腾讯研究院
Tencent
Research Institute

Tencent
AI Lab


# 智能时代的技术伦理观——重塑数字社会的信任

腾讯研究院 腾讯 AI Lab

2019 年 6 月

# 执行摘要

随着人工智能的发展应用，智能时代的大幕正在拉开，无处不在的数据和算法正在催生一种新型的人工智能驱动的经济和社会形式。与此同时，隐私保护、虚假信息、网络安全、网络犯罪、电子产品过度使用等问题成为全球关注焦点。旨在对科技行业的狭隘的技术向度和利益局限进行纠偏和矫正的人工智能伦理开始从幕后走到前台，正如华裔 AI 科学家李飞飞所言，要让伦理成为人工智能研究与发展的根本组成部分。在此背景下，从政府到行业再到学术界，全球掀起了一股探索制定人工智能伦理原则的热潮，欧盟、德国、英国、OECD、G20、IEEE、谷歌、微软等诸多主体从各自的角度提出了相应的人工智能伦理原则，共同促进 AI 知识的共享和可信 AI 的构建。要言之，各界已经基本达成共识，人工智能的发展离不开对伦理的思考和伦理保障。

从最早的计算机到后来的信息再到如今的数据和算法，伴随着技术伦理的关注焦点的转变，技术伦理正在迈向一个新的阶段。为此，我们提出新的技术伦理，探索技术、个人、社会三者之间的平衡。包含三个层面：技术信任，人工智能等新技术需要价值引导，做到可用、可靠、可知、可控（“四可”）；个体幸福，确保人人都有追求数字福祉、幸福工作的权利，在人机共生的智能社会实现个体更自由、智慧、幸福的发展；社会可持续，践行“科技向善”，发挥好人工智能等新技术的巨大“向善”潜力，善用技术塑造健康包容可持续的智慧社会。

# 目录

# 引言：智能时代迫切需要面向数据和算法的技术伦理

## （一）人工智能引领第四次工业革命加速变革经济社会

互联网的前三十年，是一个连接一切的故事，从连接人到连接商业再到连接万事万物，互联网技术以前所未有的速度和规模改变着人类和人类社会。截至 2019 年 3 月 31 日，全球网民规模达到 43.8 亿，占世界总人口 56.8%。中国以逾 8 亿网民规模成为全球最大互联网市场，个人应用、产业应用、政府应用等各类互联网应用蓬勃发展，为人们创造着效率和便利。连接带来的创新，提升了个体生活，促进了经济发展与社会进步。在这一发展过程中，互联网行业始终引领创新发展的浪潮。

再者，人工智能（包括未来可能出现的强人工智能和超人工智能）、机器人、大数据、物联网、区块链、云计算、虚拟现实、基因编辑、脑机接口、3D 打印等新技术集群加速涌现和发展，有望引领第四次工业革命。尤其是人工智能，有望像历史上的火种和电力一样，重塑人类生活和人类社会的未来——加速的自动化和智能化，无处不在的连接，物理与数字世界的融合，甚至人类与技术的融合。我们正在步入高度依赖技术的社会，生物层、物理层、技术层有可能融合成为三位一体。未来，包括企业在内的所有组织机构都会数字化、智能化。

智能时代的大幕正在拉开，在其中，无处不在的数据和算法正在催生一种新型的人工智能驱动的经济和社会形式。人们常以“数据化”

“智能化”“算法决定论”等词汇描述这一趋势，但核心都是一样：人工智能算法正在改变这个世界。比如，算法已在决定向你展示的广告、新闻资讯等数字内容，可以评估你能否得到面试机会、获得贷款、拿到救助金等，还能预测犯罪、犯罪人的危险性等。当然，人工智能的应用并不局限于这些方面。基于大量的数据，人工智能有潜力做出比人类更优的预测和决策，比如自动驾驶有望比人类驾驶更安全，智能医疗影像诊断比医生的诊断结果更准确，智能语音识别的出错率也比速记员更低。人工智能是一项通用技术，只要有数据，就有望普遍应用于各行各业，进而提高生产力并促进经济增长。

### （二）人工智能的潜力和价值巨大，但也带来需要积极应对的问题和挑战

以人工智能为代表的这一轮新技术无疑拥有巨大的潜力和价值，值得在研发和应用上持续投入，但任何有望变革人类社会的新技术都必然会带来社会伦理影响。例如，互联网技术带来的用户隐私、虚假信息、算法“黑盒”、网络犯罪、电子产品过度使用等问题已经成为全球关注焦点，引发全球范围内对互联网技术及其影响的反思和讨论，探索如何让新技术带来个人和社会福祉的最大化。此外，2018 年底发生的基因编辑婴儿事件在国内外引发激烈的伦理争议。人工智能也是如此，其在隐私、歧视、安全、责任、就业等经济、伦理和社会方面的问题正在显现。未来可能出现的通用人工智能和超级人工智能则可能带来更深远而广泛的安全、伦理等影响。

隐私方面，Facebook-剑桥分析数据丑闻引爆了社会对数据的泄露、不正当二次利用、滥用等问题的担忧，而这些问题已是现今互联网用户普遍面临的问题——很多应用软件会过度收集非必需的用户数据、冗长的“用户须知”或“使用协议”诱导用户授权厂商过度使用甚至出售用户数据等。此外，不法分子利用人工智能技术，可以非法窃取、识别个人信息，甚至影响、操纵用户行为和认知。

歧视方面，“大数据杀熟”成为 2018 年度社会生活类十大流行语之一，反映了人们对算法歧视的担忧，滴滴、携程、飞猪等都被质疑大数据杀熟。此外，语音识别、人脸识别、精准广告工具等算法应用都可能存在偏见和歧视，甚至还曾引发了诉讼。

安全方面，自动驾驶汽车的安全问题尤其引人关注：Uber 的自动驾驶汽车在测试过程中撞死行人、特斯拉的 Autopilot 系统在运行过程中造成致命事故。人工智能在事关健康安全的医疗领域也出现过问题，比如 IBM 的“沃森医生”给出过“不安全且错误”的癌症治疗建议。除了人工智能系统本身可能不完善的问题，人工智能还面临着被外部攻击风险，比如图像识别和语音识别技术的对抗攻击问题一直都是相关应用头顶上的一片乌云，自动驾驶汽车和智能家居等连接网络的物理设备也存在被网络远程干扰或操控的风险。此外，人工智能还可能被用作攻击手段——比如用于生成虚假视频和图像（所谓的 deepfake 即深度伪造），生产和传播假新闻，甚至被用于攻击人工智能系统。

责任方面，人工智能算法由于可能存在不透明、不可理解、不可解释等特性，在事故责任认定和分配上也存在有待讨论的法律难题，比如由于行人横穿公路而造成交通事故时自动驾驶汽车运营方是否应该承担责任、智能监控摄像头厂商是否应该为摄像头被网络攻击的问题负责。

就业方面，人工智能可能造成大规模失业的风险一直备受社会关注，甚至有人认为人工智能的普及将会在人类社会中产生一批史无前例的“无用阶层”；但同时也有人认为，与过去的蒸汽机和计算机等技术一样，新技术在夺走一部分工作岗位的同时也会创造更多更好的新型工作岗位。所以，人工智能对劳动力市场的长远影响究竟是积极的、消极的还是中立的，目前还很难预测。尽管如此，已经有一些政府机构和研究者在思考和探索潜在的解决方案了，比如全民基本收入以及教育改革等。

### （三）各界高度重视人工智能伦理，多举措促进新技术健康发展

总体来看，我们现在就有必要对人工智能等新技术进行更多的人文和伦理思考，正如华裔 AI 科学家李飞飞所言，要让伦理成为人工智能研究与发展的根本组成部分。因为正如基辛格所言，面对人工智能的兴起，人们在哲学、伦理、法律、制度、理智等各方面都还没做好准备，因为人工智能等技术变革正在冲击既有的世界秩序，我们却

# 每日报告

不要错过让你洞察整个商业世界的每日报告

如何免费入群？扫码加好友后回复【入群】

每日精选3份最值得学习的资料给您，不定期分享顶级外文期刊

撩他！ 撩他！

无法完全预料这些技术的影响，而且这些技术可能最终会导致我们的世界所依赖的各种机器为数据和算法所驱动且不受伦理或哲学规范约束。

显然，在当前的人工智能等新技术背景下，我们比历史上任何时候都更加需要“科技向善”理念，更加需要技术与伦理的平衡，以便确保新技术朝着更加有利于人类和人类社会的方向发展。一方面，技术意味着速度和效率，要发挥好技术的无限潜力，善用技术追求效率，创造社会和经济效益。另一方面，人性意味着深度和价值，要追求人性，维护人类价值和自我实现，避免技术发展和应用突破人类伦理底线。因此，只有保持警醒和敬畏，在以效率为准绳的“技术算法”和以伦理为准绳的“人性算法”之间实现平衡，才能确保“科技向善”。

因此，对伦理的强调和重视成为了当前人工智能领域的一大主旋律，社会各界纷纷制定相应的伦理准则或框架。例如，德国于 2017 年为自动驾驶汽车提出了 20 条伦理原则。英国已经成立了数据伦理中心，视伦理为人工智能创新的核心之一并考虑制定普适的人工智能伦理框架。欧盟人工智能战略的三大支柱之一即是确保欧盟具有与人工智能发展和应用相适应的法律和伦理框架，为此欧盟委员会已经起草人工智能伦理指南。中国的人工智能顶层政策要求制定促进人工智能发展的法律法规和伦理规范。美国、加拿大、新加坡、印度、法国、意大利等国家和地区都有类似的规划或政策。在国际层面，2019 年 5 月 22 日，OECD 成员国批准了人工智能原则即《负责任地管理可信赖

的 AI 的原则》，该伦理原则总共有五项，包括包容性增长、可持续发展和福祉，以人为本的价值和公平，透明性和可解释，稳健性和安全可靠，以及责任。2019 年 6 月 9 日，G20 批准了以人为本的 AI 原则，主要内容来源于 OECD 人工智能原则，相当于为 OECD 人工智能原则背书。这是首个由各国政府签署的 AI 原则，有望成为今后的国际标准，旨在以兼具实用性和灵活性的标准和敏捷灵活的治理方式推动人工智能发展。

以欧盟为例，2019 年 4 月欧盟发布《可信 AI 伦理指南》(Ethics Guidelines for Trustworthy AI，以下称“伦理指南”)，提出了可信 AI 框架，包含三个层次：一是可信 AI 的根基，从基本权利（尊重人类尊严，个体自由，尊重民主、正义和法治，平等、非歧视和团结，公民权利）出发，提出 AI 必须遵循的四个伦理原则，即 AI 必须尊重人类自主性，必须防止造成损害或者不利地影响人类，必须确保公平，必须透明（针对 AI 的能力和目的）、可解释（针对 AI 作出的决定）。二是可信 AI 的实现，从七项关键要求来衡量 AI 是否可信，即人类能动性和监督，技术稳健性和安全（包括安全能经受攻击，后备计划和一般安全，准确性，可靠性和再生性），隐私和数据治理（包括尊重隐私，数据质量和完整，数据访问），透明性（包括可追溯，可解释，信息透明），多样性、非歧视和公平（包括避免不公平的偏见，普遍可用的设计，利益攸关方的参与），社会和环境福祉，问责（包括可审计，负面影响最小化及报告，权衡和救济）。三是可信 AI 的评估，基于前

述 7 项关键要求,《伦理指南》提出了试点性的可信 AI 评估清单。评估清单的目的在于为具体落实这些关键要求提供指引,帮助公司或组织内不同层级如管理层、法务部门、研发部门、质量控制部门、HR、采购、日常运营等共同确保可信 AI 的实现。欧盟委员会鼓励所有利益攸关方落实这七项关键要求。

在业内,电气电子工程师学会(IEEE)已在推进制定人工智能伦理标准(即 IEEE P7000 系列标准)和认证框架,AI 白皮书《合伦理设计》(Ethically Aligned Design)提出了八项基本原则。谷歌、微软、Facebook、DeepMind 等科技公司也多举措推进人工智能伦理研究,包括发起成立行业组织(比如 Partnership on AI)、成立伦理部门(比如 DeepMind 的伦理与社会部门,谷歌和微软的伦理委员会)、提出人工智能伦理原则(比如谷歌的 7 条正面原则,包括 AI 应对社会有益,AI 应避免造成或加剧不公平歧视,AI 应安全可靠,AI 应对人们负责,AI 应融入隐私设计原则,AI 应维持高标准的学术卓越,AI 应按照这些原则来使用;以及 4 条底线涉及谷歌不从事的 AI 应用,包括可能造成普遍伤害的技术,造成或直接促成人员伤亡的武器或其他技术,违反国际准则的监控技术,与国家法和人权原则相悖的技术。微软的六大人工智能原则,公平:AI 系统应公平对待每一个人,可靠:AI 系统应可靠、安全地运行,隐私安全:AI 系统应是安全的并尊重隐私,包容:AI 系统应赋能每一个人并让人们参与,透明:AI 系统应是可以理解的,责任:设计、应用 AI 系统的人应对其系统的运行

负责）。

要言之，各界已经基本达成共识，人工智能的发展离不开对伦理的思考和伦理保障，以让人工智能的发展和应用能够遵循负责任的、安全的、普惠的实践道路。因为我们正处在一个正在实现数据化和智能化的时代，所以当前强调的技术伦理主要面向数据和算法。在计算机与信息技术的发展史上，技术伦理经历了三次重大转变，每个阶段都有其特殊的伦理问题并产生了相应的法律规制。第一阶段的关注焦点是计算机，各国围绕计算机的安全、犯罪、欺诈、滥用等问题制定了一系列法律，这个阶段的典型立法包括美国 1984 年的《计算机欺诈与滥用法案》等。互联网兴起之后，技术伦理发展到第二阶段，信息大爆炸趋势下，信息成为关注焦点，法律规制围绕信息的隐私、保护、传播、滥用等问题展开，这个阶段的立法包括欧盟 1995 年的《个人数据保护指令》、美国 1996 年的《通信规范法》等。当前，技术伦理已经发展到了第三阶段，作为关注焦点的数据和人工智能算法带来新的伦理问题，预计将出现一系列人工智能法律，例如，欧盟的 GDPR 已经针对人工智能应用作出了一些制度安排，欧盟议会发布的《算法责任与透明治理框架》则在考虑建立算法治理框架。

### （四）探索智能时代的技术伦理观：信任、幸福与可持续

在人工智能伦理方面，腾讯公司董事会主席兼首席执行官马化腾在上海“2018 世界人工智能大会”上发表演讲时指出，我们需要充分

考虑未来人工智能发展可能带来的社会影响，并从问题的角度思考未来人工智能如何做到可用、可靠、可知、可控。2019 年 3 月，马化腾以全国人大代表的身份向全国人大提交建议案《关于加强科技伦理建设 践行科技向善理念的建议》。马化腾认为，科技伦理是创新驱动发展战略、数字中国建设、数字时代商业竞争的重要保障。他呼吁在全社会、全行业积极倡导“科技向善”“负责任创新”“创新与伦理并重”等理念。并建议加强科技伦理的制度化建设，针对相关新技术制定伦理准则并积极参与、推动新技术领域的全球治理；加快研究数据、人工智能、基因编辑等新兴技术领域的法律规则问题；加强科技伦理的教育宣传并鼓励全社会践行“科技向善”理念。他在接受媒体采访时表示，对于个人数据利用与隐私保护、个性化推荐算法的规制、AI、基因编辑等科技伦理、机器人的责任等前沿热点问题，需要公众、学界、业界、监管部门等社会各界积极进行交叉学科、跨界的研究和研讨，但对前沿技术的治理需要避免两个极端：放任不管和过早过度监管。

以此为基础，在新的发展阶段，我们提出新的技术伦理（technology ethics），探索技术、个人、社会三者之间的平衡。就 AI 技术自身而言，AI 需要价值引导，应做到可用、可靠、可知、可控，从而让人们可以信任 AI，让 AI 可以给个人和社会创造价值。就 AI 与个人之关系而言，幸福是人生的终极目的，需要构建和谐的人机关系，保障个人的数字福祉和幸福工作权利，实现智能社会人机共生，

让个体更自由、智慧、幸福地生活和发展。就 AI 与社会之关系而言，AI 可以成为一股“向善”的力量，发挥出巨大的“向善”潜力，应当鼓励社会各界践行“科技向善”，助力经济社会健康包容可持续发展。

## 一、信任（trust）：人工智能需要价值引导

如前所述，互联网在过去二十多年中持续带来的问题，连同最近的人工智能等新技术带来的伦理和社会影响，归根结底是技术信任方面的挑战。在人类的交往活动中，社会习俗、伦理道德、法律规范等构建起来的信任机制，构成人类合作、交易及表达不同意见的基础，可以让我们信任陌生人和陌生事物。我们常常将自己及至亲的生命财产安全托付给陌生人（比如乘坐公共交通工具以及将钱存入银行），正是由于这一社会信任机制的存在，而非相信“人性本善”。

那么，如何让人们同等地信任人工智能呢？现在人们无法完全信任人工智能，一方面是因为人们缺乏足够信息，对这些与我们的生活和生产息息相关的技术发展缺少足够的了解；另一方面是因为人们缺乏预见能力，既无法预料企业会拿自己的数据做什么，也无法预测人工智能系统的行为。此外，AI 本身不够可靠、AI 的可解释性不够、AI 提供商不值得信任等也是造成人们不信任 AI 的主要理由。所以，当前迫切需要塑造包括 AI 信任在内的数字信任，一般认为，数字信任体现在四个维度：一是安全的维度，产品服务安全可靠，包括网络安全、个人隐私安全等；二是透明的维度，保障用户的参与和知情；

三是责任的维度，确保相关主体负责任地提供产品服务，并为其行为承担责任；四是伦理的维度，秉持正确的价值观，不作恶。

就 AI 而言，虽然技术自身没有道德、伦理的品质，但是开发、使用技术的人会赋予其伦理价值，因为基于数据做决策的软件是人设计的，他们设计模型、选择数据并赋予数据意义，从而影响我们的行为。所以，这些代码并非价值中立，其中包括了太多关于我们的现在和未来的决定。而在技术与社会生态系统互动的过程中，技术发展常常具有环境的、社会的和人类的影响，这些影响远远超过了技术设备和实践自身的直接目的。因此，我们需要构建能够让社会公众信任人工智能等新技术的规制体系，让技术接受价值引导。

作为建立技术信任的起点，我们认为，人工智能等新技术的发展和应用需要遵循伦理原则。为此，秉持“负责任研究与创新”（responsible research and innovation）、“科技向善”等理念，我们进一步阐述“四可”原则，用以引导负责任地发展和应用工智能技术，使其可以造福于人类和人类社会。并将“四可”翻译为“ARCC”（available，reliable，comprehensible，controllable，即 ARCC，读作 ark）。正如传说中保存人类文明火种的诺亚方舟，人工智能的健康发展需要以“伦理方舟”为保障，确保将来友好、和谐、共生的人机关系。

## （一）以人工智能“四可”原则促进人工智能信任

**第一，可用（available）。**

**发展人工智能的首要目的，是促进人类发展，给人类和人类社会带来福祉，实现包容、普惠和可持续发展。为此，需要让尽可能多的人可以获取、使用人工智能，让人们都能共享技术红利，避免出现技术鸿沟。**

可用性，一方面意味着人工智能的发展应遵循以人为本的理念，尊重人的尊严、权利和自由以及文化多样性，让技术真正能够为人们所用，给人们带来价值。另一方面意味着人工智能与人类之间不是非此即彼的取代关系，而是可以成为人类的好帮手，增强人的智慧和创造力，实现和谐的人机关系。此外，可用性还意味着包容性，要求技术赋能于人，尤其是残障人士等弱势群体及少数族裔。

确保公平正义，防止偏见和歧视，是可用性的一个基本要求。这意味着人工智能应公平、平等地对待每一个人，避免造成不公平歧视，加剧或者固化社会不公平。一方面，应践行“经由设计的伦理”（ethics by design）理念，即将隐私、公平、安全、非歧视、福祉等伦理价值的保障融入到 AI 产品、服务的设计当中，确保算法的合理性和数据的准确、时新、完整、相关、无偏见和代表性，并采取技术方式识别、解决和消除偏见。另一方面，应制定解决歧视和偏见的指南和原则，并可通过伦理审查委员会等方式发现、解决潜在不公平歧视。

**第二，可靠（reliable）。**

**人工智能应当是安全可靠的，能够防范网络攻击等恶意干扰和其它意外后果，实现安全、稳定与可靠。**

在该原则之下，一方面人工智能系统应当经过严格的测试和验证，确保其性能达到合理预期；另一方面人工智能应确保数字网络安全、人身财产安全以及社会安全。

就数字网络安全而言，可靠性意味着，人工智能应遵守隐私法律要求，加强隐私保护和数据安全，保障个人对数的控制，防止数据滥用。在技术上，人工智能应遵循“经由设计的隐私”（privacy by design）理念，采取加密、匿名化等技术措施加强隐私保护，探索隐私友好型的机器学习方法如联邦学习（federated learning）。

**第三，可知（comprehensible）。**

**人工智能应当是透明的、可解释的，是人可以理解的，避免技术“黑盒”影响人们对人工智能的信任。**

人们对技术的信任来源于对技术的理解。然而，现代人工智能系统越来越成为一个“黑盒”，甚至有时连研发人员都无法理解他们的造物。《科学》杂志将破解这一“黑盒”视为人工智能领域的重大难题，为此，研发人员需要致力于解决人工智能“黑盒”问题，实现可理解、可解释的人工智能算法模型。

为了实现可知性，不同主体如消费者、政府等需要的透明度和信息是不一样的，还需要考虑知识产权、技术特征、人们的技术素养等事项。一般而言，对于由人工智能系统做出的决策和行为，在适当的时候应能提供说明或者解释，包括背后的逻辑和数据，这要求记录设计选择和相关数据，而不是一味追求技术透明。

换句话说，技术透明或者说算法透明不是对算法的每一个步骤、算法的技术原理和实现细节进行解释，简单公开算法系统的源代码也不能提供有效的透明度，反倒可能威胁数据隐私或影响技术安全应用。更进一步，考虑到 AI 的技术特征，理解 AI 系统整体是异常困难的，对理解 AI 作出的某个特定决策也收效甚微。所以，对于现代 AI 系统，通过解释某个结果如何得出而实现透明将面临巨大的技术挑战，也会极大限制 AI 的应用；相反，在 AI 系统的行为和决策上实现有效透明将更可取，也能提供显著的效益。

此外，在发展和应用人工智能的过程中，应为社会公众参与创造机会，并支持个人权利的行使。一方面，由于人工智能对个人和社会的潜在影响，确保社会公众能够参与到人工智能的研发和应用活动中，是至关重要的。为此，社会各界应为公众参与创造机会，诸如用户反馈、用户选择、用户控制等，也包括利用人工智能系统的能力促进公平赋能和公众参与。另一方面，公众参与也可以通过行使个人权利的方式来实现。为此，需要尊重数据保护和隐私权、表达和信息自由、非歧视等个人权利。同时，对于人工智能做出的决策，在适当的时候

提供异议和申诉机制以挑战这些决策；对于人工智能造成的损害，提供救济途径并能追究各参与方的法律责任。

最后，还需要保障个人的信息自决，比如，确保个人能够合理知晓其在与人工智能系统直接互动、其为人工智能系统提供了个人信息等，并就人工智能系统的目的、功能、限制、影响等向用户提供相关信息，以保障用户对人工智能的预期和一般性控制。

**第四，可控（controllable）。**

**人工智能的发展应置于人类的有效控制之下，避免危害人类个人或整体的利益。**

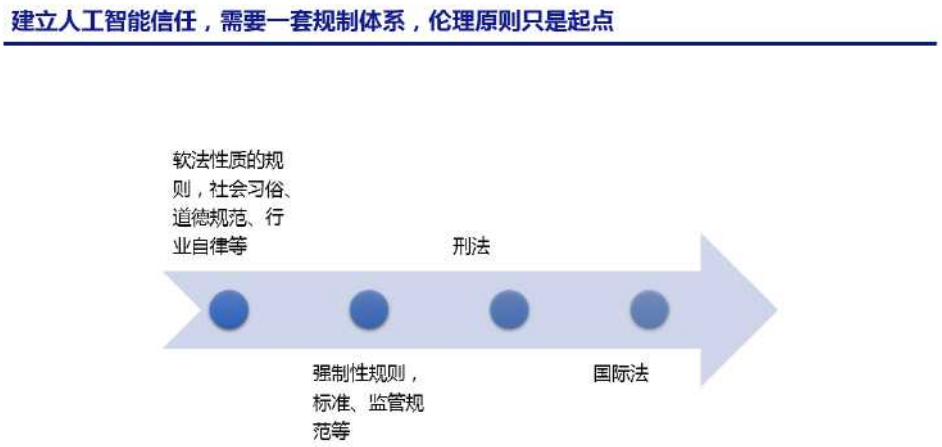
短期来看，发展和应用人工智能应确保其带来的社会福祉显著超过其可能给个人和社会带来的可预期的风险和负面影响，确保这些风险和负面影响是可控的，并在风险发生之后积极采取措施缓解、消除风险及其影响。只有当效益和正面价值显著超过可控的风险和消极影响时，人工智能的发展才符合可控性的要求。此外，“可控”还意味着，在人与机器关系中，对于机器所做的决定及其后果，需要考虑由人来承担最终责任，以确保人机共生的社会不失范。

长期来看，虽然人们现在还无法预料通用人工智能和超级人工智能能否实现以及如何实现，也无法完全预料其影响，但应遵循预警原则（precautionary principle），防范未来的风险，使未来可能出现的通用人工智能和超级人工智能能够服务于全人类的利益。

此外，可控性要求人们必须考虑人工智能应用的边界和底线问题，对人工智能技术的武器、监控、人类增强等特殊用途进行限制和规制。

## （二）构建塑造人工智能信任的规则体系

当然，信任的建立，需要一套规则体系。最左端，是软法性质的规则，包括社会习俗、道德规范等；往右，是强制性规则，比如规制组织内行为的法律（如公司法），适用于工业产品的标准，针对医疗、金融、贸易等细分领域的监管规范，等等；再往右，是刑法的威慑；最右端，是国际法，比如禁止生化武器的国际公约。




对于人工智能，人们需要一套类似的规则系统。在这些原则之下，人们可以探索制定标准、法律、国际公约等。在技术伦理的最新发展阶段，未来将可能出现一系列围绕数据和人工智能算法及其应用的法律规范。我们认为，需要遵循以下几个原则：

**第一，避免采取统一的专门监管。**企图为所有的人工智能应用制定统一的监管框架是不现实的，各个行业的专门监管机构最能评估人

工智能对特定领域的影响并采取适宜的监管措施。

**第二，采取包容审慎、敏捷灵活的规制方式。**一方面，由于技术以及商业模式快速发展和迭代，草率的立法不可能期待会产生正面的效果，而且成文或专门的立法恐难跟上技术步伐，故应避免严格、细致的法律要求，而是可以采取事后监管或者通过出台标准、行业公约、伦理框架、最佳实践、技术指南等调整人工智能等新技术的发展应用，支持行业自律。

另一方面，需要为自动驾驶汽车、人工智能医疗服务等人工智能新事物革除既有的监管法律政策障碍，避免过时的法律要求阻碍人工智能创新发展。因为随着人工智能技术持续渗透到各行各业，带来新的产品和服务，未来新事物与旧制度的冲突会越来越多，呼吁监管法律政策的革新和创新。

更进一步，作为对技术应用的“软法”规制，可以通过科技伦理来对科技行业的狭隘的技术向度和利益局限进行纠偏和矫正。而且考虑到当前技术发展的特征，科技伦理在预防和缓解新技术、新业务带来的风险挑战方面将发挥越来越重要的作用。因为法律规范本身具有滞后性，难以跟上技术发展的步伐，造成许多规范真空地带，而科技伦理可以更好地发挥事前引导作用，将技术发展引向向上向善的健康发展轨道。因此，人们需要超越狭隘的技术向度和利益局限，通过预警性思考、广泛的社会参与和多学科评估来充分讨论可能存在的风险

和危害，制定出切实可行的指导方针和伦理准则来引导、规范新技术研发应用，以更好地应对前沿领域技术应用可能引发的社会治理危机。

**第三，遵循分阶段的监管思路。**例如，在落实具体监管措施时，应首先考虑人工智能给公共安全和消费者带来的风险是否可以通过既有的监管框架来调整；其次应当考虑这些既有的监管框架是否可以充分且有效地规制这些风险；如果一项风险游离于既有监管框架之外，最后才应当考虑是否需要对监管框架作出修订或增加，以更好地应对人工智能带来的独特问题。

**第四，采取多利益相关方协同治理的模式。**考虑到人工智能等新技术的复杂性，相应的治理需要广泛听取行业、专家和公众意见，避免决策者与从业者脱节。更进一步，前沿技术领域的社会治理离不开广泛社会参与和跨学科研究。44 年前的阿西洛马重组 DNA 会议彰显了科学共同体在应对新技术发展的不确定性决策中的重要作用。现代科学技术与经济社会以异乎寻常的速度整合和相互建构，但其高度的专业化、知识化和技术化使圈外人很难对其中的风险和不确定性有准确的认知和判断，没有来自科学共同体内部的风险预警和自我反思，任何一种社会治理模式都很难奏效。

因此，一方面需要通过多利益相关方协同参与的方式，让监管机构、决策者、学术界、行业、社会公共机构、专家、从业者、公众等等都参与到新技术治理中来，各方协同治理好新技术。另一方面，通

过科技伦理教育宣传增进科研人员和社会公众在伦理上的自觉，使其不仅仅考虑狭隘的经济利益，而且对技术发展应用的潜在影响及其防范进行反思和预警性思考（precautionary thinking），才有可能通过广泛社会参与和跨学科研究的方式来实现对前沿技术的良好治理。

## 二、幸福（happiness）：智能社会人机共生

互联网、云计算、人工智能等前沿技术融合发展，加速了更加成熟的信息社会的到来，人类正在进入更加彻底的技术型社会。各种智能机器正在成为人类社会中不可或缺的一部分，和我们的生活和生产息息相关。这给人类与技术之间的关系提出了新的命题，需要人们深入思考智能社会如何实现人机共生（human-computer symbiosis）。因为无论技术怎么发展和应用，其都应当服务于人，服务于人的发展和幸福。正如古希腊哲学家亚里士多德所言，幸福是人生的终极目的。构建和谐的人机关系，实现智能社会人机共生，需要思考以下两个方面。

### （一）人人都有追求数字福祉的权利

在互联网发展二十多年后的今天，互联网创新需要从获取用户注意力，向促进用户数字福祉转变。数字福祉（digital wellbeing）有两大内涵，一方面是人人都可享受到互联网技术带来的便利和红利，但如今技术鸿沟和数字鸿沟依然存在，全球还有接近一半人口没有接

入互联网，老年人、残疾人等弱势群体未能充分享受到数字技术带来的便利。对此，国内外的科技公司一直在着力解决这些问题。例如，在国内，腾讯一贯重视用户的数字福祉，一直致力于让所有人都能平等、方便、无障碍地获取并利用信息，几乎全系列产品都有了“信息无障碍”版本。腾讯为此获得了联合国教科文组织颁发的“数字赋能残疾人奖”。

另一方面是减小、防止互联网技术对个人的负面影响。如今，在人手一部智能手机的年代，我们每天花在网络上的时间越来越多。于是人们开始担心，我们是不是在屏幕上花了太多时间，以致于可能会对我们的健康、生活、工作等产生负面影响？我们越发依赖于技术，将我们的自主性让渡给了技术？技术会对我们的思维产生什么样的影响？

在注意力经济模式下，互联网公司持续收集用户数据，并借助推荐算法实现个性化的内容推荐，从而更好地影响用户的日常交往，并持续吸引用户的注意力。但推荐算法也可能给用户带来信息茧房、算法偏见等负面影响，比如，限制用户对信息的自由选择，将用户置于算法建立起来的泡沫之中，只接触到自己喜欢或认同的内容，从而可能给用户造成信息茧房、自我封闭和偏见，进而影响用户的思维模式并可能扭曲用户的认知。此外，借由虚拟现实、增强现实等技术进行的交流和交往甚至会给我们的人际关系带来更深的影响，远程交流不断增强，而面对面交流则可能趋于式微。

可喜的是，国外发起的“Time Well Spent”运动逐渐获得主流科技公司的认可。馆 2018 年以来，科技公司已经开始采取措施应对互联网等数字技术应用对个人的影响，将保障、促进用户数字福祉作为其产品和服务的核心价值追求。比如，Android P 版推出的 Dashboard 功能，通过统计用户的屏幕时间，帮助用户控制手机和网络使用，防止过度沉迷，实现数字福祉（digital wellbeing）。iOS 12 推出了“屏幕使用时间”功能，同样意在帮助用户将手机和网络使用控制在合理的限度，防止过度使用。Facebook 则发布了“数字福祉”工具，允许用户监测自己花在社交网络上的时间并设定时长限制，同时更容易地关闭推送通知。在未成年人网络保护方面，腾讯从 2009 年起就持续推进相关措施，包括投入使用网游防沉迷系统，推出成长守护平台，推出未成年人消费提醒服务，不断升级健康游戏并启动最严格实名认证等等。此外，包括腾讯视频的护眼模式在内的多项创新模式旨在切实保障用户尤其是青少年用户的数字福祉。

总之，从关注互联网创新向关注、促进数字福祉的转变，要求科技公司依循“经由设计的数字福祉”(digital wellbeing by design)理念，将对用户数字福祉的促进融入到互联网产品、服务的设计中。

### （二）人人都有幸福工作的权利

人们对人工智能将如何影响就业和工作，存在显著的分歧。有人悲观，认为人工智能将导致大规模失业，无用阶层等概念甚嚣尘上，

呼吁对机器人和人工智能征税,并实行普遍基本收入制度(universal basic income)。有人乐观，认为人工智能将创造许多新工作和更多就业机会，但工作的内容、性质、方式和需求等可能发生很大变化，需要人们掌握新的知识和技能。纵观整个历史，技术创新从未带来大规模失业,反倒在经济活动中创造了新的、更多的就业机会,机器人、人工智能、3D 打印等新技术是否是个例外，还不得而知。

但就目前而言，人工智能的经济影响依然相对有限，不可能很快造成大规模失业，也不可能终结人类工作，因为技术采纳和渗透往往需要数年甚至数十年，需要对生产流程、组织设计、商业模式、供应链、法律制度、文化期待等各方面做出调整和改变。虽然短期内人工智能可能影响部分常规性的、重复性的工作。长远来看，以机器学习为代表的人工智能技术对人类社会、经济和工作的影响将是深刻的，但人类的角色和作用不会被削弱，相反会被加强和增强。比如，人工智能辅助诊疗工具可以帮助医生提高医疗诊断的效率和准确性并制定更有效、更个人化的治疗方案，医生则有更多时间来从事与其他医生交流、安慰病人、制定诊疗计划等需要判断力、创造力、同理心、交流和情商的非结构性工作任务。

未来，人类将以新的方式继续与机器协同工作，进入人机协同的新时代,人工智能将成为人类的强大帮手和助手,极大增强人类体力、智力等。因为机器有机器的用处，人有人的用处，两者配合才能够发挥最大化的价值。人们现在需要做的，就是为当下和未来的劳动者提

供适当的技能教育，为过渡期劳动者提供再培训、再教育的公平机会，支持早期教育和终身学习。《英国数字战略》显示，未来二十年内，90%以上的工作或多或少都需要数字技能。而且，人工智能、机器人等新技术正在从 ICT 领域向实体经济、服务业、农业等诸多经济部门扩散、渗透，未来中国将成为机器人大国，这些变化将在很大程度上影响当前的就业和经济结构。为了确保劳动者能够应对这场科技巨变，高效地为他们提供数字技能至关重要。

## 三、可持续（sustainability）：践行“科技向善”，塑造健康包容可持续的智慧社会

技术创新是推动人类和人类社会发展的最主要因素。而这一轮技术革命具有巨大的“向善”潜力，将对人类生活与社会进步带来突破性的提升。因此，在二十一世纪的今天，人类拥有的技术能力，以及这些技术所具有的“向善”潜力，是历史上任何时候都无法比拟的。换言之，这些技术本身是“向善”的工具，可以成为一股“向善”的力量，用于解决人类发展面临着的各种挑战，助力可持续发展目标。与此同时，人类所面临的挑战也是历史上任何时候都无法比拟的。联合国制定的《2030 可持续发展议程》确立了 17 项可持续发展目标，实现这些目标需要解决相应的问题和挑战，包括来自生态环境的，来自人类健康的，来自社会治理的，来自经济发展的，等等。将新技术应用于这些方面，是正确的、“向善”的方向。例如，人工智能与医

疗、教育、金融、政务民生、交通、城市治理、农业、能源、环保等领域的结合，可以更好地改善人类生活，塑造健康包容可持续的智慧社会。

因此，企业不能只顾财务表现，只追求经济利益，还必须肩负社会责任，追求社会效益，服务于好的社会目的和社会福祉，给社会带来积极贡献，实现利益与价值的统一。包括有意识有目的地设计、研发、应用技术来解决社会挑战。如今，“人工智能造福人类”（AI for Good）已经发展成为全球发展趋势，呼吁与行动并存。例如，2018 年 10 月，谷歌推出“人工智能向善”（AI for Social Good）项目，通过开放其人工智能能力，与外部机构合作利用人工智能解决全世界最重大的社会、人道和环境问题，并在过去几年针对野生动物保护、洪水预测、野火防范、婴儿健康等探索解决方案。

自 2018 年 1 月在国内首次提出“科技向善”以来，腾讯已将“科技向善”作为新的愿景与使命，并身体力行地践行“科技向善”理念。例如，将人工智能应用于医疗健康领域，致力于打造“救命的 AI”，用人工智能赋能医院和医务人员，改善我国医疗资源分布不均衡的状况。将计算机视觉技术应用于失踪人口寻找，协助警方打拐寻人，包括基于“跨年龄人脸识别”助力警方寻回被拐十年儿童，这在人工智能之前依靠人力几乎是不可能实现的。此外，发起 FEW 项目，将人工智能应用于食物、能源和水资源（FEW），致力于解决人类所面临的最大挑战。

# 结论：以智能时代的技术伦理重塑数字社会的信任

最后作为总结，互联网、人工智能等数字技术发展到今天，给个人和社会带来了诸多好处、便利和效率，未来还将持续推动经济发展和社会进步，但也引发了需要积极应对的法律、伦理和社会问题，加剧了人类与技术之间的矛盾和不信任关系。这是缺乏伦理关切的技术发展和应用的副产物。如今，步入智能时代，为了重塑数字社会的信任，我们需要呼吁以数据和算法为面向的新的技术伦理观，实现技术、人、社会之间的良性互动和发展。

因此，在“科技向善”理念之下，倡导并践行新的技术伦理观，包含三个层面：

一是人工智能等新技术的健康发展离不开人类正确价值观的引导，人工智能发展应用需要做到可用、可靠、可知、可控；

二是促进人类发展是任何技术发展应用的终极目的，需要确保人人都有追求数字福祉、幸福工作的权利，实现智能社会人机共生，让个体更自由、智慧、幸福地生活和发展；

三是这一轮新技术革命所具有的巨大“向善”潜力是历史上任何技术都无法比拟的，社会各界要善用人工智能等新技术解决人类发展所面临的社会、经济、环境、健康等方面的难题，助力可持续发展和美好社会。

最终，我们希望以新的技术伦理观增进人类对于技术发展应用的信任，让人工智能等技术进步持续造福人类和人类社会发展进步，塑造健康包容可持续的智慧社会。

**顾问团队：**

司　晓 腾讯研究院院长

张钦坤 腾讯研究院秘书长

程明霞 腾讯研究院助理院长

**研究团队：**

曹建峰 腾讯研究院高级研究员

腾讯 AI Lab 顾问组

王　融 腾讯研究院专家研究员

付　涛 腾讯研究院高级研究员

**学术支持：**

李　伦 大连理工大学大数据与人工智能伦理法律与社会研究中心主任

杨庆峰 上海大学哲学系主任、上海市自然辩证法研究会秘书长

李真真 中科院战略咨询研究院研究员、中科院人工智能产学研创新联盟伦理标准组组长

陈昌凤 清华大学新闻传播学院教授、常务副院长、博士生导师

刘永谋 中国人民大学哲学院教授、博士生导师

岳亚丁 腾讯公司专家研究员

李一凡 腾讯 QQ 音乐商业智能组高级工程师、博士

沈念祖 腾讯研究院高级研究员

刘金松 腾讯研究院高级研究员

腾讯研究院是腾讯公司设立的公共战略研究机构，旨在依托腾讯公司多元的产品、丰富的案例和海量的数据，围绕互联网发展的焦点问题，通过开放合作的研究平台，汇集各界智慧，共同推动数字经济与社会健康、有序的发展。为数字经济和社会健康、有序发展提供前沿性思考。坚守开放、包容、前瞻的研究视野，致力于成为现代科技与社会人文交叉汇聚的研究平台。

腾讯 AI Lab 是腾讯的企业级 AI 实验室，于 2016 年 4 月在深圳成立，目前在中国和美国有 70 位顶尖研究科学家及 300 位应用工程师。借助腾讯丰富应用场景、大数据、计算力及一流人才方面的长期积累，AI Lab 立足未来，开放合作，致力于不断提升 AI 的理解、决策与创造力，向“Make AI Everywhere”的愿景迈步。腾讯 AI Lab 强调研究与应用并重发展，基础研究关注机器学习、计算机视觉、语音识别及自然语言处理等四大方向，研究论文覆盖国际顶级学术会议；技术应用聚焦在社交、游戏、内容与平台型 AI 四大领域，在微信、QQ 等 100 多个产品中落地；行业应用不断取得突破，研发出屡获国际大奖的围棋 AI“绝艺”，并支持国家级 AI+医疗标杆产品“腾讯觅影”。