AI治理:寻求效率与安全的统一

人工智能:技术条件、风险分析和创新模式升级*

陈小平

(中国科学技术大学计算机学院)

摘要: 本文从两个方面探讨人工智能(AI)伦理和治理。一是现有AI成果的技术条件——封闭性。现有AI技术的应用要求场景是封闭的,所以AI的强大能力并非完全通过自主学习而获得。本文研究表明,受到广泛关注的用户隐私等问题,源于行业规则滞后、传统设计范式的伦理非封闭性和长期效应的预测能力不足。二是技术应用的外部机制。目前主要依靠熊彼特模式,其先天局限性是诸多伦理风险的深层原因,并限制了新技术在重大社会问题中的应用。本文基于对熊彼特创新向公义创新升级的时代趋势的分析,提出构建公义创新动力机制的若干路径,并以养老产业为例,阐述公义创新在创造性解决重大社会问题中所具有的巨大潜力和效能。

关键词: 人工智能伦理和治理,人工智能技术条件,创新模式升级,公义创新,创新动力机制

中图分类号: N01/TP18 **文献标识码:** A

DOI:10.19524/j.cnki.10-1009/g3.2021.02.001

作者简介: 陈小平,中国科学技术大学计算机学院教授。研究方向为人工智能基础理论及智能机器人关键技术。近年来对人工智能伦理与治理进行了系统性研究。

*本文是根据作者在2021年中国科学院伦理治理小型研讨会(第二期)上的报告整理而成。

一、前 言

从图灵于1950年提出“图灵测试”[1]以来，人工智能(artificial intelligence，以下简称AI)研究取得了丰硕成果，并应用于不同行业，对人类社会的发展具有巨大的直接作用和潜在的正面和负面的影响。当前人工智能伦理和治理的中心任务，是对所有正面、负面、短期和长期的作用、影响进行全面的分析和评估，并在可能的情况下制订合适的对策。[2][3]

“技术”(technology)是一个多义词，而且一种词义又有多种不同的理解。在一些讨论中，技术被理解为“知识的实际应用”(practical application of knowledge)，①这种理解实际上将技术研究与技术应用混为一谈，具有极强的误导性，比如引出了技术非中性的判断。[4]技术的另一种解释是“工艺和应用科学的科学研究和运用”(scientific study and use of mechanical arts and applied sciences)。②这个解释反映了技术一词的“技术研究”词义。本文认为，在AI伦理的讨论中，有必要明确区分技术研究与技术应用。

对AI及相关学科而言，技术研究与技术应用有明确界限。例如，作为训练法研究成果的深度学习算法通常不能被普通用户使用。为了应用这些算法，首先必须给定具体应用场景的设计规范，并收集该场景的一组数据，再给数据加上人工标注，然后用于训练某种选定的神经网络，训练的结果如果符合设计规范的要求，才成为产品/服务(或其部分)。同样，推理法的应用也必须首先给定具体应用场景的设计规范，并人工编写该场景的知识库，从而得到产品/服务。所以，以强力法和训练法为代表的AI技术研究及其直接成果，与这些技术成果的实际应用(产品/服务)，是有根本区别和分工的。整个AI的发展是建立在这种区别和分工的基础之上的。

因此，人工智能伦理和治理的探讨应该从两个方面展开。一是技术研究方面，从AI研究成果的技术条件出发，分析人工智能的伦理现状和治理对策；另一个是技术应用方面，从AI研究成果的产业落地机制出发，分析人工智能的伦理现状和治理对策。

①参见 https://www.merriam-webster.com/dictionary/technology: Definition of technology 1a: the practical application of knowledge especially in a particular area : ENGINEERING sense 2。

②参见《牛津高阶英汉双解词典》(Oxford Advanced Learner’s English-Chinese Dictionary)(第四版)，商务印书馆、牛津大学出版社，1997: 1569.

二、现有人工智能成果的技术条件

人工智能研究产生了大量不同的技术途径，其中得到最多研究、占据主导地位的两大类途径是强力法和训练法。[5]2-4 强力法利用显式表达的知识进行推理来解决问题，所以是可解释的。常用的知识表示和推理方法基于形式化逻辑、概率论、决策论规划、状态空间上的搜索等。训练法利用人工标注的数据训练人工神经网络（或其他类型的隐式知识表示模型），用训练好的人工神经网络来解决问题。[6][7] 深度学习是训练法的一个著名例子。以强力法、训练法为代表的现有人工智能技术能不能大规模实际应用？这种应用会引起哪些风险？这些问题引起了社会的广泛关注。

一个学科往往经历数千年的持续发展，即使部分理论或技术成果已经成熟，仍然只在一定范围内有效，即只在一定条件下可以应用，本文将这种条件称为科研成果的技术条件。对科技推动社会进步和经济发展而言，一个至关重要的课题是，及时识别现有科研成果的技术条件，从而及时推动部分成果的实际应用，以把握发展先机。认为一种新技术只处于两种极端状态——或者能够无条件应用，或者根本不能应用，这种观点不仅是完全错误的，也是极其有害的。一个社会如果不能率先识别某项新成果的技术条件，就不可能及时发现该成果的产业化时机和正确路径，于是只能等待其他社会实现成果转化之后，自己去“山寨”别人推向市场的产品或服务。这是“山寨”的深层原因之一。

目前，人工智能发展遇到的一个重大机遇和挑战，正是现有人工智能研究成果的技术条件的识别。事实上，人工智能强力法、训练法在一定条件下已经可以广泛应用。[6][7] 因此，只要正确把握强力法、训练法的技术条件，就能抓住人工智能产业落地的先机，避免重蹈“山寨”的覆辙。

新技术的应用总是落脚于一定的场景，人工智能产品/服务同样如此。在传统设计范式中，一个产品/服务的应用场景主要包括三方面的内容：（1）应用需求，即该应用所满足的用户需求，包括功能范围与性能指标；（2）应用场合，即该应用的使用场合；（3）应用条件，即保证该应用能够正常使用的条件。三方面内容的严格描述称为这个应用场景的设计规范。任何产品/服务的研发过程，必然包含设计规范的制定；而最终完成的产品/服务，必须完全遵守设计规范。以手机导航应用为例，其设计规

范的主要内容包括:(1)应用需求——为用户提供从起点到终点的路线及前进方向,但不包含用户移动的其他服务功能,比如如何避让周围行人和车辆等;(2)应用场合——用户在地面上步行或驾车,或者提前在手机、电脑屏幕上查看路线;(3)应用条件——使用导航服务时,用户的手机和环境网络处于正常工作状态,保持网络畅通和一定的带宽。手机导航功能的研发必须符合其设计规范,即研发出的产品/服务必须满足设计规范的全部要求。[5]8

如果一个应用场景符合以下三个条件,则称该场景相对于强力法是封闭的:(1)该场景的设计规范可以用有限多个确定的因素(变元)完全描述,而其他因素可以全部忽略;(2)这些因素共同遵守一组领域定律,而这组定律可以用一个人工智能模型充分表达;(3)相对于该场景的设计规范,上述人工智能模型的预测与实际情况足够接近。[5]8-9 这三个条件合称强力法封闭性条件。一个应用场景相对于训练法是封闭的,如果下列三个条件成立:(1)存在一套完整、确定的训练评价准则,这套准则充分反映了该应用场景的设计规范;(2)存在一个有限确定的代表性数据集,其中数据可以代表该场景的所有其他数据,即只用代表性数据进行训练就足够了;(3)存在一个人工神经网络 ANN 和一个监督学习算法,用该算法和代表性数据集训练 ANN 之后,ANN 将满足评价准则的全部要求。[5]10-11 这三个条件合称训练法封闭性条件。

现有人工智能成果的技术条件概括为两个封闭性准则。第一,强力法封闭性准则:如果一个应用场景符合强力法封闭性条件,那么人工智能强力法技术就可以应用于该场景。第二,训练法封闭性准则:如果一个应用场景符合训练法封闭性条件,那么人工智能训练法技术就可以应用于该场景。两个准则可以同时起作用,如果一个应用场景既符合强力法封闭性条件,又符合训练法封闭性条件,那么人工智能强力法和训练法技术就可以同时应用于该场景。短期内,其他人工智能技术途径的实用性程度远远低于强力法和训练法,所以用强力法和训练法的封闭性作为现有人工智能成果的技术条件具有普遍意义。

下面以围棋 AI 程序阿尔法狗为例,进一步阐释封闭性的意义和作用。众所周知,第二代阿尔法狗战胜了李世石;第三代阿尔法狗匿名为 Master,战胜了几乎全部现役的世界围棋高手;第四代是阿尔法狗零

（AlphaGo Zero），以 100∶0 完胜 Master。阿尔法狗零没有直接和人类棋手比赛，因为人类棋手已经无法与阿尔法狗零抗衡了。因此，AI 已经在特定领域超过了人类。尤其令人瞩目的是，虽然阿尔法狗早期使用了部分人类围棋知识，而阿尔法狗零则没有使用围棋规则以外的任何人类围棋知识。不过，这种简化的描述引起了诸多严重误判，比如误以为阿尔法狗零完全通过自学掌握了远超人类的围棋能力，并由此引申出更多没有科学依据的臆测和恐慌。

根据报告阿尔法狗零研究成果的原始论文，[8] 该程序基于四项核心技术：简化的决策论规划模型（整个程序的理论框架）、蒙特卡洛树搜索（自博中落子的自动决策机制）、强化学习（落子胜率估计的反推机制）和深层残差网络（反推结果的存储结构），其中前两项是典型的强力法技术，第三项同时属于强力法和训练法，第四项是典型的训练法（深度学习）技术。流行观点认为，阿尔法狗仅仅是深度学习的成功，与其他 AI 技术无关，这是又一个严重背离科学事实的误判。

分析表明，[5]10 阿尔法狗零的研发同时遵守了两个封闭性准则（尽管阿尔法狗零的原始文献并未提及封闭性）。具体地说，人类围棋是非封闭的，因为棋手决策时会考虑不同对手的特点，却无法提前掌握所有对手的信息。阿尔法狗零彻底放弃了这种做法，改为不考虑对手，只考虑棋盘上的 362 个落子（包括 pass 作为一个特殊落子），并根据自博产生的数据（包括依据围棋规则自动判断自博的胜负），用强化学习反推每一个落子[①]的胜率估计，从而实现了围棋问题的封闭化（将非封闭问题转化为封闭问题）。上述分析表明：作为人工智能发展最近一个里程碑的阿尔法狗零，也是遵守封闭性准则的，而且封闭化是其成功的关键。这对现阶段人工智能技术的实际应用具有普遍意义。

阿尔法狗零是完全依靠自主学习而获得其围棋博弈能力的吗？根据对原始论文[8] 的分析可知，阿尔法狗零的设计者做出的决策包括：决定只考虑落子，不考虑对手的不同特点；决定通过自博收集训练数据；决定用蒙特卡洛树搜索作为自博中的落子决策机制；决定用强化学习算法反推落子胜率估计；决定下棋时仅仅根据胜率估计进行落子决策；决定自博总局数为 2900 万局。阿尔法狗零程序自主完成的任务主要有：自动完成

①每一个落子是一个“变元”，在不同棋局下可以有不同的胜率估计。

2900 万局自博；自博中每一步棋的决策（从 1600 次试错中选出最好的一步棋）；自动记录每一局自博产生的数据（包括根据围棋规则判断一局棋的胜负）；根据自博数据用强化学习算法自动反推所有落子的胜率估计，并保存在深层残差网络中。由此可见，决定阿尔法狗零成功的所有主要决策，都是由其设计者做出的，而阿尔法狗零只是自主地完成了这些决策的具体执行。因此，阿尔法狗零的成功更多地归因于它的设计者，而不是它自身；也就是说，阿尔法狗零并不是完全通过自主学习而获得其高超围棋能力的。假如阿尔法狗零能够替代它的设计者，自主地做出相关的所有决策，那才可以说它自主地学会了下围棋。

上述分析表明，现有 AI 的“自学”只在其设计者规定的范围内起辅助作用，决定 AI 系统性能的主要是其设计者。因此，现有 AI 技术不会自主地形成“价值判断”，更不会基于这种价值判断做出决策。这个事实对于正确分析现有 AI 技术的风险状况及责任归属[①]具有决定性意义。

三、现有人工智能技术的风险状况及对策

人工智能成果的技术条件在人工智能伦理和治理中具有不可或缺的作用。本文依据 AI 技术条件分析其风险状况，并提出相应对策。

不同领域有不同的风险类型和特点，比如人工智能与生命科学的风险类型差异很大。总体上看，人工智能的可能风险主要有三种。第一种：技术失控，指的是人工智能技术超越人类的控制能力，使得人类失去对人工智能技术的控制。例如，假设将来出现了可以通过自主学习，同时在不同领域普遍超越人类能力的人工智能，则必然出现技术失控，人类将沦为 AI 技术的奴隶。[②]人们对人工智能发生技术失控的担忧远超历史上任何其他技术。根据上述分析，以强力法、训练法为代表的现有人工智能成果（包括深度学习）的技术条件是封闭性准则，即现有 AI 技术只在封闭性场景中可以应用，所以只要人类不把不成熟的 AI 技术投入使用，就不可能出现技术失控。预期这种情况在未来 15 年内都将持续存在。

---

①有关责任归属的初步讨论见文献 [9]。

②这种人工智能往往被称为“通用人工智能”“强人工智能”“超人工智能”。但这几个术语的严格定义几乎没有被讨论过，更谈不上达成共识，有些用法并不意味着“可以通过自主学习，同时在不同领域普遍超越人类能力的人工智能”。

迄今出现的断定人工智能必然发生技术失控的各种“预测”，都是脱离人工智能技术条件的主观推测；而脱离技术条件预测技术风险，是缺乏科学依据的，并且对于人工智能的健康发展是极其有害的。人工智能伦理和治理必须建立在人工智能发展规律的科学判断的基础上，而这种规律集中体现为技术条件。基于技术条件的判断方法，不仅适用于现有AI技术成果，也适用于未来成果。假如未来出现了超越封闭性准则的AI技术，则必须相应地识别出新的技术条件，并依据新的技术条件判断风险，在保证可控的前提下开展研究。

第二种风险是AI技术的非正当使用，包括技术误用和滥用，前者是无意的，后者是故意的。必须指出，人工智能及相关新技术的非正当使用，是当前存在的主要风险，尤其用户隐私、数据安全、算法公平性等问题，已引起较大反响。这些现象显然违背了社会公认的伦理原则，因此国内外提出了大量通用性伦理原则，并主张通过立法让这些原则得到遵守。然而，这一主张不符合新兴产业的客观实际和发展规律，不能真正解决问题。[3] 事实上，技术的非正当使用是一个“老问题”，新技术在最初应用时往往都出现过类似情况，主要治理手段不是通用的伦理原则，而是实际可操作的行业规则（如技术标准）的制定及监督执行，而且都取得了不错的效果。本文认为，在一定条件下，传统治理模式对于AI技术非正当使用仍然是有效的。例如，对于目前反映强烈的用户隐私问题，应由相关行业管理部门在充分调研、征求各方意见、完成专业论证的基础上，制定统一的行业规则，决定哪些用户隐私数据可以/不可以采用、在什么条件下可以/不可以采用、什么样的使用方式是允许/不允许的，等等，并相应地实施行业监管。这样才能整体上合理地兼顾用户、企业和社会的利益，并得到有效的落实。[9][3]

不过，仅仅依靠行业规则，并不能完全解决问题，还需对传统设计范式进行制度性升级。基于人工智能等新技术的产品/服务可能产生的社会效应，是以往的产品/服务不可比拟的，超出了传统设计范式的考虑范围。具体来说，传统设计范式依惯例不考虑伦理因素，只考虑功能性因素，即与产品/服务的使用功能直接相关的因素；也就是说，传统设计范式在伦理方面是制度性非封闭的。例如在网络信息服务领域，为了有效应对信息过载问题而出现的个性化推荐技术，让网络媒体平台主动呈现用

户感兴趣的信息，从而显著改善了用户的信息获取体验。可是，这种技术同时也带来了“信息茧房”效应，这是个性化推荐技术的研发者始料未及的，而且不应简单地责怪研发者考虑不周，因为只考虑功能性因素是传统设计范式的惯例。因此，在人工智能产品/服务的未来研发中，应将伦理因素制度性地纳入设计规范之内，而哪些伦理因素必须纳入，则由行业规则决定。当然，如果研发企业主动考虑行业规则之外的更多伦理因素，也是值得鼓励和尝试的。

由此进一步引出一个新问题，将伦理因素纳入设计规范之后，现有的设计、研发能力是否足以有效应对伦理因素？与隐私数据使用、算法公平性直接相关的伦理因素，往往是现有能力可以有效处理的。但在其他很多情况下，伦理因素涉及产品/服务的长期效应和“跨界效应”，比如“信息茧房”效应就横跨了信息服务和社会文化两大领域，于是产品/服务的技术成分与这些伦理因素之间的关联变得十分曲折、复杂。为了把握这种关联，强力法所需的相关知识和训练法所需的相关数据，通常都难以获得，所以现有设计、研发能力不足以有效处理这些伦理因素。这对现有人工智能技术构成一个全新挑战。由此可见，为了实现“符合伦理的设计”，必须升级现有人工智能技术。也就是说，AI 治理是以 AI 产品/服务的设计研发能力的升级为前提条件的。

面对技术非正当使用，存在三大挑战和机遇：可执行的行业规则的制定、研发体系制度建设（消除传统设计范式的伦理非封闭性）和能力建设（发展能够处理长期效应、跨界效应的设计研发能力）。率先把握这些机遇的国家、企业和个人将获得新型竞争优势。

第三种风险是社会效应风险，即在不发生第一种和第二种风险的情况下，新技术应用产生了严重的负面社会效应。例如，假设新技术的广泛应用引起工作岗位总体上的大量减少，即使没有发生技术失控和技术非正当使用，仍然是一种严重的负面效应，将对社会产生巨大、深远的影响。对于这种潜在风险，人类现有预测能力不仅严重不足，而且预测方法也不适应新形势的需要，从而给判断和决策带来困难。本文认为，有必要建立一种风险与机遇一体化的预测能力，并满足下列要求：把握 AI 等新技术的科学原理、技术条件和适用范围；把握相关学科领域之间的普遍性关联；把握科技成果与产业需求之间普遍性关联和匹配关系；对直接效应

和跨界效应、正面效应和负面效应进行一体化预测；信息来源和成果发布的渠道畅通、覆盖充分。同时应认识到，准确地预测一切是不可能的，社会发展也不可能完全建立在预测的基础上。

四、人工智能时代的创新模式升级

一般而言，技术研究的成果不具有“自我实现”的能力，即任何技术成果都不能自行实现应用，无论这些成果多么成熟、社会需求多么强烈。这就是技术的“被动性”。[2] 事实上，半个多世纪以来，技术应用主要是借助于熊彼特创新模式才得以落地，该模式也成为技术应用中诸多问题（尤其是技术伦理问题）的深层原因。[9]

熊彼特创新（innovation）的核心内涵是市场要素的商业化组合，即人力资源、技术成果、生产资料、金融投资、商业模式等要素的市场化配置。四个主要特性是：有足够的商业利益、满足用户需求、有税收、高效率；终极标准是市场接受，即只有被市场接受、获得足够商业利益的产品/服务才可以成功。[9] 熊彼特模式最强大之处在于拥有内生动力——在适宜的社会经济体制下，熊彼特模式能够依靠自身的力量持续地生存和发展。作为对比，科学技术研究及传统公益事业则不具有这种独立生存能力，必须依靠外界源源不断的投入。

熊彼特模式的根本局限性在于，没有足够商业利益的社会需求无法通过熊彼特创新得到满足。如科学研究、救灾、环保、卫生、扶助失能人群等事业，难以应用熊彼特模式。还有一些公益事业，如技术研究、教育、医疗、文化艺术等，应用熊彼特模式往往带来严重弊病。值得注意的是，熊彼特模式要求产品满足用户需求，而传统设计范式所考虑的用户需求限于功能性因素，不包括伦理要求，所以熊彼特模式实际上不要求产品满足功能性因素之外的伦理要求。

因此，在熊彼特模式下出现伦理风险是不可避免的。例如，由于不同群体消费能力的差异，在熊彼特模式下，智能产品的开发“自然地”定位于消费力强的群体，从而导致老龄群体难以享受科技进步的成果，甚至形成一种数字鸿沟。又如，“机器换人”最初是由制造业部分行业严重缺工[10] 引起的，由此推动了我国工业机器人技术的研究和应用。可是随着机器人和人工智能技术的不断进步，机器的生产效率必然越来越普遍地

高于人工，于是在熊彼特模式下，必然越来越普遍地用机器替代人工，从而出现众多行业的少人化趋势。目前仍不确定，这种趋势会不会导致大规模就业难题，可以确定的是，如果出现这种情况，在熊彼特模式中则是无解的。[9] 因此，熊彼特模式在客观上构成了第二、第三种风险的单向催化剂和加速器。

中国的快速发展是举世公认的，可这种发展正在导致熊彼特模式与时代需要的符合度快速下降，这一深刻变化却没有引起关注。具体表现为下列三种趋势：第一种趋势，社会需求重心变化。例如，老龄化、少子化、产业少人化、阶层分化等议题正在引起越来越普遍和强烈的关注，表明这些议题相对于其他议题的重要程度正在快速提高，也反映了社会需求重心的快速调整。显然，这些议题都无法通过熊彼特模式加以解决，AI 等新技术也难以借助于熊彼特模式为解决这些重大社会问题发挥作用。第二种趋势，新技术的商业瓶颈变化。新技术往往短期内难以获得足够的商业利益，从而在熊彼特模式下出现商业瓶颈，这是一种长期存在的现象。然而，人工智能、机器人等一批新技术的快速发展不仅让商业瓶颈的出现范围扩大了，也让这种瓶颈的消极后果的严重程度提高了。例如，养老已成为当前中国社会的一个紧迫问题，并且越来越严重，而机器人技术已经可以满足养老的部分需求，却由于商业瓶颈而无法应用落地。这种情况下，熊彼特模式不仅未能推动技术应用，反而制约了重大社会需求中的技术应用。第三种趋势，二元融合趋势。主要表现为物质追求与精神追求趋向融合（典型案例如软件开源运动）、工作与娱乐趋向融合（典型案例如社交媒体内容创业[11]）、生产与消费趋向融合（如粉丝经济、依单定制、用户设计）。这些动向都隐含着二元融合的经济-社会动力机制，而熊彼特模式是典型的一元化机制——市场要素的商业化组合。

以上分析表明：人工智能时代需要新的创新模式，这就是公义创新模式。[9] 公义创新的核心内涵是市场要素和非市场要素的公义性组合，即基于公义原则的组合。公义原则是商业原则和人性原则中有效成分的提炼、整合和升级。人性原则是人的生存、生活和发展多重需要的统一性原则。

公义创新包含着熊彼特创新和传统公益，却不是它们的简单合并。例如，公民基本收入制虽然得到了一些支持，也符合传统公益的思路，却不符合公义创新的原理，因为违反了人性原则。一般地，一种社会举措受

到目标群体的普遍欢迎，不必然意味着该举措符合目标群体的根本利益，并将得到目标群体的长期支持。事实上，公民基本收入制的实施将极大地助推和加速产业少人化进程，以至引起“无用阶层”的大量出现，并严重侵害人的发展权，干扰健全社会的有序发展。

公义创新是熊彼特创新的升级，包含六个方面的内容。第一，评价体系升级。从经济效益与社会效益的分离评价，升级为统一评价；第二，创新要素组合机制升级。从市场、生产要素的组合，升级为市场和非市场要素的组合，从而开辟更广阔的创新空间；第三，社会参与模式升级。从工作、公益和娱乐休闲的分离式参与，升级为融合式参与；第四，风险预研与应对体系升级。在界内效应和短期效应预测的基础上，发展界外效应和长期效应预测机制、应对体系；第五，职业道德升级。从职业道德与社会公德的相互分离，升级到相互融合；第六，管理体系升级：从现行管理体系，升级为适应公义创新的管理体系。[9]

动力机制的构建是公义创新面临的一个核心挑战，可从下列路径入手。第一条路径，软件开源机制的升级。软件开源运动[12]是第一个突破熊彼特模式而取得巨大成功的技术创新案例。开源彻底颠覆了熊彼特模式的内核机制——商业利益驱动，实现了生产方式、管理方式直到知识产权保护方式的一系列根本性变革，并达到了与熊彼特模式下的对标竞品（主流操作系统）势均力敌、平分天下的应用效果，充分展示了物质-精神二元融合机制的强大生命力和市场存活力。从中可以提炼、发展出适用于更多行业的公义创新动力机制。第二条路经，工时机制的升级。这里的工时制指的不是按时计酬的传统工资制，而是正在部分发达国家/地区兴起的一种介于计时工资和公益劳动之间的新型工作模式。例如，劳动者通过在养老院的服务获得“工时”，却不转化为工资或其他任何资金形式的收入，而是获得未来享受同等工时养老服务的回报。显然，工时制普遍适用于众多行业，不限于养老。在短期内不产生足够商业利益的应用场合，对包括劳务、知识产权在内的多种社会贡献实行工时制，有望为化解商业瓶颈发挥关键性作用。第三条路径，融合路径的升级，主要包括工作与娱乐融合、生产与消费融合的新型社会参与机制，如社交媒体内容创业、软件开源运动等。在这种机制中，商业利益不是参与者的唯一目的，个人精神追求同样发挥着关键作用。于是从社会层面看，这种机制是

落实经济效益与社会效益协调统一的一条可行路径。

在实际落地中，三种路径可以综合运用，并与熊彼特模式相结合。例如，养老服务业目前面临的第一难题是人力不足，尤其是护工，因为多数情况下养老院不可能产生足够的商业利益，所以熊彼特模式无效，传统公益也是杯水车薪。在公义创新模式下，义工的服务将不再是传统的无偿付出，也不转化为传统的付费劳务，而是让义工享有未来获得等量服务回报的机会，从而形成一种不同于传统劳务和公益服务的社会贡献形态——“公义服务”。由于养老是每一个人的刚需，而且多数人无法依靠商业化养老，在相关制度保障下，养老公义服务将吸引大量人力资源投入养老业，从而化解人力短缺问题。随后，商业资本将以新的商业模式进军养老产业，推进机器人等新技术在养老业的普及应用，并全面实现养老产业的现代化，让全社会的养老机能变得足够强大，最终解决所有人的养老问题。所以，公义创新将为养老问题创造一个熊彼特模式下根本不存在的解决方案——“公义养老”，从而达到“义者养其年”的功效，其意义堪比土改实现的“耕者有其田”。

总之，通过伦理规范和传统手段对熊彼特模式进行监管，只能让熊彼特模式不做某些不好的事，却不能让它做它原本不能做的好事(比如公义养老)，而公义创新却可以二者兼得，因为公义创新不是对熊彼特创新和传统公益做“加法运算”，而是在社会效益和经济效益两个维度上做“升维操作”——从两个相互独立且冲突的一维空间整合升级为一个自洽的二维空间，该空间不仅保留熊彼特模式和传统公益，而且创造出大量前所未有的创新可能性。这也说明，来自 AI 等新技术应用的科技红利需要改革红利的必要支撑，起这种支撑作用的是创新模式的升级，即从熊彼特创新升级为公义创新。

志谢：本文撰稿过程中，作者与梁正教授、顾淑林教授、眭纪刚研究员、赵延东教授就相关主题进行了交流，在此一并感谢。

参考文献

[1] A. M. Turing, Computing Machinery and Intelligence. Mind 49: 433–460, 1950.

[2] 陈小平. 人工智能伦理建设的目标、任务与路径: 六个议题及其依据. 哲学研究,

2020, (9): 79–87/107.

[3] 陈小平. 人工智能伦理体系: 基础架构与关键问题. 智能系统学报, 2019, (4): 605–610.

[4] Noe¨mi Manders-Huits. What Values in Design? The Challenge of Incorporating Moral Values into Design. Sci Eng Ethics, 2011, (17): 271–287.

[5] 陈小平 (主编). 人工智能伦理导引. 合肥: 中国科学技术大学出版社, 2021: 1−60

[6] 陈小平. 人工智能中的封闭性和强封闭性——现有成果的能力边界、应用条件和伦理风险. 智能系统学报, 2020, (1): 114–120.

[7] 陈小平. 封闭性场景: 人工智能的产业化路径. 文化纵横, 2020, (1): 34–42.

[8] David Silver, Julian Schrittwieser, et. al. Mastering the game of Go without human knowledge. Nature, 2017-10-18.

[9] 常宝丽. 公义创新: 人工智能时代的创新模式——专访中国科学技术大学陈小平教授. 信睿周报, 46. (1).

[10] 赵灵敏. 用工荒, 一个时代的终结. 今日南国, 2010, (7): 26–28.

[11] 亚文辉. 自媒体内容创业的时代来了?. 中国文化报, 2015-11-27. (6).

[12] Siobhan Clare O ’Mahony. The emergence of a new commercial actor: Community managed software projects. Stanford, CA: Stanford University, 2002: 34–42.

Artificial Intelligence: Applicability, risk analysis, and innovation mode upgrading

CHEN Xiao-Ping
(Computer School, University of Science and Technology of China)

Abstract: Ethics and governance of Artificial Intelligence (AI) are explored from two aspects in this paper. The first aspect is from the applicability of the existing AI achievements——the closedness, which requires that the application scenarios should be closed and implies that the remarkable capabilities of the existing AI technology is not completely obtained through autonomous learning. This study also shows that the ethical problems which are widely concerned such as that of user privacy are due to the lag of code of practice, the ethical non-closedness in the traditional design paradigm and the insufficient ability to predict long-term effects. The second aspect is from the external mechanism of technology application, Schumpeter's innovation, whose inherent limitations are the deep cases of many ethical risks and restrict the application of AI technology in major social problems, as revealed in this paper for the first time. Based on the analysis of the trend of upgrading from Schumpeter's innovation to Gong-Yi innovation, some ways to construct the dynamic mechanism of Gong-Yi innovation is put forth. In addition, a solution to the plight of aged care is proposed to illustrate the tremendous potential and effectiveness of Gong-Yi innovation in creatively solving major social problems.

Keywords: AI ethics and governance, applicability of AI technology, innovation mode upgrading, Gong-Yi innovation, dynamic mechanism of innovation

(责任编辑　肖 利)