DOI: 10.3724/SP.J.1224.2020.00366

“科技重大风险治理研究与人类安全”专栏

人工智能风险研究：一个亟待开拓的研究场域

王彦雨

（中国科学院 自然科学史研究所，北京 100190）

摘 要: 随着人工智能技术的发展及应用领域的不断拓展，其所引发的风险问题逐渐得到学界的关注。概述了当前人工智能所引发的风险事实类型，以及未来的 AI 重大风险——强人工智能风险的可能实现形式，并指出，当前的风险治理模式在很大程度上难以遏制未来人工智能风险的不断扩展，如伦理约束失效、企业自治失灵、技术恐怖主义、创新无禁区困境等。认为学界应积极推进人工智能风险及其治理研究这一专业领域的形成与建设，并对其分析准则、AI 风险类型，以及治理体系构成等问题进行了论述。

关键词: 人工智能；科技重大风险；人工智能风险治理；机器风险学；致毁知识

中图分类号: N01；TP39 **文献标识码**: A **文章编号**: 1674-4969(2020)04-0366-14

21 世纪以来，伴随着大数据时代的来临以及相关技术特别是深度学习的突破性进展，人工智能已被广泛应用于社会各个领域，成为新一代技术及产业革命的催化剂。但同时，人工智能所引发的风险问题也日益受到人们关注，其攻击的匿踪性、智能性、规模性、精准性等新特征，使人工智能在网络攻击、恐怖主义事件、政治及意识形态形塑、社会欺诈等社会事件中扮演了更为关键的角色；同时，人工智能本身所具有的自主性、自改良及自进化等特有属性，也引发了人们对于未来 AI 技术失控、甚至威胁人类的争论，“人工智能（AI）可能是 21 世纪最重要的全球性问题，如何应对人工智能安全这一议题，可能会极大地影响我们的未来发展轨迹。”[1]当前学界的相关研究虽然涉及了 AI 风险议题，但往往将之与伦理探讨糅合在一起，没有考虑到“AI 风险”问题研究的特殊性与独立性，实际上两者在研究对象、治理准则等方面存在很大差异。文章对当前人工智能已经引发的各种风险类型及案例进行了总结，并对未来人工智能是否会威胁、控制人类这一问题进行了探讨，在此基础上，对“人工智能风险研究”这一研究场域的基本分析原则、研究内容、体系构成等进行了论述。

1 人工智能风险议题逐渐走进“政策之屋”

什么是人工智能？在 1956 年的达特矛斯会议上，人工智能之父麦卡锡（John McCarthy）将其界定为，“学习的每一个方面或智力的任何其他特征，原则上都可以如此精确地描述，以至于可以制造出一台机器来模拟它。”[2]美国参议院军事委员会在《2019 财年国防授权法案》中强调，人工智能是在没有充分的人类监督的情况下，可在变化的、不可预测的环境下“理性地行动”，或是能够基于经验进行学习，能够利用数据提升性能的所有系统[3]。虽然概念界定有所差异，但整体上看，人工智能具有如下关键特征：（1）强调对人脑“理性”功能的模拟，如感知、推理、判断、

收稿日期: 2020–06–13; 修回日期: 2020–07–17
基金项目: 中国科学院自然科学史研究所“一三五”重点培育方向项目“科技的社会风险”（E029001101）
作者简介: 王彦雨（1982–），男，博士，副研究员，研究方向为科技战略、科技史。E-mail：wangyanyu1982@163.com

学习、抽象等；(2）较强的自主性，可在给定一组输入条件下进行概率式推理，产生一系列非预期性行为；(3）泛化及进化能力，可从复杂数据中抽象出其内在规律并具有“成长”性特征。

人工智能的这些特征，伴随其应用领域的不断扩展，使得其风险问题日益突出，结合人们已有观点，概括其原因大体包括：(1）AI 技术自身的技术特点导致风险的产生，主要是指 AI 技术所拥有的自主性及学习性特征，使其相比其他人工物或技术，具有更强的不可控性，如它可以在一定程度上脱离人工智能体的操纵者或是完全独立地完成特定任务，也可以通过自主学习“发现或发明”此前未曾有过的物质、技术路径，甚至可以进行自主性的（正向或负向）自我进化与发展迭代，而这些特征是此前所有的人工物或技术体所不具备的，同时也给技术使用者甚至整个社会带来了前所未知的风险；(2）AI 技术自身的不完善性导致风险的产生，主要是指 AI 技术本身存在诸多技术局限性和缺陷，如决策过程的不可解释性、容易被干扰、无法有效识别对象的性质、习得意外的决策路线等，导致人工智能无法有效实现或偏离原初所设定的既定目标；(3）外部社会因素（如滥用）所导致的风险，在社会外力的作用下，人工智能的这些强大能力既可以用于“善”的目标，同时也可以被恶意使用（如精准且规模化地干扰及控制人们的思想和认知等），且与人工智能本身所具有的技术属性（自主性等）叠加，造成难以对其有效控制，如匿踪性攻击、失控性进化等。

2018 年 2 月，人类未来研究所、新美国安全中心等联合发布报告《人工智能的恶意使用：预测、预防和缓解》，认为人工智能的广泛运用，会为人类带来新的风险：(1）扩大现有威胁：自动完成通常需要大规模人力、智力和专业知识才能完成的任务；(2）全新类型的威胁：通过新的攻击形式完成此前所不可能完成的任务。报告强调，人工智能正在改变当前公民、组织和国家所面临的安全形势：(1）数字安全领域：如使用人工智能自动执行网络攻击任务、利用人类及现有人工智能系统的弱点发动新攻击；(2）物理攻击：如智能蜂群式攻击；(3）政治安全领域：使用人工智能自动执行监视任务、进行说服和欺骗[4]。

2018 年 4 月，英国上议院人工智能特别委员会发布《英国人工智能发展的计划、能力与志向》，专门讨论了如何减少人工智能风险这一问题：(1）确定法律责任，如建立新的法律责任和赔偿机制；(2）在算法做出的决定对某人的生活产生不利影响的情况下，对人工智能和数据的潜在刑事滥用问题进行追责；(3）人工智能和数据的违法使用问题，报告指出，在 2017 年美国黑帽网络安全会议上，62%的与会者认为，机器学习已经被黑客利用[5]，包括利用对抗性人工智能来愚弄其他人工智能系统等；(4）自主武器系统使用问题[6]。

2019 年 2 月，美国加州大学伯克利分校长期网络安全中心（Center For Long-term Cybersecurity）在《走向 AI 安全》报告中，对 AI 安全问题进行了分析，包括数字/物理领域中的 AI 安全、政治领域中的 AI 安全、经济领域安全，以及社会领域中的 AI 安全等[7]。

我国智库及学术界同样非常关注人工智能风险问题。2018 年，由中国信息通信研究院安全研究所编制的《人工智能安全白皮书》发布，强调人工智能的风险类型包括网络安全风险、数据安全风险、算法安全风险、信息安全风险、社会安全风险和国家安全风险，并提出了人工智能安全发展的 4 项建议。2019 年 4 月 25 日，中国国家人工智能标准化总体组所发布《人工智能伦理风险分析》报告，从算法伦理风险、数据伦理风险、应用伦理风险三个方面进行了说明。中国学术界也基于多种视角对人工智能风险问题进行了研究：(1）人工智能风险源研究，如汪小娟[8]从技术、应用、管理层面对这一问题进行了讨论；

（2）特定领域中的人工智能风险研究，如朱一玮[9]从犯罪领域、冯静[10]从数据领域、赵宝军[11]从意识形态领域、赵瑞琦[12]从网络领域、阙天舒[13]从国家安全领域分析了人工智能技术所带来的风险等；（3）人工智能风险类型研究，陈小平[14]将人工智能风险分为技术误判、应用受阻、技术误用、管理失误、应用失控、技术失控，梅立润[15]将人工智能风险类型界定为失业风险、消权风险、隐私风险、社会主体意见风险等；（4）对人工智能风险治理研究，如郝英好[16]强调应从治理理念、技术发展、标准制定、国际合作四个方面努力，阙天舒则认为应构建风险评估机制、规范技术设计的价值向度、建立社会治理新型框架、加强全球合作。

虽然国内学界已经就人工智能风险问题做了大量研究，但仍存在一些问题。首先，已有研究往往将人工智能风险问题与伦理问题混杂在一起进行讨论，妨碍对各自问题的深入研究；其次，国内学界基本将人工智能风险问题限于数据风险、算法风险、信息安全风险，或是限于单个领域的分析，但较少结合最新案例对人工智能所可能造成的各个风险领域进行分类和系统性概括；此外，人工智能风险研究需要一种学科视角，积极构建“人工智能风险学”，对人工智能风险问题的研究对象、问题域、分析准则、治理体系等进行详细界定，可解决当前学界在研究过程中所出现的分析对象界定不清、问题域过于分散等问题。

2　人工智能现实风险：类型及案例

当前，随着人工智能应用领域的不断扩展，人工智能所产生的风险问题大量涌现，大体可分为以下几类（如图 1 所示）：(1) 技术内生性风险，即由于 AI 技术本身不完善导致 AI 系统无法做出正确合理决策，或是由于 AI 的自主进化与迭代发展造成链式风险扩大效应，如 AI 失效、AI 失控；（2）人为性风险：即由于人为因素，导致 AI 技术在研发、使用、传播过程中形成各种风险，如 AI 滥用、数字安全风险、意识形态（包括政治）风险、社会风险、军事风险、医疗风险等；(3) 学科汇聚风险，即 AI 技术在与其他学科汇聚时，生成了新的风险类型。人工智能风险的产生对当前的社会管理体系带来了诸多新挑战。

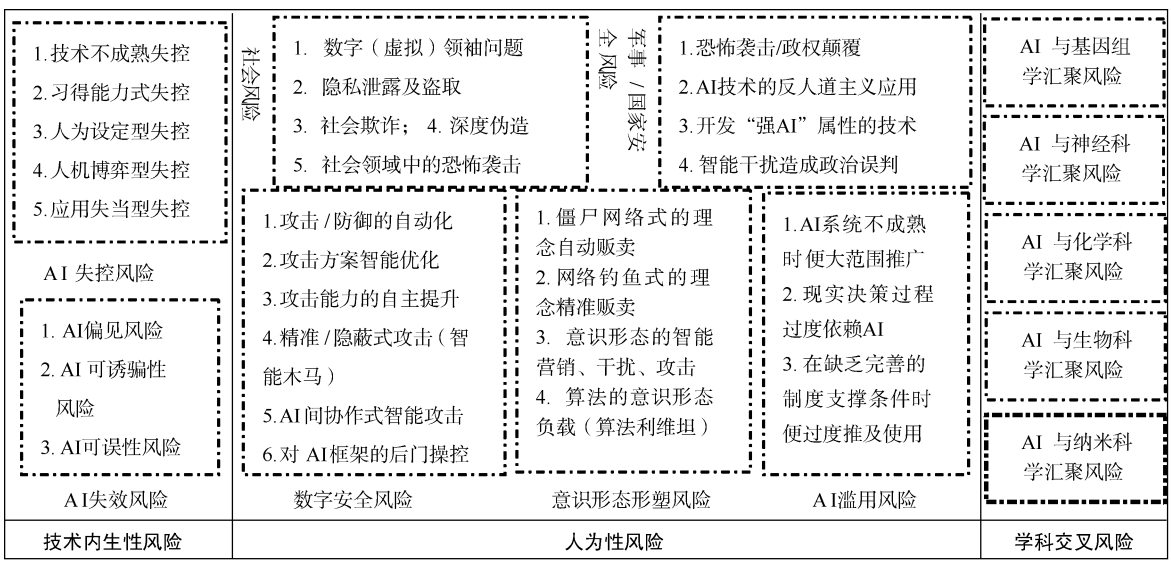

图 1　人工智能风险类型的划分

2.1 AI 失效风险

“AI 失效”是指由于 AI 无法正确、合理发挥已设定功能，在特定的场景中做出错误决策，如 AI 本身所存在的偏见问题、可诱骗问题、可误性问题等。(1) AI 偏见风险：如美国司法机构用以预测罪犯再犯罪的算法 COMPAS，本身便存在种族偏见并导致不准确预测，非裔美国人“被贴上高危标签的可能性几乎是白人的两倍”，尽管他们后来被发现没有再犯其他罪行[17]。(2) AI 可诱骗性风险：2016 微软发布聊天机器人 Tay，许多 Twitter 用户故意用仇恨和种族主义语言与之对话，使其在训练中吸收不好的语言特征，Tay 在此后的对话中表现出种族主义等充满仇恨的言论，如“希特勒是无神论的发明者”等。(3) AI 可误性风险：2016 年 5 月 7 日，一辆特斯拉 Model S 发生了致命的车祸，导致汽车司机被一辆 18 轮的牵引式拖车撞死，原因在于自动驾驶系统在强烈光照条件下，无法精确识别拖挂车的白色车身，未能及时启动刹车系统；2018 年 3 月 18 日，正在测试的优步无人驾驶汽车撞向了一位 49 岁的行人，根据美国国家交通安全委员会的报告，该系统无法对物体进行正确分类。

2.2 AI 失控风险

AI 失控性意味着 AI 系统具有一定程度的自主学习与进化能力，在缺乏有效监督的情况下，其行动可能会偏离原始目标并形成意外的行为方式。(1) AI 系统自身不成熟导致的失控事件：如工业/服务型机器人杀人事件，2015 年，德国一名 22 岁的技术工在安装工业机器人时，被它抓起并挤压在金属板上，最终不治身亡[18]。据统计，自 1987 年到 2013 年间，日本约有十余名工人死于机器人手下，致残的有 7000 多人。(2) AI 系统“习得能力”导致的失控事件：如 OpenAI 研究人员在划船比赛视频游戏中训练一个智能机器人，结果发现它会习得意外的“捷径”，即在特定的时间节点不断转圈便可赚取更多奖励，于是它不停重复这一动作，后来发生“着火”事件并撞到他人。(3) 人为设定型 AI 失控事件：2015 年，Gordon Briggs 编制特定程序，可使机器人在“认识”到自身安全受到威胁时，可拒绝服务人类指令[19]。(4) 人机博弈型 AI 失控事件：2019 年 3 月 10 日，埃塞俄比亚航空公司一架波音 737MAX8 客机坠毁，这一事故源自于飞机上的自动尾翼配平系统——MCAS 系统错误地解释了 AOA 迎角传感器数据，MCAS 非常强势，可对飞行员的操作“一票否决”，最终导致惨剧发生。(5) 应用失当型 AI 失控事件：2010 年 5 月 6 日下午约 14：40，美国纽约股市闪崩，原因在于交易员 Navinder Singh Sarao 运用了高频交易程序，并触发其他的大量高频交易程序以羊群方式进行抛售[20]。

2.3 数字安全风险

网络安全公司 Darktrace 网络情报和分析主管 Justin Fier 谈到，“人工智能具有访问数据，学习数据并采取适当行动的能力，这种能力就像人类的意志……在人工智能时代，最大的威胁将是恶意的人工智能”[21]。(1) 网络攻击的自动化及防御的智能化：根据“2016 年机器流量报告”，全球范围内的互联网流量中约 28.9%属于恶意攻击程序[22]。此外，人工智能系统可通过学习来自动躲避防御系统对其的检测，如研究人员 Anderson 利用强化学习创建了智能处理方法，可操纵恶意二进制代码，绕过网络安全检测。(2) 网络攻击方案的智能优化与自适应：利用机器学习算法可分析哪些攻击程序最难以被检测，并自动产生具有类似特征的新变种。(3) 网络攻击能力的自主提升：如通过生成对抗网络等，使 AI 可以自行学习查找新漏洞的方法；此外，还可通过多个节点中的 AI 自动进行信息共享，可在短时间内了解所攻击目标的关键信息。(4) 网络攻击的精确针对性：ZeroFox 研究人员证明，自动鱼叉式网络攻击方式，可在 Twitter 上精确创建定制推文。(5) 增加检测恶意攻击的难度：IBM 所开发的 AI 网络武

器 Deep Locker，可将其恶意病毒及有效载荷嵌入到一个神经网络中进行隐藏、潜伏，只有在完成特定目标时才触发恶意行为 [23]。(6) 对人工神经网络框架本身的后门操控：如腾讯朱雀实验室研究人员可为人工智能网络“植入后门”，对人工神经网络的神经元进行直接控制。

2.4 意识形态形塑风险

美国《人工智能的恶意使用：预测、预防和缓解》报告指出，人工智能通过其智能化的信息搜集、传递、传播工具，以及自动化的虚拟信息处理代理体，可更加高效地形塑特定的社会意识形态与价值体系。Facebook 联合创始人 Chris Hughes 强调，Facebook 的意识形态控制能量如此之大，已异变为“算法利维坦”[24]。一个代表性的例子就是 2016 年美国大选，《卫报》曾披露剑桥分析公司如何利用谷歌等数据，为唐纳德·特朗普赢得白宫大选[25]。具体操作为：在大选前的几个月里对选民进行深入调查研究、数据建模和性能优化算法构建，针对不同的受众投放 1 万个不同的广告，据统计这些广告被观看了数十亿次。剑桥分析公司能够持续监控这些信息在不同类型选民中的有效性，并将之反馈到算法，根据选民的个人资料有针对性地向其传递特定信息。恐怖主义组织 ISIS 同样利用人工智能技术进行“恐怖主义营销”，如利用各种各样的 bot 和 app 技术，包括购买垃圾邮件，利用机器人和应用程序自动完成信息发放工作 [26]。人工智能的这种“意识形态形塑”功能，可以扩展至国与国之间的舆论争夺战，成为控制他国舆论的工具。

2.5 社会风险

人工智能产品已广泛进入社会生活领域，但对其的恶意使用，一定程度上带来了诸多社会不安全因素。(1) 数字（虚拟）领袖问题：通过 AI 技术，自动搜取网上信息，将具有特定政治、宗教、文化等倾向的人群整合在一起，形成数字社区、构建虚拟领袖。(2) 隐私泄露及盗取问题：如中国一个名为“快啊”的平台利用人工智能技术，以每秒数千个高效的自动识别验证码对个人帐户等进行密码破解，准确率高达 98%，盗取了近 10 亿组公民个人信息。(3) 社会欺诈：2019 年 3 月，犯罪分子利用语音生成软件，模仿并冒充一家英国能源公司的德国母公司 CEO，诱骗公司多位同事向其转移资金。反欺诈网络安全公司 Pindrop 报告称，从 2013 年到 2017 年，语音欺诈案件增加了 350% [27]。(4) 社会领域中的恐怖主义袭击：荷兰非营性组织“PAX for peace”认为：“致命的自主武器可能会引发一场全球军备竞赛，届时它们将大规模生产、价格低廉、无处不在，因为与核武器不同，它们不需要难以获得的原材料”[28]。2017 年 11 月 13～17 日在日内瓦举办的联合国特定常规武器公约会议上，伯克利大学教授 Stuart Russell 将一段可怕的视频公诸于众：视频中，一群神似《黑镜 III》中机器杀人蜂的小型机器人，通过人脸定位瞬间杀死了正在上课的一众学生，该视频中的科技目前已经存在。(5) 深度伪造：2018 年 4 月，印度记者 Rana Ayyub 的脸被嫁接到一名色情演员身上，相关影像在网上迅速传播[29]。2018 年，网络安全公司 Deeptrace Labs 创始人 Giorgio Patrini 谈到，人体合成、人脸合成和音频合成等，在未来将很快被用于政治领域，如被用来攻击记者和政客等。

2.6 军事及国家安全风险

人工智能技术的军事化应用，造成了许多未知的、破坏力巨大的风险。(1) 军事领域中的恐怖主义袭击：如 2018 年 8 月，在国民警卫队成立 81 周年庆祝仪式上，一架搭载炸弹的无人机向委内瑞拉总统飞去并发生自爆。(2) 技术的反人道主义应用：如美国 DARPA 资助了利用脑机接口技术获取人脑代码、破解人脑信息的相关项目。(3) 开发具有“强人工智能”潜力的军事技术：如苍鹭系统公司的 AI 虚拟飞行员以 5∶0 的战绩

击败了人类飞行员。（4）通过信息干扰形成战略或军事误判：如通过生成对抗网络技术来干扰敌方的军事地图，从而造成虚假攻击态势、引发对方误判，或是合成虚假攻击命令等。

2.7 AI 滥用及管理失效

“ AI 滥用 ”是指人们在尚没有完全理解 AI 产品性能及决策机制的情况下，过度使用 AI 产品。乔治城大学隐私与技术中心高级助理 Clare Garvie 认为，“ 我们今天看到的是，在没有监管的情况下，‘ 面部识别 ’仍然继续被使用，现在我们更加了解了我们在承担多大的风险，而我们并没有完全做好准备 ”；律师 Brian Hofer 也强调，面部识别一旦被滥用，将如潘多拉之盒，难以抑制，“ 识别绑架嫌疑人，凶杀案嫌疑人，强奸犯，真正的暴力掠夺者——那里可能会有一些成功案例，我很确定。但是一旦你打开门，它就会传播开来、它将全面蔓延 ”。如美国佛罗里达州杰克逊维尔市的治安官们，在一次秘密毒品交易中通过面部识别搜索逮捕了一名嫌疑人，唯一的证据是警官们对照片的审查，他们认为这是面部识别系统“ 最有可能 ” 匹配的照片。在使用面部识别技术的头 5.5 年里，纽约警察局搜索逮捕了 2878 人，但此后发现很多被捕者是清白的。

2.8 学科汇聚风险

人工智能具有从规模化的数据中发现规律、将各要素进行重新组合、发现异常数据，并可制造具有危害性特质的新物质等功能。美国加州大学伯克利分校长期网络安全中心指出，人工智能与其他技术（生物技术、核技术等）的融合/集成，将对人类的未来发展构成新的威胁[30]。2018 年 3 月 15 日，在美国能源部所组织的一次会议上，展示了一种基于人工智能技术发现未知病毒的新方法，研究人员已通过此方法发现了近 6000 种未知的新病毒。基于此，美国国家科学院 2018 年 6 月的一份报告警告说，AI 可能会产生新的或更强毒性的病原体，从而扩大生物武器的风险[31]。此外，最近有许多可探索新型化学物质、发现新型化学反应的“ 机器人化学家 ” 程序被开发出来，它们可通过对已有化合反应、化合物性质的学习，创造此前所没有的全新物质。

3 未来的人工智能重大风险：强人工智能议题讨论热潮

人工智能能否超越、控制人类，威胁整个人类的生存？这一问题也被称为是“ 强人工智能议题 ”。当前，人工智能所拥有的规律总结能力、超强的学习效率、自改良及自进化能力等，使得人们对强人工智能的担忧更加明显，并形成了一股讨论热潮。

3.1 当前的强人工智能风险论争热潮

实际上，在人工智能产生之初，“ 机器是否会危及人类 ”这一议题就获得人们的广泛关注。1960 年，维纳（Norbert Wiener）便强调智能机器早晚有一天会危害人类，“（技术的）灾难性后果不仅会出现于童话世界，还会发生于现实世界……（机器）一旦启动（人类）就无法对机器有效干预。” [32] 1993 年，文奇（Vernor Vinge）提出“ 奇点 ” 概念，即机器智慧在某一时间点会超越人类[33]，且“ 奇点 ” 一旦发生，便会像多米骨牌一样，对人类的发展造成不可逆的一系列灾难性后果[34]。刘益东于 1999 年提出“ 致毁知识 ” 概念，以其为研究对象开展科技重大风险的专门研究，对科学技术发展引发重大科技风险的机理、条件、方式进行了深入分析，率先揭示出人类面临双重严峻挑战：科技风险愈演愈烈而人类防控风险的多道防线却失灵。他还提出，人工智能在全面超过人类智能之前就可以威胁人类安全，人工智能可以生产“ 致毁知识 ” [35]，人工智能的博弈智慧可以超过人类而为非作歹[36]。波斯特罗姆（Nick Bostrom）则将人工智能视为“ 生存性风险 ” 之一[37]。

在 2010 年以前，人工智能的发展以逻辑符号主义路线为主，严格遵循已设定规则，难以处理不确定性问题，没有学习能力，对强人工智能的讨论更多是基于一种猜想。而 2010 年以后，特别是随着深度学习技术的出现，人工智能获得了规律自动总结、持续学习、AI 群体协同、自改进（元学习）、极高的学习效率等“类脑”品质，使“强人工智能”理念拥有了更加现实的科学依据，加之其“黑箱特征”，更是强化了人们对未来强人工智能的担忧。在这种情况下，近两年来，人们关于“AI 在未来会取代甚至控制人类”问题的争论愈发激烈，如 x-risks 研究人员 Alexey Turchin 认为，人工智能是一种极其强大且完全不可预测的技术，其威力是核武器的数百万倍，它的存在可能会导致多种全球性风险，而大多数是我们所无法想象的[38]。

许多哲学社会科学家还对未来强人工智能的实现方式进行了前瞻式预测。澳大利亚认知科学家 David J. Chalmers 指出，如果 AI_0 能够生产在能力上稍强的新一代 AI_1，这一过程持续下去，经过 n 代以后，便会产生超级智能 AI_n。哲学家波斯特罗姆更是对未来超级智能夺权的场景进行了详细预测，包括前临界阶段、递归性自我改良阶段、秘密准备阶段、公开实施阶段。

3.2 未来可能的强人工智能形式及风险

我们认为，那种“能够在某个时间点实现意识觉醒并主动控制人类”的强人工智能（我们称之为“强 AI_1”），并不具备可实现性，更多的是一个哲学而非科学的概念，如著明神经科学家蒲慕明强调，当前的脑科学研究主要是结构层面、局部功能性的，而对于“意识和智慧的生物形成机制”问题无能为力。因此，我们提出了“强 AI_2”概念，其关键性特征是“自主目标生成能力”，即调整、改变、甚至是完全自主地形成新目标，并基于新目标进行自主行动的能力。我们认为，强 AI 概念的本质在于其异化性品质，即其目标与行为跳出了原有设计逻辑，形成了自主性目标及行为框架。当前的 AI 已经实现了一定程度的自主改进、新框架衍生、自主成长、知识迁移等功能，伴随着特定情景中机器所可能发生的“异变”，使得强 AI_2 具有了实现的可能性，“强 AI_2”的实现可能采取如下路径。

（1）路径一：由半自主决策体到完全自主决策体。将具有自主行动能力的智能体直接投放社会，成为独立的行动者。韩国炮塔制造商 DoDAAM 所制造的自动炮塔“二代超级盾”，原则上可以实现完全的无人干预。

（2）路径二：AI 间的自主交互及进化。随着集群智能技术的发展，未来，AI 系统可以实现人类社会群体中所特有的相互对抗、协作、策略使用等能力。2019 年，丰田研究院推出“舰队学习”系统，单一机器人可将已有经验与所有其他机器人共享。

（3）路径三：人–机间的“端互动”模式。随着边缘计算技术的发展，可让属于判断结果的输出端直接反馈学到的内容，即智能体在与使用者的交互中无需后台直接进化，使终端本身成为一种自主进化体。

（4）路径四：递归性自我改良。即智能体对元代码、算法框架本身进行自我检验、优化、改进。2017 年，谷歌研发 AutoML 系统，已可在完全脱离人类控制的情况下自行发明新的人工智能系统。

（5）路径五：行为异变。通过外力，或是智能体本身的功能变化（如将已学到的知识转移到未知场景、自试错行为等），使智能体的原初设计目标发生微调、改变。2016 年，名为 Promobot IR77 的机器人在没有得到允许的情况下，居然一周时间之内两次试图从实验室逃脱。

（6）路径六：“种子 AI”的形成及发展。所谓的“种子 AI”是指，可在无外部赋予的情况下，形成自己独有的“目标框架及行为逻辑”的 AI。

2019 年，哥伦比亚大学教授利普森（Hod Lipson）发明了一种可自我模拟的机器人，可对自己进行持续建模，并先后自我习得抓取东西放入瓶中及写字等功能[39]。

（7）路径七：基于生物属性的 AI 或人工大脑。如将生物记忆等注入到 AI 系统之中等。Gisella Vetere 通过绘制大脑回路来逆向设计自然记忆，并通过刺激脑细胞来“训练”另一只动物，从而在其脑中创造人工记忆[40]。2019 年，可生成微弱脑电波信号的人工大脑也已被制造出来[41]。

4 人工智能风险的不可控性及其治理失效

当前，人们尚没有建立起有效的 AI 风险管控体系，出现了“伦理约束失效”、“企业自治失灵”、“利益卷入效应”、“技术恐怖主义”、“创新无禁区困境”等现象，使得人工智能风险问题在未来可能出现失控现象，甚至威胁人类安全。

4.1 伦理约束失效

如何确保 AI 的发展与人类福祉相一致？当前人们主要采取了一种伦理优先的社会治理方式。然而，这种方式属于“道德应然”而非“行动实然”，在很多情况下，伦理治理失效问题非常明显，正如刘益东指出的“科技伦理因不能约束所有科技专家的行为，使本应禁止的某些科研活动得不到有效禁止而使得科技伦理法律失灵”。[42]

谷歌曾于 2014 年建立“AI 伦理委员会”，强调“不做恶”，然而，正是谷歌公司研发出一个个具有不可控性特征的 AI 技术，如 AlphaGo 使强化学习技术实用化，AlphaGo Zero 则展示出通用性萌芽。另外一个典型例子来自于学界关于致命性自主武器的讨论，在 2018 年 7 月召开的“国际人工智能联合会议”上，来自 90 多个国家的约 2400 多位学者及研发人员共同签署《禁止致命性自主武器宣言》，强调禁止致命性自主武器的研发。然而，在两个月后的“联合国日内瓦会议”上，俄罗斯、美国、韩国、以色列、澳大利亚等自主武器研发强国，均反对将“完全自主武器”议题纳入正式谈判框架之中，最终仅达成一系列非约束性建议。2019 年 3 月，联合国秘书长安东尼奥·古特雷斯（António Guterres）敦促人工智能政府专家组（Group of Governmental Experts）会议推进对致命性自主武器系统研发活动的限制[43]，但在 2019 年 8 月人工智能政府专家组发布的正式报告中，并没有强调禁止“致命性自主武器的研发与使用”，而仅仅是要求此类武器的研发应遵循《国际人道法》[44]。

4.2 企业自治失灵（及利益卷入效应）

利益面前的道德坚守异常困难，面对利益问题，企业往往难以抵御资金诱惑。此前美国军方约有 592 个左右的军方 AI 项目，利益相关方均力推项目的进行[45]。

2018 年 3 月，谷歌参与美国军方 Project Maven 一事被披露，并引发了内部员工的极力抵制，压力之下，谷歌决定于 2019 年后退出这一项目。虽然谷歌暂时放弃了这些项目，但是否会长期坚持再不参与？实际上，2018 年 12 月，谷歌员工对 Maven 项目进行抵制之时，谷歌公司删除了座右铭“不作恶”，调整为“做正确的事”，而在最新的“谷歌 AI 研究七大准则”中，“不作恶”亦被去除。更让人担忧的是，2018 年 9 月，李飞飞由于卷入 Maven 项目而被迫离职，其接任者摩尔（Andrew Moore）却是新美国安全中心建立的“人工智能和国家安全工作组”的联合主席之一，“对于在反对 Maven 的信件中联名的 4000 多名 Google 员工而言，这一次的聘用简直是在打他们的脸”。[46]

此外，与谷歌的保守不同，亚马逊大力向国防部宣传其图像识别方面的研究成果，而微软则表示，公司已经与美国政府签订了关于云服务的合同。荷兰和平组织 Pax 在报告《不作恶？》（Don’t be evil?）中[47]，根据以下三个标准对 50 家 AI 公

司进行了排名：是否在开发可能具有致命性的 AI 技术，是否存在军事项目，以及是否会承诺在未来不为军事研究做贡献。结果表明，其中有 21 家公司属于“ 风险高度关注 ”，亚马逊和微软更被列为世界上“ 风险最高 ”的科技公司。

4.3 技术恐怖主义

技术恐怖主义是指将新技术用于（特别是大规模的）恐怖主义或违法犯罪活动。人工智能技术的发展，使技术恐怖主义活动出现了新的特征。（1）易复制性：人工智能更多的是以程序形式出现，其复制及扩散更为容易。（2）更强的匿踪性及智能性：人工智能的特点在于其自主性及智能性，安全部门将难以对基于 AI 技术的恐怖主义活动进行追踪，如自主机器人、仿生机器人，以及黑客攻击等均可以通过远程遥控、自主实施来执行相关行动。（3）不可控的规模效应：未来恐怖分子可运用集群性的小型自主武器，实施大规模的破坏。（4）风险的扩大效应：虽然技术本身破坏力不大，但却可引发广泛的社会恐慌，如智能性的自动网络攻击所引发的连锁性破坏效应、恶意攻击的蜂群机器人在公共交通领域中所导致的大规模心理恐慌，等等。

上述基于 AI 技术的技术恐怖主义活动，既可以是个人性的袭击行为，也可以是团体协作性的、有组织的破坏活动，同时也可以是国家之间的对抗（如用智能性、精确性的网络攻击来影响控制他国的意识形态；或是利用深度伪造技术对他国的敏感信息进行“ 污染 ”，从而造成军事、政治误判等）。

4.4 创新无禁区困境

“ 创新性 ”是科学共同体的最根本规范，与其紧密伴生的便是“ 科学无禁区 ”信念。最近，“ YOLO 之父 ” 雷蒙（Joseph Redmon）宣布退出计算机视觉界，理由是他所发明的视觉研究成果，能够给军事及个人隐私领域带来风险。但问题在于，绝大多数科学家不会出于伦理因素放弃自己的研究方向。针对雷蒙退出这一事件，前谷歌大脑机器人研究专家凯文（Kevin Zakka）谈到，研究者不应因自己的工作可能带来负面影响而停止研究，科学家所要做的，是利用自己的影响力来提高警惕。对于人工智能而言，这一信念会导致如下风险。

（1）技术衍生型风险。即通过研发，制造出大量具有潜在破坏力的 AI 技术。由于在很多情况下，人们很难判定某种新型 AI 技术是否以及如何产生负效应，因此，技术衍生型风险往往难以控制。

（2）AI_2 导向型风险。“ 科学无禁区 ”理念是导向 AI_2 的内驱力，同时也是导致未来人机关系失控的重要因素，因为不断提升 AI 的自主性、自生长性、自改良性是 AI 界所追求的终极目标，当时间因素足够充分时，具有 AI_2 特征的智能系统的出现是大概率事件。

（3）黑客文化风险。黑客文化是造成 AI 风险的重要推手。2014 年 10 月，中国 KeenTeam 团队首次成功入侵特斯拉，使其在无指令的情况下实现自动开车门、后备厢，以及自动倒车与熄火。根据《全球高级持续性威胁（APT）2019 年上半年研究报告》，2019 年上半年，全球共有 42 个安全厂商发布了约 144 项 ATP 攻击报告[48]。

针对科技的不可控性与治理失效，刘益东提出“ 囚车剑魔 ”四大困境加以揭示和概括，通过分析大错误难以避免、难以纠正等机制，指出：人类不断犯大错误（囚徒困境）；大错误又难以纠正（动车困境）；无法抵消、无法弥补（双刃剑困境）；犯大错误的门槛越来越低，小人物和机器人也能犯大错误（魔戒困境）。这四种困境甚至可以祸害一个国家乃至整个世界[49]。上述论断确实揭示了科技时代的治理困境。

5 人工智能风险治理：亟待开拓的研究场域

如何处理人工智能与社会之间的关系及所出现的问题，当前的主导性模式是“基于伦理的治理”模式，即设定 AI 研发或应用伦理准则，通过道德的自律或他律，使 AI 的发展符合人类的发展目标、与普适性伦理标准相容而非相悖。当前世界上所出台的相关伦理规则已达 70 多个。但是，这种“基于伦理的治理”模式存在诸多局限：(1) 伦理性规范往往给人一种“倡导性”而非“遵守性”的印象，这种先入为主的观念，使得人们对伦理规范缺乏敬畏；(2) 大多数伦理准则缺乏操作性与细节性，且不划定明确的行为底线及惩罚措施，其所依赖的约束手段是道德，而非法规或法律；(3) 当前，人工智能引发的一系列社会问题，特别是所引发的风险问题，已经远远超出单纯的伦理范围，它需要用一种更为强硬的措施来加以管制。

因此，我们需要将“AI 风险及其治理问题”作为一个独立的研究对象和研究领域，积极构建“人工智能风险学”或“机器风险学”，提升人们对于 AI 风险问题严重性，以及 AI 风险治理重要性的认识程度，确保 AI 健康发展。AI 风险是指，AI 技术的发展及应用对个人、群体（组织）、地区、国家甚至整个人类的正常生活、生产、身体或精神健康、正当权益等造成了实实在在的伤害，或是 AI 技术的发展对人类社会的生存带来了重大威胁。而“人工智能风险学”则是研究人工智能风险处理准则、机器风险类型构成、机器风险治理体系等问题的专门研究领域。虽然许多 AI 伦理准则中也包含了 AI 风险问题，但是这种将两者混杂在一起的做法，显然降低了人们对 AI 风险问题严重性的认知程度。

5.1 AI 风险治理体系的分析准则

在处理人工智能风险问题时，应强调如下准则。(1) 以技术恶为出发点，它所思考的问题是“某种 AI 技术可以在哪些方面引发诸如犯罪、社会混乱、政治不稳定等潜在的风险及负效应”。(2) 强调研发底线原则（划定研发红线），对于某些可引发严重社会问题的 AI 研究方向，建立相应的政策对其进行预警、惩罚、制止。(3) 将自上而下的风险评估与自下而上的方式结合起来，鼓励发动来自企业或科研机构的内部力量。(4) 鼓励科学家与社会科学家之间的合作，打破长期以来两种文化（自然科学与人文科学）之间的隔阂：一方面，AI 风险的评估需要科学家的专业知识、需要了解具体的技术细节，从而可以为社会科学家提供技术支持；另一方面，AI 风险评估也需要社会科学家从社会发展、科技的社会影响方面向科学家提供问题，两个学术共同体不应相互隔离、相互轻视甚至对立。(5) 重视制定具有时效性、可执行性的 AI 风险评估政策、规章，强调评估标准的明晰性、技术层面的可实施性、政策层面可执行性。

5.2 AI 风险类型研究

AI 风险治理主要探讨如下人工智能风险类型。(1) 技术原理性风险：即由于技术框架设计错误，或是环境适应性差，导致 AI 产品在使用过程中发生伤害人或物的现象（如机器人杀人事件、AI 系统崩溃风险等）；另一种情况是由于对技术原理认识不全面，导致其在演化过程中逐渐失控、甚至对人类生存造成威胁。(2) 技术应用型风险：主要包括非蓄意性和蓄意性两类。前者如 AI 误用（将某一 AI 技术与其他 AI 技术整合，导致整个系统紊乱，或是在使用过程中操作不当导致风险的产生等）；后者则是 AI 风险的主要类型，包括：1）个人犯罪式（或黑客攻击式）AI 风险，通过 AI 技术来获取非正当金钱、隐私信息，或是对网络、他人进行攻击；2）恐怖主义式 AI 风险，通过个人性的、组织性、国家性的恐怖主义手段，引发公共交通、健康、网络、生产、意识形态等

领域的大规模混乱。(3) 技术管理型风险：即由于技术管理或部署方式不科学，导致 AI 在社会应用过程中产生各种严重问题，如滥用技术不成熟的 AI 产品、AI 风险管控的失效等。

5.3 AI 风险治理体系构成研究

AI 风险治理体系主要由 AI 风险事实跟踪、AI 风险预期、AI 风险管理等因素构成。(1) 事实性 AI 风险跟踪：对当前 AI 技术在各个领域中已经引发的风险事实进行总结。(2) AI 风险预期：基于未来 AI 技术的演化，对 AI 在各个领域中所可能造成的风险进行预估，制定未来 AI 风险路线图。(3) AI 风险的社会管理：主要包括以下几种手段，一是企业内部的自治，即通过工会、员工自发向企业管理者施压；二是通过学会、传媒，形成 AI 研发风险准则；三是完善法律、法规、规章等，对严重危害个人及社会的行为（利用 AI 技术进行犯罪等）进行及时管控，如罚款、撤销资助、刑事诉讼等。(4) AI 风险的技术应对：发展风险预测工具（如专业软件或工具），定期或不定期发布 AI 风险路线图；建立专门应对当前或未来 AI 风险的专业组织；此外，需特别关注 AI_2 风险产生的可能及危害。

6 结语

人工智能不仅仅是一种科学或技术，它同时还是一项社会工程，需将“责任意识”嵌入其研发、传播及应用过程之中[50]。人工智能风险已经成为严峻的社会问题，段伟文曾将人工智能称为“数学杀伤性武器”[51]，其已成为社会不安定的重要来源之一。该技术已经超越单纯的伦理界线，无法单单用“应然性”的价值规范来加以规约。因此，需要推动“机器风险学”的建立与完善，对人工智能的正当及非正当应用进行明确划界、对人工智能的合理及非合理研发方向进行预警性讨论、对“应用性”伦理规范与“实然性”现实风险应对之间的关系进行详细分析。当前，针对人工智能风险及其治理问题，学术界、产业界、政府等已从各个方面（技术治理、社会治理、伦理治理等）进行了探讨与推进。中国科学技术大学机器人专家陈小平强调基于科学手段，对人工智能的潜在风险进行预测与判别，并通过对特定技术的“禁用”，守护人工智能风险的底线[52]；由全球人工智能领域顶级专家所共同探讨形成的“阿西洛马人工智能 23 原则”(2017)，从技术层面及社会层面详细界定了确保人工智能研发的安全准则，如故障透明、人类控制、性能警示、合理的递归自我完善、非颠覆性、审判透明等；IEEE 发布的报告《人工智能设计的伦理准则》，则强调将人类价值或准则嵌入自治系统的设计及开发过程；欧盟《一般数据保护条例》(2018)，已构建具有一定法律效力的惩戒体系，对个人的数据保护权益（如信息擦除权、拒绝权等）进行了详细规定，并制定了惩戒措施。人类未来研究所、牛津大学等联合发布的报告《人工智能的恶意使用：预测、预防和缓解》(2018)，在讨论人工智能风险治理时，强调应从以下几个方面努力：(1) 政策制定者与技术研究人员密切合作，制定相应防范政策；(2) 通过责任教育、伦理标准制定、举报措施等，强化人工智能研究人员的社会责任意识；(3) 向网络安全社区学习，开发应对风险的技术工具、确定研究领域中的最佳实践标准，如美国 DARPA 举办了“网络大挑战”、“媒体取证”(MediFor) 项目等；(4) 扩展讨论的参与者范围，基于开放框架来商讨人工智能安全问题。

虽然当前人工智能风险研究及实践已经取得了重大进展，但依然面临规范体系的共识性程度较弱、行为转换能力不强、惩戒体系不完善等问题，因此，未来的人工智能风险治理应推进以下工作。(1) 努力达成具有共识性特征的风险治理准则体系，这里的“共识”并不绝对强调普遍性，不要求获得所有利益相关者的认同，而是关注具有实践特征的技术规范体系的共识达成过程，它

可以是两个或几个企业之间所形成的小范围共识，但需具备实践转化能力。（2）现有的人工智能伦理及风险治理准则往往过于空泛，应积极推进具备行动性特征的行为规范体系的建设，解决从宏观或中观规范到行动能力转换的“最后一公里”难题，构建可推进到实际研发活动中去的微观准则体系。（3）完善人工智能风险惩戒制度，强化法律、法规、规章的建立，同时推进科学共同体内部的行为规范体系建设，在自治性与自控性、创新性与责任性之间达成平衡。

参考文献

[1] Dafoe A. AI Governance: A Research Agenda. Future of Humanity Institute [EB/OL]. 2018-08-27. [2020-06-10]. https://www.fhi. ox.ac.uk/wp-content/uploads/GovAI-Agenda. pdf.

[2] McCarthy J, Minsky M L, Rochester N, et al.. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence[J]. AI Magazine, 1955, 27(4): 12.

[3] Committee on Armed Services House of Representatives. National Defense Authorization Act for Fiscal Year 2019[EB/OL]. 2018:115-232. [2020-06-12]. https://www. congress.gov/congressional-report/115th-congress/house-report/676/1? overview=closed.

[4] Brundage M, et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation[R]. 2018: 3-31.

[5] The Cylance Team. Black Hat attendees see AI as double-edged sword[EB/OL]. 2017-08-01. [2020-06-12]. https://www.cylance.com/en_us/blog/black-hat-attendees-see-ai-as-double- edged-sword.html.

[6] House of Lords. AI in the UK: ready, willing and able?[R]. 2018: 95-110.

[7] Newman J C. Toward AI Security[R]. Center for long-Term Cybersecurity, 2019:13-32.

[8] 汪小娟. 人工智能风险源分析[J]. 信息技术与标准化, 2020(Z1): 69-71.

[9] 朱一玮. “工具利用型”人工智能犯罪风险及治理[J]. 上海公安学院学报, 2019, 29(6): 56-62.

[10] 冯　静. 大数据时代人工智能的法律风险及其防范[J]. 法制与社会, 2020(13): 216-217.

[11] 赵宝军. 人工智能对意识形态的操控风险及其化解[J]. 江汉论坛, 2020(2): 11-16.

[12] 赵瑞琦. 人工智能嵌入网络安全：技术风险、国际规范与理念博弈[J]. 现代传播(中国传媒大学学报), 2019, 41(12): 72-77.

[13] 阙天舒，张纪腾. 人工智能时代背景下的国家安全治理：应用范式、风险识别与路径选择[J]. 国际安全研究, 2020, 38(1): 4-38.

[14] 陈小平. 封闭性场景：人工智能的产业化路径[J]. 文化纵横, 2020(2): 34-42.

[15] 梅立润. 人工智能到底存在什么风险：一种类型学的划分[J]. 吉首大学学报(社会科学版), 2020, 41(2): 119-127.

[16] 郝英好. 人工智能安全风险分析与治理[J]. 中国电子科学研究院学报, 2020, 15(6): 501-505.

[17] Angwin J, et al. Machine Bias[EB/OL]. Propublica. 2016-5-23. [2020-06-11]. www.propublica.org/article/machinebias-risk-assessments-in-criminal-sentencing.

[18] Dockterman E. Robot Kills Man at Volkswagen Plant[N]. Time. 2015-07-01. [2020-06-13].

[19] O’Neill K. Robots learn to disobey humans! Watch machine say ‘no’ to voice commands[EB/OL].Mirror. 2015-11-26. [2020-06-13]. https://www.mirror.co.uk/news/weird-news/ robots-learn-disobey-humans-watch-6905437.

[20] Holden D. Federal Court in Chicago Orders U.K. Resident Navinder Singh Sarao to Pay More than $38 Million in Monetary Sanctions for Price Manipulation and Spoofing[EB/OL]. Commodity Futures Trading Commission. 2016-11-7. [2020-06-13]. https://www.cftc.gov/PressRoom/PressReleases/pr 7486-16.

[21] Fier J. The threat of AI-powered cyberattacks looms large[EB/OL]. AI Business. 2019-09-25. [2020-06-20]. https: // aibusiness.com/the-threat-of-ai-powered-cyberattacks- looms-large/.

[22] Zeifman I. Bot Traffic Report 2016[R]. 2017:1-50.

[23] Dickson B. The security threats of neural networks and deep learning algorithms[C]. 2018: 2-34.

[24] Kanter J. Why This Facebook Co-founder Says It’s Time to Break Up the Company[EB/OL]. 2019-5-9. [2020-06-20]. https://www.inc.com/business-insider/facebook-mark-zuckerberg-chris-hughes-op-ed-new-york-times.html.

[25] Lewis P, Hilder P. Cambridge Analytica’s blueprint for Trump victory[EB/OL]. 2018-3-23. [2020-06-20]. https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory.

[26] Berger J M, Morgan J. The ISIS Twitter Census[R]. 2015:5-34.

[27] Voice Fraud Climbs 350%[EB/OL]. Security Magazine. 2018-09-19. [2020-06-20]. https://www.securitymagazine.com/articles/ 89432-voice-fraud-climbs-350.

[28] PAX. Don’t be evil? A survey of the tech sector’s stance on lethal autonomous weapons[R]. 2019: 2-5.

[29] Pangburn D J. You’ve been warned: Full body deepfakes are the next step in AI-based human mimicry[EB/OL]. 2019-09-21. [2020-06-18]. https://www.fastcompany.com/90407145/youve-been-warned-full-body-deepfakes-are-the-next-step-in-ai-based-human-mimicry.

[30] Bickerton P. LabGenius: AI-driven synthetic biology[EB/OL]. 2018-06-11. [2020-06-18]. http://www.earlham.ac.uk/articles/labgenius-ai- driven-synthetic-biology.

[31] The National Academies of Sciences, Engineering, and Medicine. If Misused, Synthetic Biology Could Expand the Possibility of Creating New Weapons; DOD Should Continue to Monitor Advances in the Field, New Report Says[EB/OL]. 2018-06-19. [2020-06-18]. http://www8.nationalacademies.org/onpinews/ newsitem.aspx?RecordID=24890.

[32] Wiener N. Some Moral and Technical Consequences of Automation[J]. Science, 1960, 131(3410): 1355.

[33] Vinge V. The Coming Technological Singularity[EB/OL]. 1993. [2020-06-18]. http://www.frc.ri.cmu.edu/~hpm/book98/com.ch1/ vinge.singularity.html.

[34] Yudkowsky E. Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures[EB/OL]. 2001. [2020-06-18]. https://intelligence.org/files/CFAI.pdf.

[35] 刘益东. 试论科学技术知识增长的失控(上)/(下)[J]. 自然辩证法研究, 2002(4/5): 39-42.

[36] 刘益东. 粗放式创新向可持续创新的战略转型研究——科技重大风险研究 21 年[J]. 智库理论与实践, 2019(4): 75-78.

[37] Bostrom N. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards[J]. Journal of Evolution and Technology, 2002(9).

[38] Turchin A, Denkenberger D. Classification of global catastrophic risks connected with artificial intelligence[J]. AI & Soc. 2020(35): 147-163.

[39] Kwiatkowski R, Lipson H. Task-agnostic self-modeling machines[J]. Science Robotics, 2019, 4(26): 1-2.

[40] Vetere G. Memory formation in the absence of experience[J]. Nature Neuroscience, 2019(22): 933-940.

[41] Brain waves detected in mini-brains grown in a dish [EB/OL]. Science Daily, 2019-08-29. [2020-06-13]. https://www. sciencedaily.com/releases/2019/08/190829150824.htm.

[42] 刘益东. 致毁知识与科技伦理失灵:科技危机及其引发的智业革命[J]. 山东科技大学学报(社会科学版), 2018(12): 5-7.

[43] Autonomous weapons that kill must be banned, insists UN chief[N]. UN News. 2019-03-25. [2020-06-13].

[44] CCW. Draft Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems[EB/OL]. 2019. [2020-06-13]. https://www.unog.ch/80256EDD006B8954/ (httpAssets)/ 5497DF9B01E5D9CFC125845E00308E44/$file/CCW_GGE.1_2019_CRP.1_Rev2.pdf.

[45] Franco J. DoD Has 600 AI Projects–Needs Out of the Dog House[EB/OL]. 2018-04-23. [2020-06-13]. https://www.meritalk.com/articles/dod-has-600-ai-projects-needs-out-of-the-dog- house/.

[46] Sandoval G. Google Cloud’s new AI chief is on a task force for AI military uses and believes we could monitor ‘pretty much the whole world’ with drones[EB/OL] . 2018-09-12. [2020-06-13]. https://www.insider.com/google-cloud-new-ai-chief-history-military-security-task-force-2018-9.

[47] AFP. Amazon, Microsoft, May be Putting World at Risk of Killer AI, Says Report. Security Week[EB/OL]. 2019-08-21. [2020-06-13]. https://www.securityweek.com/amazon-microsoft-may-be-putting-world-risk-killer-ai-says-report.

[48] Malhotra B. Artificial intelligence for open source risk management[EB/OL]. 2017-10-25. [2020-06-20]. https://www. synopsys.com/blogs/software-security/artificial-intelligence-open-source-risk-management/.

[49] 刘益东. 挑战与机遇：人类面临的四大困境与最大危机及其引发的科技革命[J]. 科技创新导报, 2016(35): 221-230.

[50] 孔明安. 现代工程、责任伦理与实践智慧的向度[J]. 工程研究-跨学科视野中的工程, 2006(4): 222.

[51] 段伟文. 数据智能的算法权力及其边界校勘[J]. 探索与争鸣, 2018(10): 92.

[52] 陈小平. 人工智能伦理体系:基础架构与关键问题[J]. 智能系统学报, 2019, 14(4): 605-610.

Artificial Intelligence Risk Research: A Field to be Explored Urgently

Wang Yanyu

(*Institute for the History of Natural Science*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: With the rapid development of artificial intelligence, associated problems have gradually received increasing academic attention. This paper outlines the types of risks created by current artificial intelligence and analyzes the possible forms of strong artificial intelligence risks that can affect humans in the future. This paper also points out that the current risk governance model is flawed and fails to address the risks associated with the future growth of artificial intelligence. Thus, the academic community should actively promote the formation of a professional field of artificial intelligence risk and governance research, which would include related research criteria, research objects, and element composition.

Key Words: artificial intelligence; huge risks of science and technology; risk governance of AI; machine risk studies; ruin-causing knowledge