DOI: 10.13376/j.cbls/2022109
文章编号：1004-0374(2022)08-0983-07

陈小平，中国科学技术大学计算机科学与技术学院教授，中国人工智能学会人工智能伦理治理工委会主委，教育部重点领域教学资源及新型教材建设项目专家工作组专家，全球人工智能理事会执行委员。主要研究方向为人工智能基础理论和智能机器人关键技术，近年来致力于融差性原理和开放知识的理论研究和系统性工程验证，以及人工智能伦理治理的研究与教学。曾获中国科学技术大学“杰出研究”校长奖，IEEE ROBIO 最佳论文奖，世界人工智能联合大会最佳自主机器人奖、通用机器人技能奖和助老机器人比赛第一名，以及机器人世界杯 12 项世界冠军等荣誉。

人工智能伦理治理：一种新型问题的底层逻辑和创新探索

陈小平
(中国科学技术大学计算机科学与技术学院，合肥 230026)

摘　要：本文首先梳理人工智能的基本观点、主要成果和历史性突破，以获得对人工智能学科特性的理解。依据这种理解，区分了人工智能伦理治理的三类挑战性问题 —— 可控性问题、合理性问题和重大相关问题。通过问题分析发现，人工智能伦理治理作为一个整体，不是传统的风险治理问题，而是 AI 驱动的科技、经济、社会发展向何处去的问题。针对这一新型问题，本文提出“内在追求观”，主张明确增进人类福祉为人工智能内在伦理追求。本文表明，为了有效应对三类挑战性问题，需要以人工智能的学科特性和内在伦理追求为人工智能伦理治理的底层逻辑，尝试传统治理模式的升级 (以隐私保护为例)，并摸索、建立新型研究主题 (如“可控性研究”) 和新型治理模式 (如“公义创新”)。

关键词：人工智能；人工智能伦理；风险治理；创新模式；公义创新
中图分类号：B82-057；N01；TP18　　**文献标志码**：A

Ethics and governance of AI: a new-type problem with its underlying logic and practical challenges

CHEN Xiao-Ping
(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: This paper begins with summarizing and clarifying the fundamental viewpoints, major achievements and breakthroughs of AI research, to obtain an understanding of the characteristics of AI. According to this understanding, three sorts of challenging issues in the ethics and governance of AI are identified, i.e., that of controllability of AI, soundness of AI applications and AI-related social changes. Through analyzing these issues, it is found that the ethics and governance of AI is not a traditional risk governance problem, but a new-type problem of where the AI driven economic and social development is going. Aiming at this new-type problem, the paper puts

收稿日期：2022-05-12
基金项目：国家自然科学基金项目(92048301，U1613216)
通信作者：E-mail: xpchen@ustc.edu.cn

forward the view of intrinsic ethical pursuit, which advocates taking promoting human well-being explicitly as the intrinsic ethical pursuit of AI. Furthermore, in order to solve three challenging issues, we should base the ethics and governance of AI on the underlying logic composed of the characteristics and intrinsic ethical pursuit of AI, and make the effort to upgrade the traditional mode of risk governance (exemplified by data privacy protection) and explore novel research topics (such as the controllability study) and new-type governance modes (such as Gong-Yi innovation).

Key words: AI; AI ethics; risk management; innovation mode; Gong-Yi innovation

社会各界对人工智能伦理治理的高度关注，主要源于近年来人工智能应用实践中出现的诸多挑战 [1-5]。为了有效地应对这些挑战，本文首先梳理人工智能的基本观点、主要成果和历史性突破，从而概括其现阶段的学科规律。在此认识之上，将人工智能伦理治理的挑战性问题区分为三类：可控性问题、合理性问题和重大相关问题。对三类问题的分析表明，基于风险预测和事后治理的传统风险治理模式，只能应对其中一部分相对容易的问题，其他问题不是单纯的风险问题，而是 AI 驱动的科技、经济、社会发展向何处去的问题。所以，人工智能伦理治理本质上是一种新型问题。

针对这一新型问题，本文提出人工智能伦理治理的“内在追求观”，主张人工智能带有内在伦理追求 —— 增进人类福祉，而流行看法认为，人工智能本身不带有伦理追求。本文将阐述这种新观点在人工智能发展及伦理治理中的必要性和紧迫性。

人工智能的内在伦理追求和学科规律共同构成新型治理的底层逻辑。针对三类挑战性问题，需要一方面探索传统治理模式的升级 (如面对隐私保护问题)，另一方面摸索、建立新型研究主题 (如“可控性研究”) 和新型治理模式 (如“公义创新”)。

1 人工智能伦理治理的科学基础

1.1 人工智能的基本观点

已提出的各种人工智能定义均未达成共识。根据对 70 年来人工智能发展状况的分析，人工智能研究主要立足于两种基本观点 —— 原理模拟观和功能模仿观 [6]。

原理模拟观认为，人工智能是人类思维和人类智能的原理模拟，这里的“原理”指的是人脑或人体神经系统的工作过程、机理、机制、结构等。原理模拟观要求，人工智能的工作原理必须与人脑或神经系统的工作原理相同，并通过对这些原理的模拟从而产生对应的功能。

功能模仿观认为，人工智能是人类思维和人类智能的功能模仿，这些功能包括推理、学习、理解、决策、创造、感知、行动等。功能模仿观要求，人工智能具有与人类思维和人类智能相似的功能，而人工智能的工作原理既可以与人相同，也可以与人不同。所以，功能模仿实际上包含着原理模拟。功能模仿观最初是由 AI 奠基人、创始人艾伦 · 图灵于 1950 年正式提出的。

1.2 人工智能的研究成果

人工智能研究已取得大量成果，主要包括以下技术途径。第一条途径是基于模型的强力法，包括推理法、概率法、规划法、因果法和搜索法等。这些方法主要来源于人类在长期科学实践中形成的抽象理论，如数理逻辑、概率论、马氏决策过程理论、因果理论和搜索空间表示，运用这些理论构建问题的形式化模型，并在模型的基础上进行问题求解。因此，强力法方法通常具有可解释性。强力法方法不模拟大脑的神经过程、机理或结构，所以属于功能模仿观。

第二条途径是数据驱动的训练法。运用训练法首先要建立元模型，以规定学习对象、训练目标、数据标注、训练方法和网络表示等，这些规定一般不是形式化表达的 (所以称为元模型)。然后，根据元模型进行训练，得到参数被调节好的人工神经元网络 (或其他种类的参数化网络表示)。训练法技术往往同时包含功能模仿和原理模拟，所以仍然属于功能模仿观。深度学习是训练法的典型代表，AI 早期研究者通过模拟大脑神经网络的结构及电学特性，提出了人工神经元网络 (ANN)，后来的研究者发明 (而非模拟) 了对 ANN 进行训练的有效策略，二者结合形成了深度学习技术。现有训练法技术一般不具有可解释性。

第三条途径是集成智能，运用强力法、训练法和其他方法的集成来构建 AI 系统，以解决复杂问题。迄今为止，通过了大规模工程化验证的 AI 系统都是集成智能，AlphaGo Zero 就是一个成功范例，它包含四项核心技术，其中两项是强力法技术 ——

新型决策论规划模型和蒙特卡洛树搜索，两项是训练法技术 —— 强化学习和深层残差网络。成功实现了蛋白质结构预测功能的 AlphaFold 和 RoseTTAfold 也属于集成智能。由此可见，认为 AlphaGo 仅仅是训练法的成果是严重背离科学事实的。

人工智能还发展出了很多其他方法，如演化计算、包容结构等等，它们都可以融于集成智能之中。

1.3　人工智能的历史性突破

AlphaGo Zero 标志着人工智能的历史性突破。它是第一个超过人类的围棋 AI 系统，而且不依赖于人类的围棋知识 (围棋规则除外)，仅仅通过 40 天“自博”(自己和自己下棋)，就以 100 ： 0 战胜了 AlphaGo 三代，后者之前已战胜了人类的所有现役围棋顶级棋手，从而证明人类下围棋已经远远不是人工智能的对手。之后短短几年内，这一突破已在蛋白质结构预测领域实现了实用化，从而开启了 AI 制药 ——AI 驱动的制药行业颠覆性创新，并开创了 AI for Science 的新时代。

AlphaGo Zero 取得成功的原因，不仅在于强力法和训练法的集成，更重要的是对围棋问题进行了“封闭化”[7]。在人类围棋和 AlphaGo 之前的 AI 围棋中，人类棋手和围棋 AI 的决策都是与对手的博弈策略有关的 —— 如果对手的策略不同，自己的决策也会不同。但是，对手的策略是一个无法完全预知的“变元”，这使得围棋决策是一个非封闭的问题。AlphaGo Zero 的设计者采用了一种全新思路，让围棋 AI 的决策只依据每一个落子的胜率估计，不考虑对手的策略 [1]。这就使得决策相关的所有变元都是可预知的，从而将围棋决策转化为一个封闭的问题。实际应用中，大部分问题都是非封闭的，必须首先进行封闭化，才可以成功地应用现有人工智能技术 [8]，而封闭化只能依靠人的创造，至少目前 AI 不具备封闭化的能力。

事实上，AlphaGo Zero 的成功是通过如下方式取得的：首先，人想出了一种新的围棋决策策略 (依落子胜率估计进行决策)，以及让 AI 自动掌握这种策略的训练方法 (利用自博、强化学习和深层残差网络)；然后，让 AI 通过这种训练获得这种策略，之后用获得的策略 (存储在深层残差网络中) 和人下棋。由此可见，对于以 AlphaGo Zero 为代表的人工智能而言，无论在基础研究还是在实际应用中，人都具有决定性作用。

2　人工智能伦理治理的挑战性问题

人工智能伦理治理面临的挑战性问题可以区分为三类。通过分析它们的现状和性质发现，人工智能伦理治理作为一个整体，不是传统的风险治理问题，而是一种新型问题。

2.1　可控性问题

如果一种人工智能的持续存在和未来发展方向完全不在人类的控制之下，那么这种人工智能就是不可控的。这里的“不可控”不是指人工智能在人类不干预的情况下，独立地执行某些任务并产生不符合伦理的后果；因为这种情况可以是人工智能的设计者 (人) 事先安排好的，所以事实上仍然在人类的控制之下。

还需要强调，传统意义上一种技术是不可控的，指的是该技术被人类应用之后将出现不可控情况，但人类可以选择不应用，从而避免出现不可控。然而，不可控的人工智能可以独立于人类的选择而自行应用，或以其他方式对人类和社会产生作用，并带来严重后果。这种可能性在科技史上是第一次出现，反映了人工智能及其伦理治理的独特性。

现有人工智能都是可控的。未来出现不可控人工智能的可能性理论上是存在的，但实际上会不会出现以及出现的具体形式，在现有条件下无法预测，所以也无从考虑具体的防范措施，而且事后治理显然也是无效的。以上分析表明，人工智能的可控性问题是一种新型问题，无法通过传统的风险治理加以应对。

2.2　合理性问题

人工智能技术的一种实际应用往往同时产生伦理上的正效应和负效应，很少出现只有正效应或只有负效应的情况。如果一项应用的负效应超过了容忍范围，则该应用存在合理性问题。如何确定正负效应及负效应的容忍范围，属于合理性问题的研究内容，也是制定相应的伦理规范和治理方案的必要基础。目前大众重点关注的合理性问题有：用户隐私问题、数据安全问题、算法公平性和透明性问题、数字鸿沟问题等。解决这些问题具有实践中的紧迫性。

各种现有人工智能产品 (包括服务、平台等) 都具有可控性，因为它们的研发、部署、维护和撤销都是由人类决定的。所以，合理性问题与可控性问题是性质不同的两类问题。各种合理性问题的具体表现在一定程度上是可预测的，而且也可以通过

事后治理加以消除或化解。不仅如此，在一定范围内和一定程度上，还可以通过技术手段加以解决或帮助解决。

合理性问题是人工智能伦理治理中最容易解决的类型，传统的风险治理模式仍然有效，但涉及大量复杂的新情况和新课题。

2.3 重大相关问题

现有人工智能技术的应用正带来一系列颠覆性产业变革，如 AI 制药、智能制造、智慧农业、驾驶自动化等 [6]。这些变革与同时发生的其他一些进程交互作用深刻地改变着世界。所有与人工智能相关的重大社会问题，都是人工智能伦理治理的重大相关问题，如“无用阶层”问题 (人工智能的普及应用导致大部分人失业)、老龄化 / 少子化问题 (人工智能能否帮助应对人口老龄化、少子化和普惠养老)、人的发展问题 (人在人工智能时代如何生存和发展)、气候与环境问题 (人工智能能否帮助人类应对气候变化和环境污染) 等等。

重大相关问题反映了 AI 应用的长期效应和未来社会的不确定性。长期效应由大量复杂因素的交互作用而产生，带来的一些变化可能是不可逆的，也无法及时进行预测。最具挑战性的情况是出现某些不可预测、不可接受且不可逆的变化，使得基于风险预测和事后治理的传统治理模式彻底失效。这再次表明，人工智能伦理治理是一类新型问题，需要探索新型治理模式。

与可控性问题不同，重大相关问题可由现有人工智能技术的应用而导致，并且不可预测、不可接受且不可逆的变化可能正在实际地发生着；换言之，重大相关问题中潜在的风险不是需要通过新的重大努力 (如发明、研制出不可控人工智能) 才会发生，而是不进行新的重大努力就可能无法避免。

3 人工智能的内在伦理追求——增进人类福祉

在科技与伦理关系的研究中，提出了很多观点和学说 [1]。然而在人工智能相关范围内，关于科技 - 伦理关系主要有三种观点。

第一种观点可称为“天然符合观”，即认为科技总是有利于经济和社会发展的，所以科技活动“天然地”符合伦理，创造发明不应受到任何限制。这是以往科技实践中存在的一种朴素观点。人工智能面临的伦理治理挑战对这种观点提出了强烈质疑。

第二种观点可称为“分离约束观”，即认为科技与伦理是相互分离的，科技本身不包含伦理追求。由于人工智能应用中已经出现了一些伦理问题，所以为了避免或化解严重的负效应，需要对人工智能进行伦理约束。这是目前较为流行的一种观点。

第三种观点称为“内在追求观”，即认为人工智能带有内在伦理追求 —— 增进人类福祉，但这种追求在实践中未必“天然地”得到满足，所以需要进行伦理治理，以保证人工智能的内在伦理追求得以实现。这是本文提出的观点。

本文认为，出于以下原因，应该明确地将增进人类福祉作为人工智能事业的内在伦理追求。首先，人工智能是一项有目的的人类活动，人在其中具有决定性作用。大部分从业者、相关者相信，人工智能有利于增进人类福祉；这意味着，他们有意无意地追求通过人工智能增进人类福祉。这就表明，增进人类福祉一直是人工智能的内在伦理追求，只是过去没有明确。

其次，人工智能不仅需要长期的巨大投入，而且面临着伦理风险和挑战。在此情况下，沿袭传统观点，将发展人工智能的目的局限于单纯的“科学探索”和“发展经济”，是不符合人工智能的实际情况和治理需要的。所以，对于认为人工智能不以增进人类福祉为伦理追求的少数从业者和相关者而言，亟需改变观点，明确树立人工智能以增进人类福祉为伦理追求的观念。

另外，由于传统风险治理模式无法有效地应对可控性问题和重大相关问题，可以得出一个重要判断：人工智能伦理治理不是单纯的风险问题，而是 AI 驱动的科技、经济、社会发展向何处去的问题。显然，要解决人类社会向何处去的问题，必须以增进人类福祉为根本宗旨。这就说明，人工智能带来重大相关问题，而重大相关问题的解决必须以增进人类福祉为根本宗旨，所以人工智能事业必须以增进人类福祉作为内在伦理追求。

总之，有必要通过教育、培训和大众科普等一切方式，在全社会明确树立人工智能以增进人类福祉为伦理追求的观念，提高所有从业者和相关者关于人工智能伦理治理的自觉性、主动性和创造性。

4 传统治理的新探索：以隐私保护为例

本节以用户隐私问题为例，探讨传统的风险治理用于合理性问题所面临的新情况和新课题。

公民享有隐私权，私人信息依法受到保护，从而不被他人非法获取、利用和公开。然而随着信息技术和网络应用的快速发展，利用私人信息 (包括

隐私信息) 提升服务性能，已在全球形成新潮流，由此导致隐私保护与性能提升之间产生张力，使得用户隐私保护成为技术风险治理中的一个新课题。

侵犯用户隐私的主要表现形式有：不恰当地采集用户的隐私数据 (未经用户授权、未恰当给出数据采集说明或授权方式等)；从采集的数据中不恰当地提取用户的隐私信息；不恰当地使用采集或提取的用户隐私信息，或者用其他方式侵犯用户隐私，比如未经授权地公开或出售用户隐私信息。下面从成因分析、责任归属和治理建议三方面展开讨论。

(1) 人工智能技术。目前条件下，一个 AI 产品采集哪些用户数据、对数据进行哪些分析 (提取隐含信息)、采集数据和分析结果如何使用，都不是 AI 技术决定的，目前 AI 技术没有这样的能力 [7-8]。同样由于其局限性，现有 AI 技术也不足以为隐私保护提供充分的技术支持 [9]。

从以上分析得出两个判断。在归因方面，现有 AI 技术是一种有缺陷的工具，该工具通过某些方式的使用产生了用户隐私问题，所以工具和工具使用者都与所产生的问题之间存在因果关系。在归责方面，由于人工智能不具备法律主体地位和责任承担能力 [1]，AI 工具使用中出现的问题不能归责于工具本身。这个结论具有普遍性，不仅适用于用户隐私问题。

(2) 产品研发。一种 AI 产品采集哪些用户数据、对数据进行哪些分析、采集数据和分析结果如何使用，是由产品研发者决定的，这里的研发者包括开发者和决策者。决策者拥有产品设计的最终决定权，所以负有更大的责任，但决策者可能不完全了解产品设计中的伦理缺陷，甚至开发者也不完全了解 [10]。所以，研发者有责任充分预估产品设计的伦理后果，企业有必要提升自身的伦理责任意识和能力，但不能简单地将用户隐私问题的全部责任归于研发者。

(3) 管理。根据行业管理惯例，正常情况下所有产品都应该有对应的技术标准和行业规范，它们规定了相关产品必须遵守的要求，包括伦理要求 [10]。然而，通常只有当产品研发和推广使用积累了相当经验之后，才能够制定出合理可行的技术标准和行业规范，所以新兴产业出现行业监管空白期是普遍现象，这个阶段容易产生企业之间的无序竞争，进而滋生伦理问题。所以，技术标准和行业监管的滞后或缺失也是用户隐私问题的原因之一。

(4) 法律。我国法律法规中包含公民隐私保护的多项规定，正在针对人工智能等新技术的应用情况制定更多相关规定，从而为新形势下的隐私保护提供法律依据。国外也在进行相关的立法，如欧盟的《通用数据保护规范》(GDPR) 于 2018 年 5 月开始生效，从个人数据处理模式、用户权利、各方义务、产品认证以及监管措施等方面设立了数据保护框架，以保护公民免受隐私和个人数据泄漏的影响。

(5) 理念。大众对隐私问题的关注度不断提高，相关企业、管理机构的重视程度也在提升之中，这种变化对于用户隐私保护具有极其重要的作用。不过，不同企业的伦理意识和治理力度差异很大。例如，某手机企业更新了它的操作系统，增加了有关选项让用户拒绝某些隐私信息被采集，从而受到了用户的普遍欢迎。同时，此举直接导致某社交平台的广告推送能力下降，一年内广告收入预期减少 100 亿美元。事实上，伦理治理已成为企业竞争的一个新维度，“二维企业”(进入伦理维度的企业) 对“一维企业”(未进入伦理维度的企业) 已形成了“降维打击”的能力。

为什么在同样的外部环境 (包括大众的隐私保护意识、行业监管和法律法规) 下，不同企业之间出现如此之大的差异？根本原因在于企业伦理意识之间的差异：二维企业的伦理意识已经觉醒，而一维企业的伦理意识尚未觉醒。明确人工智能的内在伦理追求，将促进企业员工增强伦理意识，进而激发企业伦理意识的觉醒。

5　新型治理的创新探索

人工智能伦理治理作为一种新型问题，需要以增进人类福祉为根本宗旨，在以下三方面进行创新。所以，人工智能伦理治理不是单纯的伦理约束问题。

5.1　来自可控性问题的挑战

人工智能可控性挑战包含着如下问题：不可控人工智能是否具有科学可能性？如果有，人类要不要发明、研制不保证可控性的人工智能？如果要，如何应对可控性丧失带来的巨大风险？如果不要，如何确保人工智能的可控性？

上述诸问题目前处于争论之中，尚未达成共识。例如，“强人工智能”最初是哲学家塞尔作为一种不可能实现的 AI 概念而提出的 [11]，之后出现了不同的观点。一种观点认为，强人工智能才是人工智能的正确目标和真正希望；另一种观点则认为，强人工智能将导致可控性的丧失，并从不同角度阐述了理由 [6,12-13]。还有“通用人工智能”和“超人工智能”，也存在类似情况。

关于强人工智能、通用人工智能和超人工智能是否具有科学可能性、会不会丧失可控性，只能通过科学研究才能够得出科学结论，而这种研究迄今并不存在。为了有效应对可控性问题，有必要在人工智能学科[6]中设立一个“可控性研究”的新分支，研究与人工智能可控性相关的课题。该分支的作用和意义不限于可控性问题本身，还将涉及人工智能理论基础研究中的一些根本性问题，从而带动整个人工智能学科的一次重大变革。

5.2 来自重大相关问题的挑战

针对重大相关问题，有必要探索、建立新的治理模式。由于重大相关问题的深层原因在于现行创新模式（熊彼特创新[14]）的局限性[7,10]，所以重大相关问题的治理涉及新型创新模式的探索。这种新型创新模式应更好地适应人工智能时代的人类需要，更好地为增进人类福祉发挥支撑和引导作用。

现阶段熊彼特模式产生的主要效益是GDP可度量的。然而大量分析表明，大众福利不仅来自GDP可度量效益，也来自GDP不可度量效益；而且历史上大众福利增长最快、社会最繁荣的时期，创新带来的GDP不可度量效益最显著、作用最大[15-16]。例如，抽水马桶带来的大众福利远非相关产品和设施所增加的GDP可度量；而人均寿命增长带来的大众福利更不是GDP可度量的[15]。可是，以往创新产生的GDP不可度量效益只是熊彼特模式顺带而来的副作用。这种副作用在当今时代已经显著减少，并有预测认为这种趋势难以逆转[15]。依据这些分析，新的创新模式应能更好地统筹经济效益和社会效益。

基于上述目标提出了公义创新[7,17]。传统创新的目标对象是满足市场需求的具体产品，并不考虑由此带来的长期效应对社会的深层影响；而公义创新的目标对象是符合经济、社会发展需要的人工系统[18]。人工系统的设计并非基于单纯的经济效益，而是基于经济效益和社会效益的统筹，并且比产品设计更加复杂，需要人工智能技术的强大支持[17]。在公义创新中，人工智能技术不仅是治理对象，也是必要的治理手段。这种情况完全不在分离约束观的考虑范围内，却与内在追求观相符合。

5.3 合理性问题驱动的技术创新

如用户隐私保护问题的分析表明的那样，技术应用的正负效应通常混为一体，无法彻底分离。于是，通过传统治理中的正负效应权衡，在化解负效应的同时往往会减少正效应，即使用性能下降。如果使用性能下降过多，将严重影响AI产品的用户接受度，甚至导致产品失败。

通过技术手段可以改变AI产品的正负效应结构，为解决上述难题提供新路径。例如，隐私计算技术[19]被用于减少隐私数据的利用，并保持产品性能；因果推理和可解释性研究被用于解决算法公平性问题[20]。其他正在探索的技术手段有伦理嵌入设计、追溯技术、可信技术、安全技术[21]、公平的机器学习技术[22]及其他数据治理技术。这些研究形成了人工智能研究的一个新方向——“技术AI伦理”(Technical AI Ethics)[9]。这再次表明，人工智能伦理治理不是一个单纯的伦理约束问题，同时也包含着伦理驱动的AI技术创新。

[参 考 文 献]

[1] 陈小平. 人工智能伦理导引[M]. 合肥: 中国科学技术大学出版社, 2021

[2] Dubber MD, Pasquale F, Das S. The Oxford handbook of ethics of AI [M]. New York: Oxford University Press, 2020

[3] Vieweg SH. AI for the good: artificial intelligence and ethics [M]. Cham: Springer, 2021

[4] Coeckelbergh M. AI ethics [M]. Cambridge: The MIT Press, 2021

[5] 沈寓实, 徐亭, 李雨航. 人工智能伦理与安全[M]. 北京: 清华大学出版社, 2021

[6] 陈小平. 科学实践中人工智能到底是什么[N]. 中国科学报, 2021-12-27

[7] 陈小平. 人工智能: 技术条件、风险分析和创新模式升级. 科学与社会, 2021, 11: 1-14

[8] 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, 15: 114-20

[9] Stanford Institute for Human-Centered Artificial Intelligence. Artificial intelligence index report 2022 [EB/OL]. [2022-03-17]. https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf

[10] 常荧丽. 专访陈小平: 公义创新——人工智能时代的创新模式[N]. 信睿周报, 2021-03-15

[11] Searle JR. Minds, brains, and programs. Behav Brain Sci, 1980, 3: 417-57

[12] 赵汀阳. 人工智能会是一个要命的问题吗？开放时代, 2018, 282: 31-48+6

[13] 周志华. 关于强人工智能. 中国计算机学会通讯, 2018, (01): 45-6

[14] 黄阳华. 熊彼特的“创新”理论[N]. 光明日报, 2016-09-20

[15] 罗伯特·戈登. 美国增长的起落[M]. 北京: 中信出版社, 2018

[16] 查尔斯·古德哈特, 玛诺吉·普拉丹. 人口大逆转: 老龄化、不平等与通胀[M]. 北京: 中信出版社, 2021

[17] 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, (09): 79-87
[18] 赫伯特·西蒙. 人工科学[M]. 北京: 商务印书馆, 1987
[19] Li F, Li H, Niu B, et al. Privacy computing: concept, computing framework, and future development trends. Engineering, 2019, 5: 1179-92
[20] Chiappa S. Invited talk: path-specific effects and ML fairness[C]. NeurIPS workshop on Algorithmic Fairness through the lens of Causality and Robustness, 2021
[21] 方滨兴. 人工智能安全[M]. 北京: 中国工信出版社, 2021
[22] 古天龙, 李龙, 常亮, 等. 公平机器学习: 概念、分析与设计. 计算机学报, 2022, 45: 1018-51