避免公权力过早介入人工智能领域而对技术及产业发展造成阻碍，以开放、宽容的态度应对其伦理风险，着力创造促进人工智能发展和创新的有利环境。系统的公平、透明与合理，促进人工智能技术及其应用的健康、良性发展。

附录：国外有关人工智能基本原则的文献

1. 《ASILOMAR AI PRINCIPLES》（《阿西洛马人工智能原则》）
2. 《Statement on Algorithmic Transparency and Accountability》（《算法透明和可责性声明 》）
3. 《The Japanese Society for Artificial Intelligence Ethical Guidelines》（《日本人工智能学会伦理准则》）
4. 《Principles for the Governance of AI》（《人工智能治理原则》）
5. 《Montréal Declaration for Responsible AI draft principles》（《可靠的人工智能草案蒙特利尔宣言》）
6. 《TOP 10 PRINCIPLES FOR ETHICAL ARTIFICIAL INTELLIGENCE》（《人工智能伦理的十大原则》）
7. 《Artificial Intelligence——The Public Policy Opportunity》（《人工智能——公共政策的机遇》）
8. 《Partnership on Al to Benefit People and Society》（《福泽人类和社会的人工智能合作关系》）
9. 《The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE)》（《IEEE 人工智能设计的伦理准则》）
10. 《AI Code》（《人工智能伦理准则》）

一、《ASILOMAR AI PRINCIPLES》（《阿西洛马人工智能原则》）

发表日期：Jan 3-8, 2017

国家或地区：United States

发布者：Future of Life Institute (FLI), Beneficial AI 2017

发布者类型：Academia, Non-profits and Non-Governmental Organizations

缩写：FLI 2017

1. 安全性：人工智能系统应当是安全的，且是可适用的和可实行的。

2. 故障透明：如果一个人工智能系统引起损害，应该有办法查明原因。

3. 审判透明：在司法裁决中，但凡涉及自主研制系统，都应提供一个有说服力的解释，并由一个有能力胜任的人员进行审计。

4. 职责：高级人工智能系统的设计者和建设者是系统利用，滥用和行动的权益方，他们有责任和机会塑造这些道德含义。

5. 价值观一致：应该设计高度自主的人工智能系统，以确保其目标和行为在整个运行过程中与人类价值观相一致。

6. 人类价值观：人工智能系统的设计和运作应符合人类尊严，权利，自由和文化多样性的理念。

7. 个人隐私：既然人工智能系统能分析和利用数据，人们应该有权利存取，管理和控制他们产生的数据。

8. 自由与隐私：人工智能在个人数据的应用不能无理缩短人们的实际或感知的自由。

9. 共享利益：人工智能技术应该尽可能地使更多人受益和授权。

10. 共享繁荣：人工智能创造的经济繁荣应该广泛的共享，造福全人类。

11. 人类控制：人类应该选择如何以及是否代表人工智能做决策，用来实现人为目标。

12. 非颠覆：通过控制高级人工智能系统所实现的权力，应尊重和改善健康社会所基于的社会和公民进程，而不是颠覆它。

13. 人工智能军备竞赛：应该避免一个使用致命自主武器的军备竞赛。

二、《 Statement on Algorithmic Transparency and Accountability》（《算法透明和可责性声明 》）

发表日期：Jan 12, 2017

国家或地区：United States

发布者：ACM US Public Policy Council (USACM)

发布者类型：Academia, Non-profits and Non-Governmental Organizations

缩写：USACM 2017

1. 充分认识：算法分析系统的所有者、设计者、制造者、使用者和其他利益相关者，应当充分认识到算法歧视的可能性；算法歧视可能出现在算法的设计、运行和适用阶段。同时，也应当充分认识到算法歧视对个人和社会可能造成的危害。

2. 救济：某些个人和群体可能受到算法决策的不利影响。监管者应当建立健全救济机制，允许上述个人和群体对算法决策结果提出质疑、获得救济。

3. 可责性：即使无法解释算法产生的结果，利用算法进行决策的部门也应当对其决策负责。

4. 解释：鼓励利用算法进行的决策的组织和机构，对算法决策遵循的程序和具体决策结果作出解释；这一原则在公共政策决策中尤为重要。

5. 数据来源：算法的制造者应当保留一份关于训练数据来源的描述，同时应当附带一份说明，阐明在收集数据（人工或算法收集）中的潜在歧视风险。公共审查能提高算法被纠正的可能性。但是以上在下列几种情况中，可以不公开数据来源，而仅对符合标准且得到授权的部分人公开数据：（1）涉及隐私问题；（2）涉及商业机密；（3）公开数据来源可能导致恶意第三人蓄意（利用输入数据）使系统产生偏差的。

6. 可审查性：模型、算法、数据和决策应当留存记录，以便在怀疑其导致损害结果产生的情况下进行审查。

7. 验证和测试：相关机构应当对其模型的有效性进行严格的验证，并记录验证方式与验证结果。特别的，应当对算法进行日常测试，以确定算法最严重的歧视性问题出现在何处。鼓励机构公开此类测试的测试结果。

三、《The Japanese Society for Artificial Intelligence Ethical Guidelines》（《日本人工智能学会伦理准则》）

发表日期：Feb 28, 2017

国家或地区：Japan

发布者：The Japanese Society for Artificial Intelligence (JSAI)

发布者类型：Academia, Non-profits and Non-Governmental Organizations

缩写：JSAI 2017

1. 贡献人类：日本人工智能学会会员应当为全人类的和平、安全、福利和共同利益做出贡献，保护基本人权，尊重文化多样性。作为人工智能专家，在设计、研发和使用人工智能的过程中，会员应当消除人工智能对人类安全的威胁。

2. 遵守法律法规：日本人工智能学会会员应当遵守有关研发和知识产权的法律法规和契约性协议。会员不得侵犯他人所有的信息与财产。不论直接或间接，会员不得以伤害他人为目的使用人工智能。

3. 尊重隐私：在研发人工智能的过程中，日本人工智能学会会员应当尊重他人隐私。会员应当合理地对待他人隐私，并遵守相关法律法规。

4. 公正：日本人工智能学会会员应当永远保持公正。会员应当承认，人工智能可能会导致此前并不存在的不平等与歧视现象。会员在研发人工智能时应避免歧视。会员应尽最大努力，保证其研发的人工智能可为人类公平平地使用。

5. 安全：作为人工智能专家，日本人工智能学会会员应当认识到对安全性的迫切需求，并承担保证人工智能可控的责任。在研发和使用人工智能的过程中，日本人工智能学会会员应当时刻注意人工智能的可控性、安全性和必要保密性；同时，会员应保证用户得到了合理、充分的信息。

6. 秉直行事：日本人工智能学会会员应当认识到人工智能可能对社会带来的巨大影响；因此，会员将永远秉直行事、赢得社会信任。作为人工智能专家，会员应当避免发表错误或模糊的意见；同时，会员负有忠实义务，应诚实、充分、合理地解释人工智能系统中存在的问题与技术局限。

7. 可责性与社会责任：日本人工智能学会会员应当核实其研发的人工智能

的表现与影响。如果在审核中发现潜在危险，应该迅速对全社会发布警告。会员应当意识到，尽管违背其主观意愿，但其研发的技术可能被用于伤害他人；会员应当努力阻止此类问题发生。如果有人发现并报告人工智能的如上错用，应当保护报告者不受损失。

8. 社会沟通和自我发展：日本人工智能学会会员应当以增进社会对人工智能的了解为目标。会员应当认识到，对人工智能，社会上存在着不同观点，会员应从其中学习。这些观点能帮助会员更深入地理解社会，并与社会保持持续、有效的联系，这将有助于人类的和平与幸福。作为人工智能专家，会员应当不断发展自我，并帮助“同行者”追逐共同的目标。

9. 人工智能伦理准则：与日本人工智能学会会员相同，人工智能也应当遵循上述伦理准则，以便人工智能成为学会的“准会员”。

四、《Principles for the Governance of AI》（《人工智能治理原则》）

发表日期：Oct 3, 2017 (unconfirmed)

国家或地区：United States

发布者：The Future Society, Science, Law and Society (SLS) Initiative

发布者类型：Academia, Non-profits and Non-Governmental Organizations

缩写：The Future Society 2017

1. 人工智能应该促进人类，其社会及其自然环境的福祉。对幸福的追求似乎是一种不言自明的愿望，但鉴于人工智能和混合智能系统滥用的日益普遍性、威力和风险，这是一项特别重要的基本原则。在使法律程序的中央实况调查任务更加有效和高效的同时，熟练设计和执行混合情报程序可减少判定有罪或无罪的错误，加速解决争端，并为那些缺钱的人提供司法的可能性。

2. 人工智能应该是透明的。透明度是能够追踪算法决策途径中的因果关系，并且在混合智能系统中追踪其操作员。例如，在发现过程中，这可能会延伸到选择用于训练预测编码软件的数据，保留用于设计和执行自动审查过程的专家的选择或用于确认的质量保证协议准确性。法院和法律智库通常倾向于在这方面提供大量的文件和披露。国际标准化组织在其电子发现行为准则中提出如下：“透明度意味着该流程易于审计，无论是参与流程的一方还是第三方，并且整个流程中的因果关系都很容易被看到并被理解。“另外，ISO 建议维护有关程序，决定和评估的”完整记录“。

3. 人工智能的制造商和运营商应该负责问责制意味着能够为人工智能或其运营商造成的影响分配责任。法院有权采取纠正措施或制裁为了使人工智能弄错事实调查任务而故意进行操作的人。

4. 人工智能的有效性应该可以在其预期的实际应用中测量。可测量性意味着专家用户和普通公民能够具体衡量人工智能或混合智能系统是否达到其目标。我们很少有人能够理解发现中使用的人工智能算法，或科学专家需要有效操作的程序。但每个人都可以理解，一个达到 80％准确度的系统要优于 50％准确度的系统。

5. 人工智能系统的运营商应具备适当的能力。如果 Netflix 的算法在星期六晚上推荐了错误的喜剧，我们都不会受到伤害。但是当我们的健康，我们的权利，我们的生活或我们的自由依赖于混合智能时，这些系统应该由具有必要专业知识的专业人员设计，执行和测量。ABA 于 2012 年通过了对“职业行为规范”的修订，其中包括对规则 1.1 关于权限的新评论。它规定：“律师应该及时了解法律和实践的变化，包括与相关技术相关的利益和风险。”ISO 进一步承认，文件审查“基本上是一种信息检索练习”，并且它必须因此需要借鉴信息检索科学（计算机科学，统计学，语言学等）的专业知识“.ISO 的决心无疑将加速文件审查和诉讼分析的专业化，并迫使美国律师协会，法院和保险公司考虑这一决心的道德和责任影响。

6. 通过与民间社会进行深思熟虑的包容性对话，应将编入人工智能系统决策的准则编纂成文。在大多数情况下，人工智能的可接受用途的编纂仍然是技术精英的领域，立法者，法院和政府正努力赶上现实的发展的领域，而普通公民仍然被排除在外。人工智能在发现中的采用反映了一个广泛的，精巧的咨询过程。它几乎涉及到所有利益相关方：法院，法律从业人员，科学家，人工智能和混合智能解决方案的提供者，还包括伦理学家，学者，广泛包容性和审议性的智囊团，以及成千上万的公民和公司，每一天，在国家的每一个角落，在现实世界的诉讼和纠纷解决中。有关在电子发现中使用人工智能的社会对话将受益于更具包容性，更多论坛寻求政治科学家，社会学家，哲学家和代表性群体的普通公民的积极参与。即便如此，电子发现的领域为有关包容性对话如何能够在确保人工智能系统在重要社会功能中有益使用方面产生广泛共识提供了一个有希望的例子。

五、《 Montréal Declaration for Responsible AI draft principles》（《可靠的人工智能草案蒙特利尔宣言》）

发表日期：Dec 4, 2018

国家或地区：Canada

发布者：University of Montreal

发布者类型：Academia, Non-profits and Non-Governmental Organizations

缩写：Montreal 2018

1. 福祉。人工智能的发展最终应该增进所有有情生物的福祉。

2. 自主。人工智能的发展应该促进所有人的自主权，并以负责任的方式控制计算机系统的自主性。

3. 正义。人工智能的发展应该促进公平正义，并设法消除所有类型的歧视，特别是与性别，年龄，精神/身体能力，性取向，族裔/社会起源和宗教信仰有关的歧视。人工智能的发展应该促进正义，并设法消除所有类型的歧视，特别是与性别，年龄，精神/身体能力，性取向，族裔/社会起源和宗教信仰有关的歧视。

4. 隐私。人工智能的发展应该保证尊重个人隐私，并允许使用者访问他们的个人数据以及任何算法可能使用的各种信息。

5. 知识。人工智能的发展应该促进批判性思维的发展，保护人们免受宣传和操纵之苦。

6. 民主。人工智能的发展应促进人们有准备地参与公共生活，合作和民主辩论。

7. 责任。人工智能发展的各方参与者应该承担起自己的责任，抵御技术创新带来的风险。

六、《TOP 10 PRINCIPLES FOR ETHICAL ARTIFICIAL INTELLIGENCE》（《人工智能伦理的十大原则》）

发表日期：Dec 11, 2017

国家或地区：International

发布者：UNI Global Union

发布者类型：Academia, Non-profits and Non-Governmental Organizations

缩写：UNI Global Union 2017

1. 要求人工智能系统透明

一个透明的人工智能系统就是一个可以发现系统如何以及为什么做出决定的系统，就机器人而言，可以研究机器人的行为方式。尤其是：

A. 我们强调开源代码对于透明度既不必要也不足够——清晰度不会因复杂性而被混淆。

B. 对于用户来说，透明性非常重要，因为透过为用户提供一种简单的方式来了解系统正在做什么以及为什么这样做，它建立了对系统的信任和理解。

C. 对于人工智能系统的验证和认证，透明度非常重要，因为它暴露了系统的审查流程。

D. 如果发生事故，人工智能将需要对事故调查员保持透明和负责，因此可以理解导致事故的内部过程。

E. 工人必须有权要求人工智能系统的决策和结果以及底层算法的透明度（见下面的原则 4）。这包括对人工智能/算法做出的决定提出上诉并由人进行审查的权利。

F. 必须就人工智能系统的实施、开发和部署咨询工作人员。

G. 事故发生后，参与审判过程的法官、陪审团、律师和专家证人需要为证据和决策提供信息的透明度和问责制。

透明度的原则是保证下述原则能够遵守的前提条件。

有关可操作性的解决方案，请参阅以下原则 2。

2. 使用“道德黑匣子”装备人工智能系统

人工智能系统中的完全透明度应该通过存在一种设备来实现，该设备可以以

“道德黑匣子”的形式记录关于该系统的信息，该设备不仅包含相关数据以确保系统的透明度和问责性，还包括清楚了解所述系统中建立的道德考虑的数据和信息。

适用于机器人的道德黑匣子将记录所有决策，这是机器人主机的决策、动作和感官数据的基础。黑盒子提供的数据还可以帮助机器人将他们的动作解释为人类用户可以理解的语言，促进更好的关系并改善用户体验。道德黑匣子的读取应该简单快速。

3. 让人工智能服务人与地球

这包括人工智能的开发、应用和使用的道德准则，以便在其整个运作过程中，人工智能系统保持兼容，并增加人的尊严，完整性，自由，隐私和文化和性别多样性的原则，以及基本人权。此外，人工智能系统必须保护甚至改善我们星球的生态系统和生物多样性。

4. 采用人为命令的方法

绝对的先决条件是人工智能的发展必须是负责任的、安全的和有用的。机器保持着工具的法律地位，法人依然一直控制和负责这些机器。

这意味着人工智能系统的设计和运行应遵守现行法律，包括隐私。考虑到系统分析和利用这些数据的能力，工人应有权访问、管理和控制人工智能系统产生的数据（参见“工人数据隐私和保护的十大原则”中的原则 1）。当人工智能系统用于人力资源程序时，例如招聘，晋升或解雇，工人也必须具有“解释权”。

5. 保证一个无性别偏见的人工智能

在人工智能的设计和维护中，系统受到负面或有害的人类偏见的控制至关重要，任何偏见（无论是性别，种族，性取向，年龄等）都被识别出来并且不会被系统传播。

6. 分享人工智能系统的益处

人工智能技术应该尽可能多地使人们受益并赋予人们权力。人工智能创造的经济繁荣应该广泛而平等地分配，以造福全人类。因此，全球性的旨在弥合经济，技术和社会数字鸿沟的国家政策是必要的。

7. 确保公平转型并确保对基本自由和权利的支持

随着人工智能系统的发展和增强现实的形成，工作人员和工作任务将被取

代。为了确保公平的过渡以及未来的可持续发展，至关重要的是公司制定政策，确保与这种流离失所有关的企业问责制，如再培训计划和改变工作机会的可能性。还需要政府采取措施帮助失业工人重新培训并找到新的工作。

人工智能系统与更广泛的数字经济转型相结合将需要工人在各个层面和各个职业都有机会获得社会保障，并持续终身学习以保持就业。各国和各公司有责任找到为所有工作形式的所有工作者提供权利解决方案。

此外，在一个临时工或个人化工作不断增加的世界里，所有工作形式的工人都必须具有相同的，强大的社会和基本权利。所有人工智能系统必须在其部署和增强与人权法律，国际劳工组织公约和集体协议中规定的工人权利相协调的情况下进行检查和平衡。反映系统内置的核心国际劳工组织公约 87 和 98 的算法“8798”可以达到这个目的。发生故障时，系统必须关闭。

8. 建立全球性的管理机制

UNI 建议在全球和地区层面建立多方利益相关者体面工作和道德人工智能的管理机构。这些机构应该包括人工智能设计人员、制造商、业主、开发人员、研究人员、雇主、律师、公民社会组织和工会。必须建立举报机制和监测程序，以确保过渡到和实施道德的人工智能。这些机构应被授予建议合规程序和程序的权限

9. 禁止机器人的责任分配

机器人应尽可能以遵守现行法律，基本权利和自由（包括隐私）来设计和运行，这与法律责任问题有关。根据 Bryson 等人，2011 年 UNI 全球联盟声称机器人的法律责任应该归功于一个人。法律规定机器人不是责任方。

10. 禁止人工智能装备竞赛

应禁止包括网络战争在内的致命自主武器。

七、《Artificial Intelligence—The Public Policy Opportunity》（《人工智能—公共政策的机遇》）

发表日期: Oct 18, 2017

国家或地区: United States

发布者: Intel

发布者类型: Industry

缩写: Intel 2017

1. 促进创新和开放式发展：为了更好地理解人工智能的影响并探索人工智能实施的广泛多样性，公共政策应鼓励人工智能研发投资。各国政府应支持人工智能系统的受控测试，以帮助行业，学术界和其他利益相关者改进技术。

2. 创造新的就业机会并保护人们的福利：人工智能将改变人们的工作方式。支持增加劳动力技能和促进不同部门就业的公共政策应该增加就业机会，同时保护人们的福利。

3. 负责任地解放数据：人工智能通过访问数据提供支持。机器通过随时间分析更多数据学习算法，而得到改进的数据访问对于实现更加强化的人工智能模型的开发和培训至关重要。消除数据访问障碍将有助于机器学习和深度学习充分发挥其潜力。

4. 反思隐私问题：像公平信息实践原则和隐私设计这样的隐私方法经受住了时间的考验和新技术的发展。但是随着创新，我们不得不“重新思考”我们如何将这些模型应用于新技术。

5.对进行道德设计和实施的责任的要求：随着更多人获得人工智能的实施，计算的社会影响不断扩大并将继续扩大。公共政策应致力于识别和减轻因使用人工授精引起的歧视，并鼓励设计这些危害的避免措施。

八、《Partnership on Al to Benefit People and Society》（《福泽人类和社会的人工智能合作关系》）

发表日期: Sep 28, 2016 (unconfirmed)

国家或地区: United States

发布者: Partnership on AI

发布者类型: Academia, Non-profits and Non-Governmental Organizations

缩写: PAI 2016

1. 我们将努力确保人工智能技术造福于尽可能多的人。

2. 我们将教育公众并倾听他们的意见，积极地与利益相关者接触，寻求他们对我们的关注问题的反馈，告知他们我们的工作，并解决他们的问题。

3. 我们致力于就人工智能的伦理、社会、经济和法律问题展开研究和对话。

4. 我们认为，人工智能研究和发展需要广泛的利益相关者积极参与并对他们负责。

5. 我们将与商界的利益相关者代表进行接触，以确保对特定领域的关注和机会得到理解和解决。

6. 我们将努力最大限度地利用人工智能技术带来的好处并解决潜在的挑战，通过以下几种方式:

A. 致力于保护个人的隐私和安全。

B. 努力理解和尊重所有可能受到人工智能发展影响的各方的利益。

C. 努力确保进行人工智能研究的社会团体对社会负责、敏感，并直接参与人工智能技术对更广泛社会的潜在影响。

D. 确保人工智能的研究和技术是强健的、可靠的、值得信任的，并在安全的约束下运行。

E. 反对发展和使用违反国际公约或人权的人工智能技术，并促进无害的保障和技术。

F. 我们认为，为了能够解释技术，人工智能系统的运作可以被人们理解和解释是很重要的。

G. 我们努力在人工智能科学家和工程师之间创造一种合作、信任和开放的

文化，帮助我们更好地实现这些目标。

九、《The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE)》（《IEEE 人工智能设计的伦理准则》）

发表日期: (v1) Dec 13, 2016. (v2) Dec 12, 2017

国家或地区: International

发布者: The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

发布者类型: Academia, Non-profits and Non-Governmental Organizations

缩写: IEEE 2017

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

十、 《AI Code》（《人工智能伦理准则》）

发表日期：Apr 16, 2018

国家或地区：United Kingdom

发布者：House of Lords, Select Committee on Artificial Intelligence

发布者类型：Governments

缩写：House of Lords 2018

1.人工智能应服务于人类的共同利益和福祉。

2.人工智能应遵循可理解性和公平性原则。

3.人工智能不应被用于削弱个人、家庭和社区的数据权利或隐私。

4. 所有公民都有权利接收相应的教育以便能够在精神、情感和经济上适应人工智能的发展。

5. 人工智能绝不应被赋予任何伤害、毁灭或欺骗人类的自主能力。