人工智能系统安全综述*

陈磊[1,2]　李雅静[3]

(1. 国防科技大学,长沙 410073;
2. 中国工程院战略咨询中心,北京 100088;
3. 麒麟软件有限公司,北京 100190)

摘要:人工智能的迅速发展使得人们越来越关注人工智能技术对社会的潜在影响。对人工智能安全技术的基本组成,及内生安全、衍生内生安全进行了研究。从技术上,分析了机器学习由于奖励函数、数据变化以及算法探索等方面的安全性挑战;从应用上,提出了伦理和法制等方面值得研究和解决的问题;从国家安全角度,提出了技术的自主可控性;最后,展望了人工智能安全未来的研究方向。

关键词:人工智能;安全;机器学习;深度学习

中图分类号:TP18　　**文献标识码**:A

引用格式:陈磊, 李雅静. 人工智能系统安全综述[J]. 信息通信技术与政策, 2021,47(8):56-63.

doi: 10.12267/j.issn.2096-5931.2021.08.009

0　引言

人工智能技术正在迅速普及,甚至是未来的主导技术。随着大数据时代的到来和计算机性能的提升,以深度学习为代表的机器学习算法在计算机视觉、自然语言处理、自动驾驶汽车[1-2]等领域的应用取得了极大的成功,预示着人工智能超过人类自然习得智能的可能性。人类文明是人类智能的产物,而通过人工智能获得更高级的智能将影响人类历史。

人工智能技术的发展方向和时间进度无法准确预测,因而人工智能技术的进步也是一把"双刃剑"。一方面,人工智能作为一种通用使能技术,在医学、科学和交通等领域[2-4]产生颠覆性的作用,也为保障国家网络空间安全、提升人类经济社会风险防控能力等方面提供了新手段和新途径;另一方面,人工智能在技术转化和应用过程中,引入的隐私、安全、经济和军事等问题[5-8],引起人们对人工智能的长期影响而担忧,尤其是对网络与信息系统安全、社会生产系统、社会就业、法律伦理等领域的冲击,并对国家政治、国防、经济和社会安全带来诸多风险和挑战[9-10]。因而,世界主要国家都将人工智能安全作为人工智能技术研究和产业化应用的重要组成部分。

本文概要介绍了人工智能、人工智能安全以及分类,并对人工智能内生安全和衍生内生安全两个方面进行了阐述,最后对人工智能安全技术的研究进展、面临的挑战以及应用前景进行了探讨和展望。

1　人工智能技术及其安全问题

人工智能是指计算机程序或机器具有思考和学习的能力,是试图使计算机"智能化"的研究领域,是计算机科学行业的顶尖技术之一。

1.1　人工智能技术的发展

"人工智能"最初是科学家们用来讨论机器模拟人类智能时提出的。1936 年,英国数学家图灵就曾在他的论文"理想计算机"提出图灵模型以及 1950 年在他的论文"计算机可以思考吗"提出机器可以思考的论述(图灵实验),为人工智能的诞生奠定了基础。

* 基金项目:中国工程院 2019 年重大咨询研究项目(No. 2019-ZD-1)资助

1956 年,美国达特茅斯大学举办了一场“侃谈会”,人工智能这个词语第一次被搬上台面,从而创立了人工智能这一研究方向和学科。1956 年,美国的两个心理学家纽俄尔和西蒙也成功地在定理证明上取得突破,于是开启了通过计算机程序模拟人类思维的道路。在 1967—1970 年代初期,科学家们想对人工智能进入更深层次的探索时,发现人工智能的研究遇到许多当代技术与理论无法解决的问题。因为当时计算机的处理速度和内存容量都已经不足以实现更智能化的发展,也没有人知道人工智能究竟能够智能化到何种程度。因此,各界科研委员会开始停止对人工智能研究的资助,人工智能技术的发展也就此跌入低谷。1980—1987 年,随着理论研究和计算机软、硬件的迅速发展,美国、英国对人工智能开始重新研究并投入了大量资金,在 1984 年启动了 Cyc 项目,目的就是让人工智能可以应用到类似人类大脑思考以及推理的工作中。随后许多研究人工智能的技术人员们开发了各种 AI 实用系统尝试商业化并投入到市场,人工智能又激起了一股浪潮。2016 年,AlphaGo 战胜了世界围棋高手李世石,人工智能成为当年热度最高的科技话题。图 1 为人工智能技术发展历史。

人工智能分为强人工智能和弱人工智能两类。强人工智能的定义为:需要具备自我意识,在遇到问题的时候需要能向人类一样进行对问题的决策。而要实现这种情况的难度很大,所以在强人工智能的研究方面始终没有很大的进展。弱人工智能并没有思维意识,只能按照程序员预编写好的程序进行相应的工作,与强人工智能相比较而言,其在这 60 多年得到了快速的发展,现如今人工智能的发展也是主要围绕着弱人工智能去进行。

1.2 人工智能安全

近十几年来,由于计算能力的高速发展、可获取数据量的急速增长以及人工智能技术的自身优势,尤其是深度学习技术的突破性进展,使得人工智能系统获得愈来愈广泛的应用并取得令人瞩目的成就。与此同时,人工智能系统在应用过程中也暴露出大量安全问题,引发了人们对人工智能安全的高度关注。

从一般的人造系统或工程系统的角度考虑,人工智能系统作为一种特殊的工程系统同样具有其同类系统所可能具有的安全性问题。例如,作为一种计算机软件或硬件系统,它同样具有计算机软、硬件系统所可能具有的安全性问题;作为机器人、无人驾驶系统,它可能具有结构、材料性质、环境适应性等安全性问题。但是人工智能系统或“智能体”与其他系统的本质区别在于其具有“智能”,能够思维,可以进行“学习”而习得知识,可以做出自主性决策,进而可以执行自主性行动。区别于与其同类工程系统的普适性安全问题,这里讨论的人工智能安全问题,集中于人工智能系统由于具有学习能力、自主决策、自主行动这些智能行为而产生的安全性问题,既包括人工智能系统本身由于设计、制造、使用环境或受到攻击所产生的安全问题,也包括由于人工智能技术的应用而衍生出来的对其他

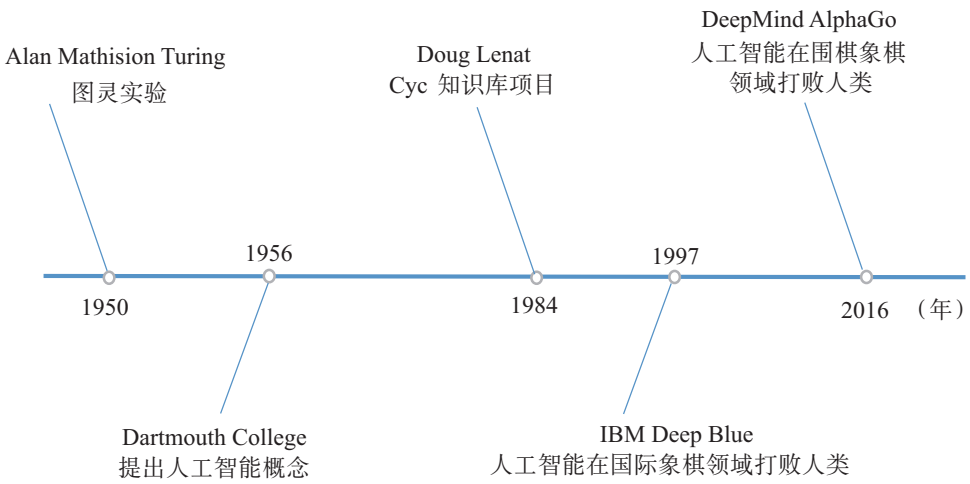

图 1 人工智能技术发展历史

系统的安全影响。

1.3 人工智能安全的分类

根据产生安全问题的原因不同,可以把人工智能系统由于外部影响或受到攻击而产生的安全称为“外生安全”或者“衍生安全”,人工智能系统由于其本身技术特点、缺陷、脆弱性而产生的安全称为“内生安全”。

人工智能系统的内生安全问题既包含智能体本身由于其技术特点、缺陷、脆弱性等可能造成的智能体本身的安全问题(确切地说,该智能体作为具有特定功能的系统而不能正确地行使其特定的功能),也包括该智能体工作环境或由于受到其他系统的攻击而不能正确地行使其特定的功能。内生安全问题是局限于智能体本身的安全性问题,而无论是由于其本身的特性产生的问题还是工作环境或遭受外部攻击所产生的问题。人工智能系统的衍生安全问题是由于智能体的应用对于其他系统所可能造成的安全性问题,既包括由于智能体出现安全问题(即不能行使其特定功能)而产生的(衍生出来的)对其他系统的安全性所造成的影响,也包含该智能体特定功能的成功行使而可能对其他系统的安全性所造成的影响。

此种划分可以区分对抗环境中系统的安全问题产生的原因,尤其是识别外部攻击的因素,有益于研究对抗环境中人工智能的安全性问题。而大量外部攻击均利用了人工智能系统本身的技术特点、缺陷、脆弱性,因此即使是由于外部影响或在受到攻击而产生的安全问题也是内因和外因共同作用的结果。

本文主要讨论人工智能本身的安全问题和由于人工智能的应用而可能造成的外在的安全影响,所以在在分类中仅包含采用“人工智能系统本身的安全性问题”(即内生安全问题)和由于人工智能技术的应用而衍生出来的安全为题(即衍生安全问题)。在分析内生安全问题的同时,也将把人工智能系统本身的缺陷和脆弱性与可能利用这种缺陷和脆弱性而进行的外部攻击的因素结合进行论述。

2 人工智能安全的国内外研究现状

下面将从内生安全和衍生安全的两个方面综述国内外人工智能安全的研究进展。

2.1 人工智能内生安全

人工智能内生安全分类参见图2。

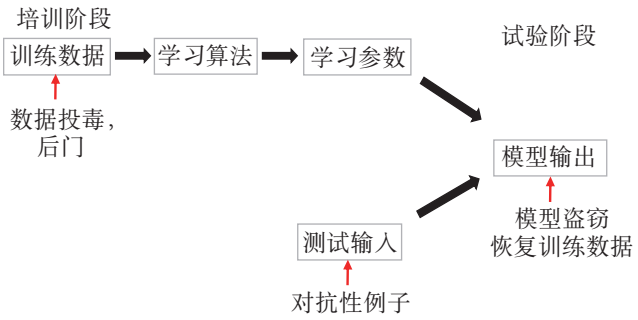

图2 人工智能内生安全分类

2.1.1 框架/组件引发的内生安全

在框架/组件方面,难以保证框架和组件实现的正确性和透明性是人工智能的内生安全问题。当前,国际社会已经推出了大量的开源人工智能框架(如TensorFlow、Caffe)和组件并得到了广泛使用。然而,由于这些框架和组件未经充分安全评测,可能存在漏洞甚至后门等风险。一旦基于不安全框架构造的人工智能系统被应用于关乎国计民生的重要领域,这种因为“基础环境不可靠”而带来的潜在风险就更加值得关注。

2.1.2 数据安全

在数据方面,用于训练的数据不完全可能使得学习算法难以发现准确反映环境和应用需求的正确模型,是造成系统安全问题的重要因素。训练中使用的数据规模、质量(准确、真实、全面)、数据分布特征等均会影响训练结果的正确性,从而影响AIA决策和行为能否正确地实现设计目的;而训练模型应对数据分布特征变化的鲁棒性也是影响系统安全性的重要因素。

训练数据对安全性的影响使得数据安全成为对抗环境中攻击和防护技术的研究热点。在对抗环境中,攻击者可以对数据实施闪避攻击和药饵攻击等。闪避攻击通过在正常的训练样本中加入人工不易觉察的少量样本数据,改变样本数据集,形成对学习系统的攻击。闪避攻击可通过对抗样本生成、利用传递性进行黑盒攻击等方式实现,主要应用在推理阶段。药饵攻击是在训练数据集中加入虚假数据(即药饵),使得学习系统生成错误的模型的人达到攻击的目的。药饵攻

击通常应用在训练阶段,破坏系统生成的模型。

2.1.3 算法安全

在算法方面,学习算法的可解释性、复杂性问题使得算法成为影响人工智能系统安全性的脆弱环节。算法的可解释性是影响用户对算法安全性的关键因素,尤其是在如智能诊断和医疗、财政金融等决策的正确性可能引发重大安全后果的领域。算法的复杂性意味着对于大数据量的学习会付出巨大的代价(如计算资源和执行时间),而动态变化的环境可能要求算法经常和反复执行基于数据学习新的动态模型,这种巨大的代价可以使得机器学习系统难以适应大规模数据以及随时间和环境动态变化的需求,产生错误的学习结果。

2.1.4 模型安全

在模型方面,模型的不透明性、脆弱性、保密性均使得机器学习模型成为影响人工智能系统安全性的脆弱环节。不正确的目标函数会使得机器学习系统获得错误的模型,造成 AIA 不能执行设计者预想的特定功能;模型的不透明性使得学习模型做出错误决策而影响行为的正确性和安全性;模型对动态变换的环境的适应性(健壮性)也是模型影响系统安全的重要因素;训练出的模型在对外服务的过程中可能被以对抗手段获取而威胁系统安全。这些均使得机器学习的模型成为影响人工智能系统安全的脆弱环节。

在对抗环境中,攻击者还可以对模型实施模型窃取攻击和后门攻击等。

(1)模型窃取攻击是指攻击者可以通过窃取存储的神经网络模型的攻击方法。在以机器学习作为一种服务而提供给用户的系统中,攻击者可以在不知道模型参数或训练数据,而通过访问以机器学习提供服务的系统黑盒子,对模型进行窃取,被窃取的模型进而可以为进一步的攻击提供模型或功能参数,从而形成安全隐患。

(2)后门攻击是指在模型中植入后门这样一种重要的模型攻击方法。由于模型本身的不透明性、不可解释性等特点,所植入的后门难以被发现,而攻击者却可伺机启动后门的功能,从而形成安全隐患。

2.1.5 协同引发的内生安全

在协同方面,多无人系统是一个复杂的综合体,其安全问题是多维度且耦合的。例如,作战单元或后勤保障单元遭入侵,间谍节点潜伏集群内加入作战任务,作战关键时刻倒戈或故障可导致集群任务失败;人类指挥与自主无人系统的决策产生矛盾时,很难改变无人系统自身决策,从而导致沟通困难甚至贻误战机;在非全局立场上,自主无人系统之间会产生机与机的矛盾;对于多无人系统协同的数据传输和通信协议过程,未进行加密处理或加密强度较低时,可能会发生两种类型的安全威胁。一种是探索性攻击(Exploratory Attack),即攻击者不会干扰通信,但会尝试提取可用信息。在这种情况下,传输的敏感信息可能泄露。另一种是诱发性攻击(Causative Attack)。此时,攻击者可能会尝试拦截、修改、删除甚至替换数据包。

2.1.6 运行环境引发的内生安全

在系统和运行环境方面,智能系统本身的结构和应用环境是影响系统安全性的重要因素。实现机器学习或智能决策和控制的计算系统中在软硬件结构不同的层次上存在安全隐患或脆弱环节,使得攻击者可以利用这些隐患或脆弱环节对人工智能系统进行攻击;在云计算体系结构中存在大量的用户,采用机器学习的方法进行模型训练和智能推理,这些用户共享分布式计算系统中的软硬件资源,如服务器、复杂的软件栈、存储资源、计算框架等,这些资源易于受到攻击而对模型的训练和智能推理产生安全性威胁;在分布式计算环境中的联邦学习框架中,多个用户在分布式系统中进行合作训练和推理,易于受到恶意参与者的模型攻击造成安全隐患。

2.2 人工智能衍生安全

人工智能系统的衍生安全性问题涵盖技术和工程系统、经济、社会、金融、医疗健康、国防和国家安全各个领域,是国家和社会在人工智能技术发展和应用中必须严肃面对的重大问题。

2.2.1 人工智能系统存在安全隐患、引发安全事故

人工智能系统因算法不成熟或训练阶段数据不完备等原因,导致其存在缺陷,这种缺陷即便经过权威的安全评测也往往难以全部暴露出来,人工智能系统在投入实际使用时,就容易因自身失误而引发人身安全问题。当前,具有移动能力和破坏能力的智能体,可引发的安全隐患尤为突出。2018 年 3 月,由 Uber 运营的自动驾驶汽车在美国亚利桑那州坦佩市(Tempe)撞倒了一名女性并致其死亡,经调查分析认为,这是因为自动驾驶的汽车“看到”了这名女性但没有刹车,同时自

动驾驶系统也没有生成故障预警信息。仅根据 AI 事故数据库 AIID 最新数据(截止到 2021 年 1 月 5 日),AI 智能体已经引发近百种至上千个重要事故,包括自动驾驶汽车致人死伤、工厂机器人致人死伤、医疗事故致人死伤、伪造政治领袖演讲、种族歧视言论、不健康内容等安全危害事件;而利用人工智能技术造成的系统破坏、人身杀伤、隐私泄露、虚假身份识别风险、社会影响重大的舆论等事故也多有发生。

2.2.2 人工智能给人类社会治理带来巨大冲击

人工智能的发展完善和应用领域的不断拓展,可能会对人类整体的文明产生巨大冲击。当前,人工智能已经广泛应用于医疗设备与医疗诊断、工业自动控制、交通出行、金融服务等很多领域,正在帮助人类进行一些原来只属于人类的工作。在经济冲击方面,人工智能可以帮助人类感知、分析、理解和预测,用更经济、便捷的方法执行任务,导致不再需要另行聘请有经验的专家,可以显著减少劳务开支和各种相关费用;在社会冲击方面,随着技术的发展,机器已经逐步替代人类从事部分繁琐重复的工作或体力劳动,机器在给人们带来福利的同时,也让人们越来越担忧自己的工作会被机器所替代,从而引发人们对于失业的担忧。由此产生的大量失业者也将成为社会的不安定因素,可引发社会劳动和职业结构的剧烈变化;在人类思维冲击方面,随着人工智能的发展,机器变得越来越“聪明”,人类越来越依靠机器,这在某种程度上会导致部分人群思维变得懒惰、认知能力下降。

2.2.3 智能体一旦失控将危及人类安全

智能体一旦同时具有行为能力以及破坏力、不可解释的决策能力、可进化成自主系统的进化能力这三个失控要素,不排除其脱离人类控制和危及人类安全的可能。智能体失控造成的衍生安全问题,无疑是人类在发展人工智能时最关心的一个重要问题。已有一些学者开始思考“奇点”何时会到来,即人工智能的自我提升可能将会超过人类思想,导致智慧爆炸。尽管目前尚未出现真正意义上的人工智能失控事件,但新技术的发展很难保证在将来超级人工智能不会出现,届时将如何保护人类,实现超级人工智能和人类的和谐共存,这是人工智能在未来发展道路上需要解决的主要问题之一。

对于人工智能的安全性问题,国际范围内包括学术界、国家和国际组织,以及著名的政治、经济、社会活动家,从科学与技术、社会、政治、经济国家关系和国家安全等方面进行了大量的讨论。我国非常重视人工智能的安全问题,确定了在大力发展人工智能的同时,必须高度重视可能带来的安全风险和挑战,加强前瞻预防与约束引导,最大限度降低风险,确保人工智能安全、可靠、可控发展,为我国人工智能,尤其是人工智能安全工作,确定根本的指导原则。

3 挑战与展望

目前,人工智能技术正处在高速发展时期,人工智能系统的安全面临诸多挑战。从技术的角度,机器学习在奖励函数、数据变化以及算法探索等方面均存在安全性风险,随着人工智能系统越来越自治,有可能产生无法预计、有害的行为以及安全性事故。未来,机器学习的开发应考虑安全性。从应用的角度,人工智能在技术转化和应用上对网络与信息系统安全、工程和技术系统、社会经济和就业结构、法律伦理等领域产生巨大的冲击,带来诸多风险和挑战。从国家安全的角度,由于我国人工智能研究基础薄弱,技术的自主可控性存在挑战。

3.1 奖励函数的安全性

机器学习中不合理的奖励函数将人工智能系统的运行偏离设计者的意图。达到同样的奖励函数,可以有多种解决方案,而人工智能系统可自主选择有效的、但不符合设计意图的方案执行,即人工智能系统与其设计者存在博弈。奖励函数的不合理性可以导致连贯的、意想不到的行为,并有可能对现实世界的系统产生有害的影响。例如,遗传算法[11]通常可以输出意想不到但形式上正确的问题解决方案。基于 Counterfactual Learning[12-13]方法和 Contextual Bandits[15]方法的机器学习系统[12,14]的反馈回路被研究存在上述问题。上述不同领域案例的激增表明,奖励函数的安全性已成为了普遍问题。因此,提高奖励函数的安全性是一个重要的技术难题,提高奖励函数的安全性是本领域重要的研究方向。

3.2 数据变化的安全性

人类发现自己在应对没有经验的事情上会出错,往往在意识到自己知识的缺陷时而避免风险。人工智能系统存在同样的问题,例如基于干净的语音,训练得

到语音识别系统,在处理有噪声的语音时表现很差,但人工智能系统并不能意识到出错的风险。这种错误常具有危害性,例如给出可信度很高但错误的医疗诊断。

人工智能系统没有得到正确数据的训练可能产生严重的错误和有害的行为。此外,在真实数据发生变化,与训练数据不同时,则会产生意识不到的、有害的风险。人工智能的许多领域将遇到上述问题,包括变化检测和异常检测[15-17]、假设检验和迁移学习[18-20]等领域。因此,探索更好的方法来检测数据变化引入的错误,并最终保证错误发生的概率在统计学上足够小,对于构建安全和可预测的系统至关重要,是人工智能安全重要的研究方向。

3.3 算法探索的安全性

在环境信息不足时,人工智能系统往往采取一些行动来了解所处的环境,这种行动称为探索。由于不清楚行动的后果,探索可能是危险的,例如无人机可能会撞到地面、工业控制系统可能会造成严重问题等。

常见的探索策略,如 Epsilon Greedy[21]或R-max[22]通过随机选择一个行动或所有未探索的行动进行探索,因此不会试图避免这些危险情况。更复杂的勘探策略[27]采用连贯的在长时间尺度上的行动,与随机行动相比,可预测更大的潜在危险。采用上述方法可预测危险的行为,并在探索时避免这些行为。

目前,人工智能系统相对简单,设计者可分析出所有出错的情况,通过简单地硬编码避免灾难性行为。随着人工智能系统越来越自治,在更复杂的领域中行动,预测每一个可能的灾难性失败可能变得越来越困难,探索的安全性问题在文献[23]和文献[24]中进行了讨论。

因此,提出一些更具普遍意义、原则性的算法设计规则在探索中避免危险是非常必要的,也是人工智能算法自动探索安全性的重要研究方向。

3.4 自主可控性

我国在人工智能软硬件基础理论、核心关键技术上积累薄弱,缺乏重大原创科技成果,核心算法、芯片及基础元器件等核心环节受制于人。目前,在深度学习开源平台领域,已经形成了谷歌的 TensorFlow 和脸书的 PyTorch 两家独大的格局。谷歌自研了 TPU 芯片,与其深度学习框架 TensorFlow 进行深度融合,以“深度学习框架+人工智能芯片”的模式,构建智能时代新的“Wintel”联盟,试图掌控智能时代新的话语权。华为的解决方案是通过自主设计 AI 计算框架 MindSpore,自主研发高性能 AI 芯片,自定义算子开发工具等策略来实现对 AI 体系的全方位掌控。

3.5 对我国的挑战

在核心算法方面,目前还没有提出像 CNN、GAN 等国际上流行的原创算法。在芯片方面,我国虽然也出现了寒武纪等芯片,但在影响力方面仍然与国外顶尖厂商有相当差距;芯片制造、封装等能力在国际上的竞争力更加薄弱。一方面,由于这些开源框架、组件、算法、芯片都来自于美国等发达国家,我国人工智能的安全技术要基于这些开展研究和部署;另一方面,由于没有经过我国严格的测试管理和安全认证,国外这些开源框架、组件、算法可能存在的漏洞和后门等安全隐患一旦被攻击者恶意利用,可危及人工智能产品和应用的安全性而导致重大财产损失和恶劣社会影响。

3.6 法律和伦理问题

人工智能技术在社会领域的渗透逐渐深入,给当前社会的法律法规和基本的公共管理秩序带来了新的危机。新的社会现象的出现,超出了原有的法律法规在设计时的理念边界。人工智能的发展对现有法律体系的冲击使得现行法律监管无法及时有效回应智能技术带来的挑战。人工智能技术和伦理安全的法律和法规的缺失,可能导致人工智能技术的发展和产品的开发应用失控失序,危及公民权利、社会福祉、公共秩序、国家安全等严重态势。在人工智能伦理安全方面,近期更多的是需要应对人工智能在算法歧视和决策偏见、人工智能技术滥用等风险,远期则需要防范其可能产生超级智能而带来的不可控性以及对人类主体性的冲击。在上述原则基础上,急需构建面向人民群众、面向现代化、面向世界的人工智能伦理规范和法律法规。

4 结束语

本文介绍了人工智能技术的概念、技术构成、发展历程,分析了该项技术的发展现状和存在的技术挑战,特别是安全性挑战。在技术上,分析了机器学习由于奖励函数、数据变化以及算法探索等方面设计不当而产生的非预期的有害事故,指出防止此类事故的重要性,即人们对人工智能系统的安全性如果失去信心,更大事故的风险将更加难以估量。人工智能系统自治日

益增长的趋势表明,需要制定普遍的规则来防止事故产生。在应用上,提出了伦理和法制等方面值得研究和解决的问题。并从国家安全角度,指出加强技术的自主可控性。随着计算机算力的提升,以及机器学习算法的进步,加速了人工智能技术的快速发展,已在一些领域超过人类自然习得智能,并在更多领域成为一种趋势,最终将变革人类社会发展的方向。

参考文献

[1] KUANG Lichun, LIU He, REN Yili, et al. Application and development trend of artificial intelligence in petroleum exploration and development[J]. Petroleum Exploration and Development, 2021,48(1):1-14.

[2] Li B K, Miao Q, Li M, et al. An investigation on machined surface quality and tool wear during creep feed grinding of powder metallurgy nickel-based superalloy FGH96 with alumina abrasive wheels[J]. Advances in Manufacturing, 2020, 8(2):160-176.

[3] 李旭东, 林晓珠, 房炜桓, 等. MRI 图像纹理分析在胰腺神经内分泌肿瘤病理分级中的应用研究[J]. 诊断学理论与实践, 2017,16(6):601-606.

[4] Gil Y, Greaves M, Hendler J, et al. Amplify scientific discovery with artificial intelligence[J]. Science, 346(6206):171-172.

[5] 刘海, 李兴华, 雒彬, 等. 基于区块链的分布式 K 匿名位置隐私保护方案[J]. 计算机学报, 2019,42(5):942-960

[6] 张丰菊, 李玉, 熊瑛. 强化有晶状体眼后房型人工晶状体植入术矫正近视眼和散光的安全性[J]. 中华眼科杂志, 2021,57(2):86-89.

[7] Dong X, Mcintyre S H. The second machine age: work, progress, and prosperity in a time of brilliant technologies[J]. Quantitative Finance, 2014,14(11):380-383.

[8] Autonomous weapons: not - open letter from AI & robotics researchers[J]. Intelligence the Future of Computing, 2015.

[10] Northey M, Jewinski J. Superintelligence : paths, dangers, strategies[M], 2014.

[11] Flanagan M P, Shearer P M. Topography on the 410-km seismic velocity discontinuity near subduction zones from stacking of sS, sP, and pP precursors[J]. Journal of Geophysical Research: Solid Earth, 1998, 103(B9).

[12] Park J S, Kim M S. Design and implementation of an SNMP-Based traffic flooding attack detection system[C]// Asia-pacific Network Operations & Management Symposium. Springer, Berlin, Heidelberg, 2008.

[13] Zhang Y, Niu B. Design and implementation of a lightweight distributed job management system[J]. e-Science Technology & Application, 2019.

[14] Park J S, Kim M S. Design and implementation of an SNMP-Based traffic flooding attack detection system[C]// Asia-pacific Network Operations & Management Symposium. Springer, Berlin, Heidelberg, 2008.

[15] Ding K, Ding S, Morozov A, et al. On-line error detection and mitigation for time-series data of cyber-physical systems using deep learning based methods[C]// 2019 15th European Dependable Computing Conference (EDCC). IEEE, 2019.

[16] Ramchandran A . Unsupervisedanomaly detection for high dimensional data—an exploratory analysis[J], 2018:233-251.

[17] Lee C K, Cheon Y J, Hwang W Y. Studies on the GAN-based anomaly detection methods for the time series data[J]. IEEE Access, 2021(99):1-1.

[18] Zhang X, Shen C, Guo X, et al. ASFP (Artificial Intelligence based Scoring Function Platform): a web server for the development of customized scoring functions[J]. Journal of Cheminformatics, 2021, 13(1).

[19] Wang Y, Yang Y, Yao Y. Single shot multibox detector for ships detection in inland waterway[J]. Journal of Harbin Engineering University, 2019.

[20] Adams N . Dataset shift in machine learning[J]. Blackwell Publishing Ltd, 2010, 173(1):274-274.

[21] Tokic M, Palm G. Value-difference based exploration: adaptive control between epsilon-greedy and softmax[C]// Annual Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2011.

[22] Cicali E J, Weitzel K W, Elsey A R, et al. Challenges and lessons learned from clinical pharmacogenetic implementation of multiple gene – drug pairs across ambulatory care settings[J]. Genetics in Medicine, 2019.

[23] 阙天舒, 张纪腾. 把握人工智能时代下的国家安全治

理[J]. 新华月报, 2019(3):114-116.
[24] 买合木提江·阿不都热合曼. 人工智能在网络安全防御中的应用[J]. 计算机与网络, 2020, 46(7):56.

作者简介:

陈磊 国防科技大学博士研究生,中国工程院战略咨询中心助理研究员,研究方向为科技政策、人工智能安全等

李雅静 通信作者。麒麟软件有限公司,研究方向为科技政策、新媒体等

An overview of artificial intelligence system security

CHEN Lei[1,2], LI Yajing[3]

(1. National University of Defense Technology, Changsha 410073, China;
2. Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China;
3. KylinSoft Beijing 100190, China)

Abstract: The rapid development of artificial intelligence makes people pay more and more attention to the potential impact of artificial intelligence technology on society. In this paper, the basic components of artificial intelligence safety technology are introduced, and the research status at home and abroad is expounded from the research direction of endogenous safety and derivative endogenous safety. In terms of technology, the security challenges of machine learning due to reward function, data change and algorithm exploration are analyzed. In terms of application, the ethical and legal issues worth studying and solving are proposed. From the perspective of national security, the autonomous controllability of the technology is proposed. Finally, the future research direction of artificial intelligence security is prospected.

Keywords: artificial intelligence; safety; machine learning; deep learning

(收稿日期:2021-05-24)