

# 人工智能教育融合安全警示： 来自机器学习算法功能的原生风险分析

刘梦君, 蒋新宇, 石斯瑾, 江南, 吴笛

(湖北大学 教育学院, 湖北 武汉 430062)

**[摘要]** 安全与隐私是《2021 地平线报告》头等关注议题, 针对人工智能技术的核心基础——机器学习算法存在的功能安全漏洞问题, 文章首先从原理上介绍了机器学习算法面临的投毒攻击和对抗样本攻击功能风险; 随后对服务于教师、学生与教育管理三方面对象和不同教学与管理环节的六类典型人工智能教育应用: 智能教学平台、智能课堂教学行为分析、智能评测、智能搜题、智能学习助手、智能考务应用中, 机器学习算法融合过程中的功能风险进行了深度剖析, 结果表明, 算法功能攻击动摇了所有上述应用正常运行的根基; 最后, 在当前并无良好的技术解决方案的现实下, 研究初步探讨从运行与管理层面, 对人工智能教育应用面临的算法功能攻击予以防御的建议。

**[关键词]** 人工智能教育; 机器学习; 安全风险; 算法缺陷; 功能安全

**[中图分类号]** G434

**[文献标识码]** A

**[文章编号]** 1671-6973(2022)05-0089-13

## 一、引言

“人工智能+教育”作为新一代教育信息化的核心引擎, 吸引着政、企、学三界不遗余力地推动人工智能与教育教学的深度融合。学术和产业界产出了大量关于人工智能的教学理论<sup>[1]</sup>、模型<sup>[2]</sup>、方法<sup>[3]</sup>和功能系统<sup>[4]</sup>, 人工智能仿佛无所不能, 各类新奇的人工智能教育理念<sup>[5]</sup>和应用<sup>[6]</sup>纷纷涌现在师生、家长和社会眼前。智能教学、智能课堂教学行为分析、智能评测、智能搜题、智能教与学助手、智能考务等应用越来越多地被运用到诸多教与学环节中。一时的教育界, 人工智能风头无两, 迎来高光时刻。

人工智能技术的基石是机器学习算法, 科学家一直梦想着机器学习算法(如无特别说明, 后文

**[收稿日期]** 2021-10-09

**[基金项目]** 本文为国家自然科学基金“新高考综合素质评价数据安全治理机制研究: 区块链技术赋能视角”(项目编号: 72204077)、湖北省自然科学基金“基于区块链的新高考综合素质评价数据安全治理机制研究”(项目编号: 2021CFB470)、教育部人文社会科学青年项目“基于认知诊断的网络学习能力评测与个性化推荐机制研究”(项目编号: 19YJC880093)阶段性研究成果。

**[作者简介]** 刘梦君(1988—), 男, 湖北黄冈人, 博士, 湖北大学教育学院副教授, 主要研究方向为教育数据挖掘与安全治理、人工智能教育应用。

**[通讯作者]** 吴笛(1984—), 男, 湖北潜江人, 博士, 湖北大学教育学院副教授, 主要研究方向为学习资源推荐。

单独的“算法”皆为“机器学习算法”一词的简称)能够发展到帮助人工智能具备人类智慧的水平,然而,在进展到这一步之前,一些计算机科学领域的算法科学家们对算法自身的安全可靠性忧心忡忡<sup>[7,8]</sup>。为了推动技术的发展,一直以来,研究者们对算法的运行环境,往往设定成参与各方无条件遵循算法设计者意愿,配合算法运行。这种理想化的算法运行环境,在实际尤其是牵涉到重大利益的现实应用上,往往难以存在,导致应用运行过程危机四伏<sup>[9,10]</sup>。计算机学科领域专家已经对此问题开始了大量针对性的研究<sup>[11-13]</sup>,而教育界还鲜有研究关注到此问题并意识到问题的严峻性。

正所谓“根基不牢,地动山摇”。除去人工智能教育应用作为一种信息系统面临的传统信息安全攻击风险,如信息泄露、网络攻击等传统数据隐私安全风险<sup>[14]</sup>之外,它还面临着一些非传统的功能安全攻击风险<sup>[15-17]</sup>,而现有的信息安全技术不能有效应对这些风险。由于安全是一切技术应用发展的首要前提,特别是在教育这一拥有特殊用户群体的领域。因此,为了在满足日益增长的人工智能教育教学应用需求的同时,保证教育教学活动安全开展,亟需向广大教育业界人员解读机器学习教育应用所面临的功能安全风险这一非传统安全风险。与此同时,在基础算法理论取得突破之前,教育领域各界人士务必保持必要警惕性,并在人工智能教育应用中从应用运行管理方面采取各类法律、制度、管理和组织等防御性措施。

## 二、人工智能系统教育应用安全风险研究现状

中国信通院发布的《人工智能安全框架》中,将人工智能安全框架划分为目标、管理、能力和技术四个维度,研究者们从应用目标与定位背离公序良俗带来伦理风险、应用运行管理不当带来管理风险、应用安全能力的缺失带来隐私风险和应用核心技术本身的缺陷带来功能风险四个角度展开了相关研究。

应用的目标与定位是人工智能系统教育应用实现的前提,这个前提决定着人工智能教育应用的整体走向。为避免应用目标与定位背离公序良俗对社会带来根本性的伦理风险,国内研究者依据中国国情,针对这一风险进行了许多研究。张安毅<sup>[17]</sup>立足于人工智能应用责任制度,建议在相应的责任法给出人工智能致损应用责任规定,以此对应用设计者的行为做出规范;杜静<sup>[18]</sup>基于人工智能在教育领域的应用现状,归纳出八大教育智能教育伦理原则,力求为人工智能伦理问题的解决提供规制;钱小龙<sup>[19]</sup>针对人工智能带来的教育伦理风险,提出了教育人工智能必须遵循的伦理原则和未来发展的策略建议。

在应用运行管理上,《人工智能安全标准化白皮书》明确指出人工智能应用的运行及其管理是与用户最直接相关的,在此阶段产生的所有安全管理风险都是与用户和开发者直接相关的,其重要性可见一斑。因此,教育界的研究者们纷纷展开研究,为人工智能应用运行管理建言献策。卢迪<sup>[20]</sup>从全球视野出发,着眼于宏观国家层面、中观校园层面与微观个性化层面问题的解决,为实现人工智能教育“善治”构建国际适用的治理框架;胡元聪<sup>[21]</sup>从开发者角度出发,明确了人工智能企业研发生产、缺陷应用管理等方面社会责任,通过创新人工智能应用管理制度来降低相应风险。

在应用安全能力上,杨蓉<sup>[22]</sup>、季卫东<sup>[23]</sup>等人的研究均表明人工智能应用安全能力的缺失会使系统内隐私数据暴露于各类风险之下,产生大量诸如系统安全、网络安全和数据安全等传统安全攻击风险。对此类问题,研究者们已从系统安全防护<sup>[24]</sup>、网络安全防护<sup>[25]</sup>以及数据安全防护<sup>[26]</sup>等角度展开研究并取得一定进展。特别是在教育人工智能应用所产生的涉及教育者、学习者、管理者等

多个层面、多种模态的教育大数据<sup>[27]</sup>安全方面,研究者们从传统信息安全技术<sup>[28]</sup>与新兴的区块链技术<sup>[29,30]</sup>等多个角度提出了相应的隐私防护措施。

总的来看,人工智能教育应用在伦理、管理、隐私三方面的传统安全风险,在政企学三界的努力下,已经从各方面给出了大量解决方法。然而,对于投毒攻击<sup>[31]</sup>、对抗样本攻击<sup>[32]</sup>等多种非传统安全攻击导致的功能安全问题,闫怀志<sup>[33]</sup>曾指出:“人工智能功能安全问题与传统的网络安全强调的保密性、完整性、可用性等信息安全问题,存在本质不同。”李欣姣<sup>[31]</sup>、魏立斐<sup>[12]</sup>、何英哲<sup>[11]</sup>以及陈宇飞<sup>[32]</sup>等人对机器学习算法的研究也表明,机器学习算法在训练和预测阶段遭遇到非传统安全攻击行为,算法无法实现设计功能,具备较弱鲁棒性。

### 三、机器学习及其技术风险

#### (一)机器学习技术简介

机器学习是一门通过学习“经验”来改善计算机系统性能的学科。经验常为数据,学习“经验”则是通过学习产生可以使系统性能变好的数学模型,也即算法。从机器学习定义来看,数据和算法是两大关键,机器学习应用程序依据这两大关键大致可分为数据采集、数据预处理、模型训练、模型输出四个步骤。其中数据采集主要是收集目标数据;而数据预处理是对采集数据进行清洗,然后划分为训练数据集和测试数据集;模型训练则是使用训练数据,对算法的参数进行调整,以适应目标数据集的独有特性;模型输出则是使用调好参数的算法对测试数据进行预测。

#### (二)机器学习算法功能风险

机器学习工作关键在模型训练和模型预测阶段。机器学习算法面临功能风险也出现在这两个阶段,如图1所示。攻击者多采用投毒攻击和对抗攻击来对教育应用功能进行干扰,破坏模型的完整性和可用性,使其无法正常运行。

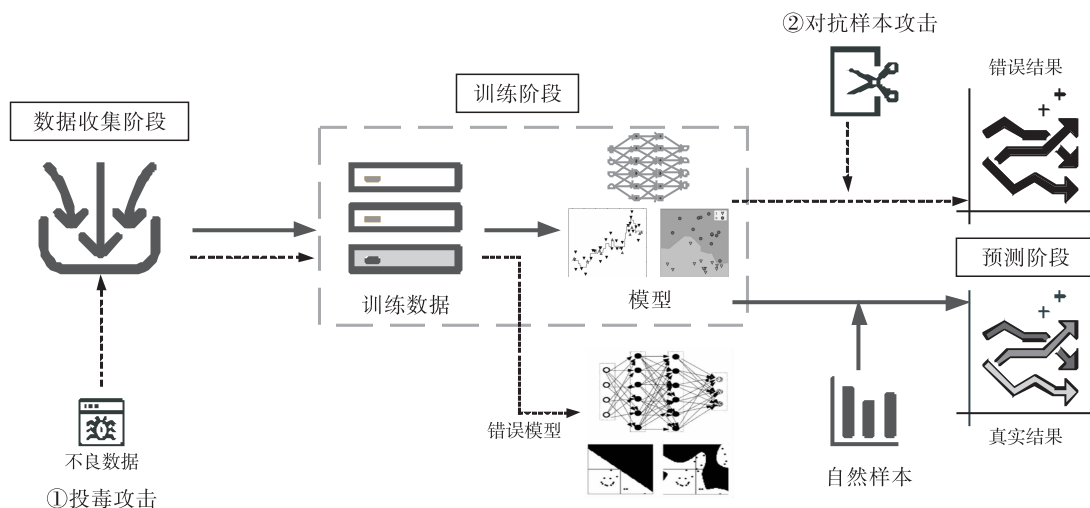


图1 机器学习算法流程、面临功能风险

投毒攻击<sup>[31]</sup>是一种典型“谎话百遍成真理”现象,主要发生在训练阶段,是攻击者利用机器学习应用需要不断更新训练数据的契机,通过修改、删除或注入不良数据,构造的恶意数据,并将此数据注入到训练数据集,在模型拥有者毫无察觉的情况下,改变原有的数据集的概率分布,来形成模型“后门”,使训练出来的模型边界发生偏移或使模型精度降低,导致模型训练结果产生偏差,得到错误结论的算法模型。投毒攻击能够成功的原因也在于,当前的机器学习算法一直假定数据都是

来源于客观世界,数据在时间轴上的分布是相同的,而一旦这个假设前提被攻击者打破,其后续算法运行结果便不可能正常。

对抗样本攻击<sup>[31]</sup>通过各种方式改变模型输入数据的特征来影响模型的预测效果,主要发生在预测阶段。攻击者常常使用优化求解技术在样本空间中搜索对抗样本、利用特征信息(如梯度信息)构造对抗样本和生成模型直接生成对抗样本等方法来产生对抗样本,并且采用 L2 范数来控制以及衡量噪声大小等方式来降低人类对对抗样本的感知度,从而实现利用人类无法察觉的对抗样本使模型以高置信度给出一个错误输出的攻击目的。例如在人脸识别系统中,眼部及其周围区域像素是算法特征数据重点提取来源,只用一副眼镜,就使得算法获得的眼部特征形成巨大差异,导致系统无法识别。它能够成功的原因主要在于理想模型和实际从数据中训练出来的模型之间特征维度具有差异。因为样本空间覆盖面相对有限的缘故,训练出来的模型,无法学习到所有特征以及特征之间的联系,这就给攻击者极大的操作空间。

### (三)机器学习算法功能风险对智能教育的危害

由于机器学习算法功能是智能教育能够实现的主要途径,是人工智能教育应用能够运行的核心支柱,更是应用在激烈的市场竞争中赖以生存的立身之本,而近期的研究表明对抗样本攻击与投毒攻击对算法功能的攻击成功率极高,造成的后果较为严重<sup>[11,34]</sup>。一来,攻击者可以利用对抗本来实施针对机器学习教育应用的对抗样本攻击和恶意侵扰来逃避检测,这会导致应用的功用大打折扣。例如,将待检测的作业处理成对抗本来逃避批改系统的检测,使系统无法做出正确的批改,应用的可用性、准确性都大大降低。再者,攻击者可以通过“数据投毒”,即在训练数据里加入伪装数据、恶意样本等破坏数据的完整性,进而导致训练的算法模型出现偏差。例如,在作业批改式应用中,被污染的数据输入模型后,导致模型决策出现偏差,并将错误的批改结果反馈给教师和学生。直接导致用户体验变差,降低客户信心和黏度。

令人不安的是,传统的信息系统安全领域下的数据隐私保护问题,可以通过诸如随机化响应以及隐私保护的数据发布机制<sup>[35-37]</sup>予以解决,而干扰应用功能的投毒攻击和对抗样本攻击作为非传统的技术攻击手段,很大程度上无法使用传统信息安全技术规避。因为即使不存在任何前置背景信息,攻击者还是能够依据行业经验进行投毒和对抗样本攻击。虽然当前计算机领域研究人员在极力研究应对方案,但目前为止,并未有特别行之有效、一劳永逸的抵御方案。

## 四、典型人工智能教育应用面临的算法功能安全风险分析

传统的教育信息系统功能被干扰的现象中,攻击者往往通过系统软件开发或网络安全设置缺陷进入系统,实施修改数据乃至关停服务等行为。而前文所述算法中的干扰系统功能的行为都是通过正常的样本输入行为,来误导或者诱导算法模型产生误判,行为具有较高的迷惑性和隐蔽性。为使人工智能教育应用从业人员认识到这种功能安全风险的迷惑性、隐蔽性和危害性,本文接下来结合具体应用场景对干扰系统功能的投毒攻击与对抗样本攻击的发动原理与攻击后果做详细分析。

### (一)典型人工智能教育应用及其采用的机器学习技术介绍

通过对市场上已有的人工智能教育应用进行系统梳理,根据其服务的对象、所服务的教学与管理环节,我们对现有典型人工智能教育应用进行归类,如表 1 所示。其应用功能逻辑关系如下:

对于教师,现有的人工智能教育应用一方面致力于帮助其对自身课堂教学行为进行分析评估,

主要通过人工智能课堂观察的视频分析功能实现;另一方面致力于减轻其现有教学负担,主要通过自动化作业批改(即作业批改及反馈)及自动生成学生过程性统计分析报告(即长期综合评价)来实现。

对于学生,现有人工智能教育应用一方面致力于在学生开展自主学习时主动提供学习建议,包括学习前的个性化学习路径规划、学习中的内容动态推送、学习后的自适应评测三方面来实现;另一方面致力于在学生被动练习时提供不同形式的信息交互反馈服务,包括图文交互的拍照搜题(有答案时的答案搜索,无答案时的人工在线解答服务)和语音交互的人机互动答疑(教育机器人)。

对于教学管理人员,现有人工智能教育应用主要致力于为其提供智能考务服务,包括考试前的自动化考场分配和智能组卷、考试后的智能阅卷和自动化统计分析。

更具体的功能、技术及其面临的安全风险见后文的详细说明。

表 1 现有典型人工智能教育应用分类

服务对象	教学与管理环节	应用类型	核心功能	核心技术	适用场景	代表性应用
教师	教学诊断	课堂教学分析类	视频分析	①关键教学行为帧提取 ②教学行为标注生成	评价教师课堂教学:通过对课堂中的教学言语和行为等交互数据的分析来进行教学诊断及评价分析	EduSense 系统
	学习诊断	全面智能测评类	作业批改与反馈	①自动批改 ②点评反馈	帮助教师批改作业;替代教师完成批改作业等机械性重复性的教学辅助工作;	批改网
			长期综合评价	①记录学习者的成长轨迹 ②学生个体画像 ③学情分析报告	帮助教师掌握学生学习状况;收集学生的全程学习数据,帮助教师评价学生能力	智慧学伴
学生	主动学习	智能教学平台类	自适应测评 个性化学习路径规划 学习内容动态推送	错误知识模型 学习活动序列 ①实时学习数据收集 ②学习内容匹配计算	学生主动进行个性化学习:系统根据学生的学习过程数据为不同学生创建个性化的自适应学习方案与学习路径,推送符合每个学生特点的差异化学习与课程规划。	松鼠 AI 英语流利说 Knewton
	辅助学习	拍照搜题类	拍照搜题	文字识别模型	帮助师生和家长解决作业难题:遇到不会的题目时可以用手机拍照并上传,系统通过图片识别和比对可以快速反馈对应的答案和解题思路;也可在线匹配教师进行 1v1 答疑	作业帮搜题 小猿搜题
			匹配师生答疑	师生需求匹配		
教学管理	考试管理	智能考务类	智能学习助手类	人机互动答疑 人机会话	陪伴学生完成学习全过程,并在期间替代教师为学生答疑解惑	Dino 机器人 阿尔法蛋
			考场与监考分配	动态规划算法		七天网络
			智能组卷	试题难度模型	帮助安排考试:从试卷编制、考场分配到试卷批阅、成绩分析,覆盖考试全过程	BookSeeit ETS
			智能阅卷	光学字符识别模型		
			成绩统计分析	分类方法的学生成绩等级划分		弘成 OTS 在线考试系统

## (二)典型人工智能教育应用面临的算法安全风险分析

### 1. 面向教师的人工智能教育应用面临的算法安全风险

#### (1)课堂教学分析类应用的安全风险

课堂教学分析类机器学习教育应用利用诊断平台对常规课堂教学视频素材进行识别,得到课堂中的教学言语和行为等交互数据,然后结合教学诊断、教学设计改进和教学评价三个场景,提供可解释的诊断结果。

视频分析是课堂教学分析应用中,教学诊断模块实现教学事件分类识别、教学设计模块完成教学法结构分析的核心技术,也是教学评价模块完成教学评语和教学事件绑定的关键支撑。典型的基于深度学习的视频分析系统主要包含两个部分。一是使用卷积神经网络完成视频中关键帧提取;二是使用循环神经网络,进行语音识别生成语句标注。

在关键帧提取阶段,系统可能受到对抗样本攻击。攻击者可以通过对关键帧图像添加少量肉眼觉察不到的扰动,如叠加一个小向量来扰动关键帧图像构造对抗样本,或通过生成特定于任务损失减数的对抗样本,甚至通过直接改变关键帧图像中的一个或几个像素,来实现对抗攻击。最终,误导卷积神经网络做出错误的分类,干扰教学诊断中的事件识别及分类功能。此外,系统也可能受到投毒攻击的安全威胁,攻击者可以通过模型后门对识别与分类进行操控,使模型无法识别某些教学行为或者对教学行为错误分类,影响教学诊断模块的行为分析。

在标注生成阶段,攻击者可以通过三种方法来构造对抗样本发动攻击。一是标注克隆攻击,对一张图片和一个目标标注句子生成一个对抗样本,使得标注系统在其上的标注与目标标注完全一致;二是标注异化攻击,对一张图片,生成一个对抗样本,使得标注系统在其上的标注与原标注无关;三是关键词攻击,对一张图片和一组关键词,生成一个对抗样本,使得标注系统在其上的标注含有所有的关键词。对标注系统的对抗攻击将直接影响对教学事件的特征分析与分类,导致系统无法完成教学事件类型分类和时间分布图生成,进而导致逆向分析评语无法实现,直接导致系统丧失教学评价功能。

#### (2)全面智能测评类应用的安全风险

全面智能测评类应用能够高效完成人类的体力劳动、脑力劳动或者认知工作,被广泛运用于两大场景,一是替代教师完成批改作业等重复性、机械性的教学辅助工作,帮助教师节约时间与精力;二是收集并分析平台中记录的学习数据,帮助教师对学习者的学习表现及效果进行评价。

全面智能测评类应用的主要功能是作业批改与反馈和长期综合评价。①作业批改与反馈即将作业与标准答案进行对比分析来完成自动判改并生成全面且精细的点评反馈。②学生长期综合评价即系统对学生的行为数据和学习数据等进行学情分析,通过时间轴的方式记录学习者的成长轨迹,形成对学生个体与学生整体的画像,并生成可视化的学情分析报告。

在作业批改和反馈系统中,攻击者可以通过将作业文本更改少量的字符,或引入一些类似于人们现实中也会写的“错别字”,再或进行不改变语义的重述来构造对抗样本,实施针对自然语言处理技术的文本对抗攻击。这种文本对抗攻击一方面会使系统在进行学习者语料与标准答案语料数据比对分析时无法正确测量出两者之间的“距离”数据,导致发生误判。另一方面可能会导致系统无法按特定的分析规则将“距离”映射转化为用户可理解的分数、总评语、按句评语等反馈内容,使系统丧失错误解析的功能。除了攻击者外,有些弄虚作假的用户也可以发动投毒攻击来提高自己的

分数。一方面,他们可以通过观察大量反馈内容推测系统的评分标准,找出“得高分”的方法,在答题时欺骗机器获得高分。另一方面,他们也可以持续向系统输入与题目毫无联系、无意义的内容作为答案来进行数据投毒,就可以将反馈系统武器化,并以此攻击其他合法用户和内容,让其他用户的答案无法获得正确的批改,从而提升自己的成绩。

在学生长期综合评价系统中,学生分类是系统进行学生画像与群体画像的重要支撑。学习行为数据和作业等学业数据在一定程度上能够反映学生多方面的能力,是对学生进行长期、综合评价的重要指标。通过对这些数据进行分析,系统能够评价学生在不同能力指标上隶属于不同的水平类别,然后将学生在不同能力上的水平标签汇聚起来进行综合的评价,再搭配学习者的成长轨迹,形成学习者个体画像,实现对学生进行长期综合评价。对分类功能产生巨大干扰的往往是投毒攻击与对抗样本攻击。在投毒攻击方面,一方面,攻击者可以通过污染学生行为数据集来让分类模型对能力水平高低的分类边界发生偏移,从而降低模型的准确率。另一方面,攻击者也可以通过后门,注入特定训练数据,操控模型的核心算法分类树,使模型的分类功能按照自己的意愿进行,比如,将自己分类到高能力水平,获得高评价为自己谋名利,或使他人获得差的评价以提升自己的排名。在对抗样本攻击方面,攻击者可以通过设计特定样本来迷惑分类器,使能力水平分类结果产生偏差,进而导致学生个体与整体画像形成受阻,无法生成可视化的学情分析报告。

## 2. 面向学生的人工智能教育应用面临的算法安全风险

### (1) 智能教学平台类应用的安全风险

智能教学平台的核心是人工智能自适应学习系统,其主要功能是运用人工智能技术分析学习者所学内容,构建学习者知识图谱,为学习者提供个性化的学习内容以及学习方案;支持自适应学习,实现学习内容的智能推荐,为教师和学生提供个性化教与学服务。包括自适应测评、个性化学习路径规划和学习内容动态推送三个模块。①自适应测评能够实时测评学生对每个知识点的掌握水平,自动识别学习者的错误并推断错误的原因,实现学生知识水平测量、错误模型搭建、学生个性特征挖掘(如认知特征、情感特征等)。②个性化学习路径规划能自动识别学习需求,根据用户特征信息(如学习偏好、知识水平等)动态呈现个性化的学习活动序列(含学习对象),从而更好地完成知识建构。③学习内容动态推送能够针对学生的个性化学习路径,匹配最合适的学习内容,使内容难度与学生能力匹配、内容类型与学生偏好匹配,从稍高于学生当前的水平逐渐增加,形成循序渐进的学习路径,让学生不断获得成就感,提升学习乐趣。

错误模型是平台实现自动识别错误和推断错误原因功能的基础,常以学习者的错误/误解数据为输入,基于摄动模型和约束模型来建模。在错误模型搭建时,攻击者可以通过构造对抗样本发动对抗攻击,使摄动模型发生分类错误,无法将正确的反应形成领域知识的规则,也无法将错误反应存入错误库中成为错误规则,从而使系统丧失自动识别错误的功能。约束模型通过分析学生求解问题时所达到的问题求解状态来推断学生的出错原因。这意味着约束模型所接收的训练数据是不断更新的学生响应数据,若此时攻击者向训练数据中投入大量的恶意数据,很可能会影响模型的预测功能,导致模型丧失推断错误原因的功能。

学习活动序列是学习路径的主要内容,能够根据学习者特征模型构建知识学习的先后次序,它的实现依赖知识图谱。知识图谱包括知识拆分、打标签两个步骤。知识点拆分是构建知识图谱的基础,它根据知识难易程度与知识点间关联程度对知识进行划分。敌手可以通过模型后门,在模型

拥有者毫不知情的情况下操控分类树进行错误的知识点分类。打标签是构建知识图谱的核心,也是依据知识图谱进行学习路径规划的关键。每个知识点都要打上标签,标签包括内容、难易度、区分度等等,知识点的颗粒越细,标签越多,匹配学习路径时就更精准。攻击者可以在打标签时发动对抗样本攻击,使知识点标签缺失、错误,进而导致正确的学习序列无法形成。攻击者也可以发动投毒攻击,在系统进行标签更新时,注入大量的恶意数据,使打标签模型对知识点的分类发生偏移,降低标签的准确性。

学习内容动态推送包括实时数据收集和-content匹配计算。数据的实时收集既是自适应系统具备自适应性的关键,也是招致安全攻击的根源,在数据的收集过程中,攻击者能够利用数据的实时性,向数据中掺入恶意数据,破坏自适应序列的形成。内容匹配计算基于回归树与关联规则算法实现,在进行内容匹配时,攻击者可以通过模型后门,对回归树进行操控,致使系统无法推测学生所处的能力层次,无法匹配出恰当难度与类型的学习内容。攻击者也可以通过构造对抗样本,使系统无法通过关联分析梳理内容间的联系、制定合适的学习顺序,导致推荐的学习内容无法形成循序渐进的学习模式。

## (2)拍照搜题类应用的安全风险

拍照搜题类应用通过大规模题库支持,基于图像识别技术,由系统匹配题库,自动为学习者返回题目答案和解题方法并调配教师进行一对一在线答疑。

拍照搜题类应用主要由拍照搜题和在线答疑两个模块组成。其主要工作流程为:当学习者在学习中遇到疑惑时,利用手机拍照功能拍下题目上传搜索即可获取题目及答案,若系统呈现的内容无法解决学生问题时,学生可以发出在线答疑请求,系统将自动根据学生情况匹配合适的优质老师进行一对一实时答疑。主要使用两方面技术:①利用图像识别技术、自然语言处理技术自动地为学生提供搜索题目中所包含的知识要点以及难点;②利用分类与回归树、遗传算法自动地为学生匹配适合的教师进行1V1讲解。

拍照搜题模块的主要任务是将拍照图片转成文字,然后进行文本分析,再由系统匹配题库,自动为学习者返回题目答案和解题方法。当攻击者对系统发起对抗样本攻击时,图像识别和文本分析功能会受到影响。对于图像识别技术来说,当上传的题目被切割为一个个的字以后,对抗样本会在单字识别模型对每个字进行多次卷集和下采样时进行干扰,导致模型无法实现题目文字的识别,即图像识别技术失效。对于文本分析技术来说,对抗样本能够对自然语言推理模型发动对抗攻击,导致其不能对识别出的文字做自然语言处理,无法完成对题目的理解与分析。在此情况下,系统将无法与题库匹配,反馈相应的题目答案和解题方法,拍照搜题功能将完全丧失。

在线答疑模块的核心功能是为学生找到最合适的讲题教师,其功能实现的基础是分别对学生与教师进行分析形成学生画像与教师画像,依据画像特征实现最合适的匹配。一方面,在学生画像的生成过程中,攻击者通过对抗样本攻击,能够干扰系统根据树状结构对题目和知识进行分类的功能,使其无法完成学生知识水平分析,导致系统无法完成知识结构建模,形成学生画像。攻击者也可以发动投毒攻击,在神经网络模型的输入中掺杂有毒数据,导致模型无法准确抽取学生的各种知识特征,对接下来教师分配和调度的策略会产生影响。另一方面,攻击者也可以利用投毒攻击干扰分类树算法,对教师画像的形成进行干扰,导致系统无法对教师进行分类;也能对回归树进行干扰,导致系统无法预测知识点讲解教师的供应量和知识点学习学生的需求量,无法完成在线师生资源



的匹配。

### (3)智能学习助手类应用的安全风险

智能学习助手是为促进智能化教与学而设计的一种人工智能应用,具有较强的互动和沟通能力,能够扮演教师、学习同伴、教学助理或顾问等多重角色,与使用者进行互动。该类应用不仅在课前备课、课堂互动和课后评价方面对教师起着辅助作用,而且在提高学生学习兴趣,培养学生分析能力、创造能力和实践能力等方面也发挥着重要的作用。

智能学习助手的主要功能是人机互动答疑功能。该类应用提供面向课前、课堂、课后的全学习场景的应用与服务,能够在学生学习全过程中随时与学生进行人机互动答疑,及时为学生解惑,提升学生学习体验。

互动答疑系统能够直接与用户沟通,回答用户问题并给予指导意见,大体分为接受、分析和反应三个阶段。首先在接受用户信息阶段,攻击者能够对语音识别中的循环神经网络发动对抗攻击和投毒攻击,使系统无法完成从所接收的信号中提取合适的特征向量来训练声学模型,导致后续语音解码和搜索算法的分析过程无法进行。其次,在NLP的语义分析功能的意图识别阶段,攻击者也可以采取投毒攻击,对文本向量化的词袋模型或词向量进行干扰,导致系统无法完成语言模型构建,无法完成文本处理与分析。此外,攻击者也可以发起对抗样本攻击,使系统无法将用户的话语分类到相应的意图种类,导致系统不能准确的理解用户的意图,不能给予用户准确的回复。在反应阶段,攻击者可以通过干扰循环神经网络,对语音合成过程中的语言处理、韵律处理和声学处理进行干扰,导致系统无法合成语音,以语音形式与用户沟通。

## 3. 面向教学管理的人工智能教育应用面临的算法安全风险

### (1)智能考务类应用的安全风险

智能考务是近年来教育人工智能领域中兴起的又一应用场景,它能够管理及规范考前、考后相关工作流程,减轻管理人员工作负担,降低考务管理成本。其主要功能包括考场与监考分配、智能组卷、智能阅卷、成绩统计分析四个方面。

考场与监考分配模块采用智能化的孤岛编排模式来自动编排考场,使每个考生前后左右的考生都来自不同学校或班级,以降低抄袭的可能性。该模块的核心是动态规划算法,当敌手发动对抗样本攻击时,状态转移方程无法寻找到合适边界条件,导致动态规划算法失效,使系统的孤岛编排模式失去效用。

在智能组卷模块中,能够进行可视化组卷,还能够进行试题难度预估,从而保证试卷质量。系统通过对历年的考试大数据<sup>[38]</sup>进行整合,经过多次训练迭代建立试题与难度之间的对应关系,并以此为基础训练出试题难度模型。当进行过分词、词嵌入、深度表征等预处理的新试题投入到模型中时,系统会自动计算出该题在题库中所对应的难度值,从而为命题人提供参考。该功能模块的核心技术支撑是分类决策树,当攻击者发动投毒攻击或对抗样本攻击时,都能破坏决策树算法的分类功能,进而影响题目知识属性标签以及难度值的生成,导致系统无法根据搜题算法在题库中选择合适的题目组卷,也无法为命题人提供难度值参考。

智能阅卷模块是智能教务系统中涉及人工智能技术最多的功能板块,包括问卷扫描和考试阅卷两个阶段。在问卷扫描阶段,攻击者可以通过投毒攻击使光学字符识别模型的分类器发生偏移,导致其无法将纸上的字符正确转化成计算机文字,从而无法完成客观题的批改。在考试阅卷阶段,

攻击者可以通过将考生的答题卡构造成对抗样本来实施对抗样本攻击,使光学字符识别使用的卷积神经网络发生误判,最直接的后果是导致客观题的判改发生错误,会对智能阅卷的准确率产生重大影响。

在成绩统计分析模块中,系统在前述模块记录的考生的答题内容及评分结果数据基础上,再通过统计分析来提供学生成绩分析和统计报告。该模块的功能核心是基于分类方法的学生成绩等级划分,当受到投毒或对抗样本攻击时,分类模型的边界会发生偏移,导致分类方法无法正确进行学生成绩等级划分,影响成绩统计分析对教学的反馈指导作用,使课堂教学安排和针对性强化训练的实施受到干扰。

### 五、人工智能教育应用面临的算法功能安全风险消解建议

与以往人类遭遇的新型安全风险一样,人工智能教育应用面临的这种算法功能被干扰的安全风险结局有两种。其一是在人工智能教育应用真正大面积落地应用之前,基础算法在未来不长的时间得到长足发展,功能安全风险被系统性消除,这是最理想的一种结局。其二,假若这种技术风险在短期内无法消除,或者就如同流感病毒与人类长期共存一般,形成道高一尺魔高一丈的对立统一关系,那人工智能教育应用便面临着不得不带病上阵的局面。在这种局面下,应当更多从运行管理层面给予补救。本文初步给出如下建议:

#### (一)建立人工智能教育应用算法风险预警机制

鉴于干扰应用功能风险的客观存在,教育用户需要与风险共存,管理部门和服务提供商,应该让用户充分知晓这种风险的存在。需要综合多方因素,客观评估应用功能与利益关联程度。如利益相关因素:考试成绩和升学直接相关,干扰成功获益巨大,招致恶意干扰的风险可能性越大;应用市场竞争状况:普及发展期的蓝海市场,各方都在增长,相互恶性竞争小,制造干扰可能性大,而存量优化的红海市场,各方零和博弈,相互恶性竞争大,制造干扰可能性强等等。最终给予不同类别、不同发展阶段的应用,标以不同级别的风险标识,在用户使用相关应用时予以告知,以避免用户对应用系统的盲目信任,而忽视了基本的防范。

#### (二)加强应用外部用户访问秩序管控。

投毒攻击是外部攻击者利用算法更新训练数据集的时机,向应用系统输送大量“有毒”数据,导致模型边界发生偏移。因此,应该严格限定数据来源,如摒弃外部来源数据,尽可能由服务提供商或用户自己收集并更新训练数据集,如拍照搜题类应用,试题数据应由服务提供商组织专班人马采集整理数据,而不应该简单依赖用户提供数据。而对抗样本攻击中,攻击者欺骗模型的对抗样本,是通过大量测试应用模型结果,试出导致模型错判的样本。因此,应限制单个用户使用服务频率和总次数,如采取类似锁屏解锁的惩罚性冻结时间机制,对于明显异于常人的用户,需要重点关注,并加以限定措施。通过上述缩小训练数据来源、限制应用访问速率的机制,尽可能减少外部攻击者干扰系统的机会。

#### (三)确立应用运营的内部监管奖惩纪律

通过最小化应用开放时空范围,一定程度上压制了外部恶意干扰行为的实施机会,但仍存在着系统内部用户变节、内外串通可能性。因此,对于训练数据,为了防范投毒攻击,应建立从采集到录入到存储到使用全过程的责任到人管理制,清晰标明数据的来源、行为、时间和目的,可以辅助以传

统数据安全技术,如区块链技术进行数据安全防护。对于一些敏感应用的测试使用,如智能考务类,对于每一次应用测试,都要执行如申请审核、伴随式记录方法,以及人工审核服务功能,并对测试样本和结果存档备案,以备复核,防范对抗样本攻击,这方面可以借鉴当前公安系统内部采用的身份证查询系统使用管理机制。

#### (四)构建应用安全运营法律防范体系

教育类应用具有巨大的社会影响,其对应的不仅仅是教育问题,更多时候往往涉及到教育公平、社会正义。干扰人工智能教育应用功能正常运行,误导教育教学行为和结果,会给教育现代化进程蒙上阴影,轻则影响教育、教学秩序,重则影响学生身心健康,激发民愤、民情。因此,针对恶意商业竞争,干扰正常运营服务商的实体和法人,一经查实,应当立法实行教育行业禁入;恶意协助外部第三方注入有毒数据、搜寻对抗样本的内部工作人员,一经确认,应当予以行政处罚并调离现有岗位,有犯罪行为的,应当移交司法部门审判;通过干扰应用功能,获取不正当利益的用户,应当取消误导应用得出偏颇结论获得的所有不当名誉和权益,并考虑计入诸如学术不端档案的个人信用库。

### 六、结语与展望

在人工智能教育应用面临着诸多传统外部风险同时,人工智能算法自身内部也蕴含着功能安全风险。或者说,人工智能教育应用应用核心技术——机器学习算法面临着投毒和对抗样本两类攻击风险。它干扰了应用功能正常运行,是一种非传统的安全风险,是机器学习算法所特有的技术缺陷,且当前并未有较好的办法予以根绝。而对当前十分热门的智能教学平台、课堂教学分析、全面智能评测、拍照搜题、智能学习助手和智能考务六类典型的人工智能技术教育应用的深入分析,也进一步论证了干扰应用功能这一非传统安全风险,动摇了所有人工智能教育应用正常运行的根基,且无一能够幸免。这种非传统安全风险一旦在关键事务上造成重大系统安全事故,形成对人工智能技术教育应用的普遍负面印象,必将影响后期正常的技术应用,亦或直接成为人工智能教育应用的阿喀琉斯之踵。

尽管人工智能技术在过去七十余年的发展,已经经历过三起两落。许多从业人员认为人工智能技术已经渡过了技术发展前期的脆弱期,不会再出现根本性的发展波折了。但辩证法告诉我们,事物的发展从来不会是一帆风顺,而是螺旋梯度、曲折上升,对人工智能教育应用中要害问题的麻痹大意,将会导致人工智能教育应用事业发展进入下降螺旋。要认识到的是,人工智能教育应用面临的要害问题很多,但很显然,算法安全问题是其中十分关键的一个。一些学者认为人工智能算法安全攻击条件高,实施难度大,造成现实危害的可能性不大,因此可以忽略。但以发展的眼光看,只要攻击成本小于回报,它就有存在空间,并且随着资金和人才的富集,算法安全攻击的成本会急剧下降。因此,人工智能算法安全攻击问题解决不好,将会进一步摧毁人工智能在教育行业应用的生态基础。由于教育涉及面的广泛性、影响力的持久性,其对人工智能技术社会信任度的打击巨大,而信任一旦失去,人工智能技术发展的前景便不复存在,最终可能导致人工智能的发展第三次跌入深渊,而这一局面是我国现代化事业发展的不可承受之重。

### [参 考 文 献]

- [1] 祝智庭,韩中美,黄昌勤.教育人工智能(eAI):人本人工智能的新范式[J].电化教育研究,2021,42(01):5—15.

- [2] 张志祯,张玲玲,李芒.人工智能教育应用的应然分析:教学自动化的必然与可能[J].中国远程教育,2019(01):25-35+92.
- [3] 戴静,顾小清.人工智能将把教育带往何方——WIPO《2019 技术趋势:人工智能》报告解读[J].中国电化教育,2020(10):24-31.
- [4] 刘邦奇,王亚飞.智能教育:体系框架、核心技术平台构建与实施策略[J].中国电化教育,2019(10):23-31.
- [5] 徐晔.从“人工智能教育”走向“教育人工智能”的路径探究[J].中国电化教育,2018(12):81-7.
- [6] 奥拉夫·扎瓦克奇-里克特,维多利亚·艾琳·马林,梅丽莎·邦德,et al.高等教育人工智能应用研究综述:教育工作者的角色何在?[J].中国远程教育,2020(06):1-21+76.
- [7] 贾珍珍,刘杨钺.总体国家安全观视域下的算法安全与治理[J].理论与改革,2021(02):135-48.
- [8] CHRISTIAN B, GRIFFITHS T. Algorithms to Live By: The Computer Science of Human Decisions[M]. Algorithms to Live By: The Computer Science of Human Decisions, 2016.
- [9] HOADLEY D S, LUCAS N J. Artificial intelligence and national security[M]. Congressional Research Service Washington, DC, 2018.
- [10] 陈全真.智能机器人权利存在的由因及对策[J].贵州师范大学学报(社会科学版),2019(03):144-51.
- [11] 何英哲,胡兴波,何锦雯,et al.机器学习系统的隐私和安全性问题综述[J].计算机研究与发展,2019,56(10):2049-70.
- [12] 魏立斐,陈聪聪,张蕾,et al.机器学习的安全问题及隐私保护[J].计算机研究与发展,2020,57(10):2066-85.
- [13] 胡圣波,朱满琴,杨露露,et al.未来无线通信与大数据、人工智能[J].贵州师范大学学报(自然科学版),2020,38(06):1-10+132.
- [14] 薛庆水,李凤英.人工智能教育应用的安全风险与应对之策[J].远程教育杂志,2018,36(04):88-94.
- [15] 刘睿瑾,陈红,郭若杨,et al.机器学习中的隐私攻击与防御[J].软件学报,2020,31(03):866-92.
- [16] 谭作文,张连福.机器学习隐私保护研究综述[J].软件学报,2020,31(07):2127-56.
- [17] 张安毅.人工智能侵权:产品责任制度介入的权宜性及立法改造[J].深圳大学学报(人文社会科学版),2020,37(04):112-9.
- [18] 杜静,黄荣怀,李政璇,et al.智能教育时代下人工智能伦理的内涵与建构原则[J].电化教育研究,2019,40(07):21-9.
- [19] 钱小龙,张奕潇,宋子昀,李强.教育人工智能系统的伦理原则与困境突破[J].江南大学学报(人文社会科学版),2021,20(06):96-104.
- [20] 卢迪,段世飞,胡科,et al.人工智能教育的全球治理:框架、挑战与变革[J].远程教育杂志,2020,38(06):3-12.
- [21] 胡元聪,李雨益.企业社会责任视域下人工智能产品风险防范研究[J].当代经济管理,2020,42(04):19-26.
- [22] 杨蓉.从信息安全、数据安全到算法安全——总体国家安全观视角下的网络法律治理[J].法学评论,2021,39(01):131-6.
- [23] 季卫东.人工智能开发的理念、法律以及政策[J].东方法学,2019(05):4-13.
- [24] 曹珍富.信息安全的新发展——为《计算机研究与发展》创刊六十周年而作[J].计算机研究与发展,2019,56(01):131-7.
- [25] 方滨兴,任奎,贾焰.网络安全的新领域[J].Engineering, 2018,4(01):7-10.
- [26] 祝高峰.论人工智能领域个人信息安全法律保护[J].重庆大学学报(社会科学版),2020,26(04):150-60.
- [27] 张坤颖,张家年.人工智能教育应用与研究中的新区、误区、盲区与禁区[J].远程教育杂志,2017,35(05):54-63.
- [28] 刘梦君,姜雨薇,曹树真,et al.信息安全技术在教育数据安全与隐私中的应用分析[J].中国电化教育,2019(06):123-30.
- [29] 刘梦君,许明雪,宗敏,et al.区块链技术助力新高考改革:问题、措施与挑战[J].中国电化教育,2020(11):104-11.
- [30] 田贤鹏.隐私保护与开放共享:人工智能时代的教育数据治理变革[J].电化教育研究,2020,41(05):33-8.

- [31] 李欣姣, 吴国伟, 姚琳, et al. 机器学习安全攻击与防御机制研究进展和未来挑战[J]. 软件学报, 2021, 32(02): 406—23.
- [32] 陈宇飞, 沈超, 王骞, et al. 人工智能系统安全与隐私风险[J]. 计算机研究与发展, 2019, 56(10): 2135—50.
- [33] 闫怀志, 胡昌振, 谭惠民. 网络安全主动防护体系研究及应用[J]. 计算机工程与应用, 2002(12): 26—8.
- [34] 纪守领, 杜天宇, 李进锋, et al. 机器学习模型安全与隐私研究综述[J]. 软件学报, 2021, 32(01): 41—67.
- [35] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning; proceedings of the proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, F, 2017[C].
- [36] 刘梦君, 贾玉娟, 姜庆. 差分隐私保护的学习资源学习热度推荐[J]. 现代教育技术, 2019, 29(05): 99—105.
- [37] 侯浩翔. 人工智能时代学生数据隐私保护的动因与策略[J]. 现代教育技术, 2019, 29(06): 12—8.
- [38] 王蕾, 佟威. 赋能教育考试新基建 助力考试战线新发展——国家题库 2.0 创新实践[J]. 中国考试, 2021(02): 34—9.

(责任编辑: 闫卫平)

## AI Education Fuses Safety Alerts: Native Risk Analysis from Machine Learning Algorithmic Functions

LIU Meng-jun, JIANG Xin-yu, SHI Si-jin, YU Gang, JIANG Nan, WU Di

(Hubei University, School of Education, Hubei, Wuhan 430062)

**Abstract:** Security and privacy are the top issues in the 2021 Horizon Report. Aiming at the problem of functional security holes existing in machine learning algorithms, the core foundation of artificial intelligence technology, this paper firstly introduces the risk of poisoning attack and anti-sample functional attack faced by machine learning algorithms in principle; Then six kinds of typical artificial intelligence education application are deeply analyzed, including intelligent teaching platform, intelligent behavior analysis, intelligent evaluation of classroom teaching, intelligent teaching assistant teaching and learning, intelligence, intelligent examination applications, functions of risk in the process of machine learning algorithms fusion depth profiling. The results show that the algorithm function attack shook the foundations of normal operation of all of the above applications. Finally, in the context that there is no good technical solution at present, the research puts forward some preliminary suggestions on how to defend against the algorithm function attack that the educational application of artificial intelligence is facing from the aspects of law, system, organization and management.

**Key words:** artificial intelligence education; machine learning; security risk; algorithm defect; function safety