

陕西省技术发明奖公示信息

(2025年度)

一、项目基本情况

项目名称	金融人工智能安全检测与服务关键技术及其应用
主要完成人	沈超；蔺琛皓；周俊；王骞；李前；张志强
主要完成单位	西安交通大学；支付宝（杭州）信息技术有限公司；武汉大学

二、提名意见（适用于部门、机构提名）

提 名 者	陕西省教育厅	提名等级	<input checked="" type="checkbox"/> 一等奖 <input type="checkbox"/> 二等奖及以上
<p>提名意见：</p> <p>金融人工智能安全检测与服务关键技术及其应用是保障人工智能算法在反洗钱、反欺诈、保险风控、身份核验等金融场景下规模化应用的关键安全防线，一直是学术界与工业界共同关注的重要基础问题。针对金融人工智能算法面临的后门攻击、数字世界对抗攻击、物理世界对抗攻击以及深度伪造音视频攻击等严峻挑战，项目组历经八年，投入 610 余名研发人员，围绕数字世界黑盒对抗攻击检测、物理世界黑盒对抗攻击检测、深度学习算法后门检测、深度伪造音视频检测这四项金融人工智能算法全周期安全防护关键目标，攻克了相应核心技术，研制出金融人工智能算法安全测评与防护平台。该平台广泛应用于金融欺诈检测、非法内容识别与用户身份核验等关键风控场景，为政府机构、金融机构、企事业单位及个人用户提供了坚实的金融安全技术支撑。项目成果已在蚂蚁集团智能金融风险平台、支付宝 AI 平台、保险风控平台等实现产业化推广，服务覆盖欧美、日韩、东南亚等 25 个国家（地区）以及我国 34 个省市自治区，累计用户数达 12 亿（其中国内用户 8 亿）。已为全球 20 家金融机构与企业、超过 1000 万家中小微企业提供安全保障，覆盖资产规模超过 2 万亿元。服务对象包括中信银行、网商银行、中国人民保险集团、华为、荣耀、vivo、德国博世集团、荷兰飞利浦集团、孟加拉 bKash、泰国 TrueMoney、菲律宾 GCash 等国内外知名企业与机构，产生了超过 10 亿元的直接销售与服务收入，取得了显著的经济与社会效益。</p> <p>项目组完成人政治立场坚定，拥护党的领导，热爱党的教育事业，爱岗敬业。我单位认真审阅了该项目推荐书及附件材料，确认全部材料真实有效，提名该项目为陕西省技术发明奖一等奖。</p> <p>说明：省科学技术奖一、二等奖项目，实行按等级标准提名、独立评审表决的机制。提名单者应严格依据省科学技术奖的标准条件，说明提名项目的贡献程度及等级建议。“仅提名一等奖”评审落选项目不再降格参评二等奖。提名项目正式提交后，提名等级建议本年度不得变更。</p>			

二、提名意见（适用于专家提名）

姓 名			
专家类型	<input type="checkbox"/> 国家最高科学技术奖获得者 <input type="checkbox"/> 中国科学院院士 <input type="checkbox"/> 中国工程院院士 <input type="checkbox"/> 国家科学技术奖获奖项目第一完成人（需注明获奖等次） <input type="checkbox"/> 省最高科学技术奖获奖人（或 xxxx 年省科学技术最高成就奖、xxxx 年基础研究重大贡献奖获奖人） <input type="checkbox"/> Xxxx 年省科学技术奖第一完成人（需注明获奖等次）	提名等级	<input type="checkbox"/> 一等奖 <input type="checkbox"/> 二等奖及以上
责任专家	<input type="checkbox"/> 是 <input type="checkbox"/> 否		
提名意见：			
<p>说明：省科学技术奖一、二等奖项目，实行按等级标准提名、独立评审表决的机制。提名单者应严格依据省科学技术奖的标准条件，说明提名项目的贡献程度及等级建议。“仅提名一等奖”评审落选项目不再降格参评二等奖。提名项目正式提交后，提名等级建议本年度不得变更。</p>			

三、项目简介

金融人工智能安全检测与服务以人工智能安全技术为核心，涵盖金融人工智能算法训练、测试、部署的全生命周期，是保障人工智能算法在反洗钱、反欺诈、保险风控、身份核验等金融场景下可靠运行和规模化应用的关键安全防线。针对金融人工智能算法面临后门攻击、数字世界对抗攻击、物理世界对抗攻击、深度伪造音视频攻击等严峻挑战，结合 10 余项国家和企业委托的课题，本着“从实践中来，到实践中去”的原则，历经 8 年投入 610 余名研发人员，围绕数字世界黑盒对抗攻击检测、物理世界黑盒对抗攻击检测、深度学习算法后门检测、深度伪造音视频检测这四个金融人工智能算法全周期安全防护应用关键目标与难题，研制出金融人工智能算法安全测评与防护平台，广泛应用于金融欺诈检测、非法内容识别、用户身份核验等关键金融风控场景，为国内外政府机构、金融公司、企事业单位及个人用户的金融安全监管与防护等方面提供了关键技术支撑。项目主要技术发明与贡献如下：

1. 取得四项主要技术发明成果。

①发明了基于“参数迭代+聚类轨迹”的多环节人工智能模型后门检测组件。核心工作包括：发明了 AI 模型后门摄动输入聚类分析方法，利用 AI 模型对正常样本和携带后门触发器样本的行为差异，构造叠加输入样本集合进行检测，有效区分金融场景中的正常样本与后门样本，实现了模型输入后门触发器的检测；提出了 AI 模型后门隐层输出轨迹分析算法，追踪和分析模型对正常样本和后门触发器样本特征识别过程中的差异，实现了金融场景中分析输出分布与追踪隐层输出聚类变化的后门触发器检测、后门检测；发明了 AI 模型后门参数逆向还原分析方法，实现了金融场景中已知模型参数下的逆向模型参数分析、迭代求解后门触发器的后门检测、后门触发器逆向技术。研究成果有效支撑了蚂蚁集团及其用户公司金融风控业务场景中用户身份识别、活体检测认证等开源算法的后门触发器的检测与拦截，对部署的黑盒或灰盒模型，后门攻击检测准确率超过 90%，对白盒模型，后门攻击检测准确率超过 80%。

②发明了“漏洞挖掘+预警防御”的多维度数字世界人工智能模型漏洞检测组件。核心工作包括：发明了基于代理模型微调的量化金融模型漏洞挖掘方法，并提出混合精度激活值量化训练，实现了模型安全漏洞的全面挖掘；提出了基于逆向微小扰动的金融模型恶意行为检测预警方法，根据正常样本与对抗样本的行为不一致性实现对抗样本检测；提出了基于困难样本挖掘的金融模型防御增强方法，引入了困难/简单对抗样本概念，并提出了基于早期丢弃机制的高效对抗训练方法，有效提高了模型的鲁棒公平性；提出了困难样本权重重分配方法，通过动态权重调整优先优化更具威胁性的对抗样本，显著缓解了鲁棒过拟合难题。研究成果有效支撑了蚂蚁集团及其用户公司金融风控业务场景中黑灰产数据识别、用户身份信息一致性校验、保险服务风险管理等金融场景下的人工智能算法安全漏洞检测和可信安全评测与防御，数字世界黑盒及白盒对抗攻击样本生成场景中，攻击成功率大于 85%，数字世界对抗攻击防御场景中，防御成功率大于 80%。

③发明了基于“对抗伪装+延迟优化”的多类型物理世界金融人工智能对抗攻击漏洞检测组件。核心工作包括：发明了基于三维对抗伪装的对抗攻击检测算法，通过

纹理转换和物理增强模块并融入环境感知与自适应学习机制，扰动金融数据或模型输入，增加金融风险管理决策延迟并影响市场预测和风险评估；发明了基于延迟优化的目标跟踪攻击检测框架。通过增加系统响应延迟并干扰物体检测与跟踪并构建全面的防御网，使得攻击者难以绕过检测系统从而提高整体安全性，用于金融系统操控市场数据或模型输入，增加决策延迟并提升攻击检测精度。此外，发明了金融场景下的风险管理与决策支持方法，基于实际金融市场的波动性和不确定性，开发动态风险模型并根据市场变化实时调整风险参数，通过引入系统延迟作为风险管理工具，模拟不同延迟条件下的市场反应，提升金融攻击检测效果，减少潜在风险。在金融场景的物理世界黑盒及白盒对抗攻击中，攻击成功率超过 70%，检测成功率超过 80%。

④发明了基于“特征提取+真伪预测”的金融多模态生成式内容检测组件。核心工作包括：发明了基于知识驱动的动静态跨域局部差异分析生成式图像检测方法，实现了具有高精度高健壮性的伪造图像检测；发明了基于语义弱化与关键帧抽取的生成式视频检测方法，从而可以学习到生成模型在频域特征中留下的纹理伪影，实现了高效率高泛化性的伪造视频检测；提出了数据驱动的多类型多尺度异常特征挖掘音频伪造检测方法，利用多种特征提取方法和特征金字塔获得多类型多尺度的异常特征，提升分析的准确性和鲁棒性。研究成果有效支撑了蚂蚁集团及其用户公司金融风控业务场景中全媒体数字场景和金融账号活体检测认证、语音交互应用等实际业务场景下人工智能多模态音视频数据的真实性与安全性保障，支持深度伪造图像、视频、语音的检测成功率不低于 80%。

上述核心成果授权国家发明专利 59 项，发布国际/国家/行业标准 36 项，出版专著/章节 5 部，在 IEEE TIFS、IEEE TDSC、USENIX Security、CCS、NDSS、CVPR、NeurIPS、ICML 等人工智能和网络安全领域的国际顶级期刊与会议上发表论文 86 篇，获得最佳论文奖 9 项。

2. 研制出金融人工智能算法安全测评与防护平台。该平台包括全生命周期人工智能模型后门检测组件、多维度数字世界模型漏洞检测组件、多类型物理世界对抗攻击漏洞检测组件、多模态生成式内容检测组件，具有深度学习算法后门检测、数字世界对抗攻击漏洞检测、物理世界对抗攻击漏洞检测、生成式音视频内容检测等功能。

3. 实现了产业化推广应用。应用于蚂蚁集团的智能金融风险分析平台、支付宝 AI 平台、保险风控平台等，实现了产业化推广应用，服务于欧美、日韩、东南亚地区等 25 个国家（地区）以及我国 34 个省市自治区共计 12 亿用户（我国 8 亿用户），为全球 20 家金融机构企业（机构）、1000 万多家中小微企业的超过 2 万亿资产（包括国内金融保险企业如中信银行、网商银行、中国人民保险集团股份有限公司、人保健康、众安保险、中国人保寿险、太平人寿、太平养老、国泰产险、阳光人寿、中国人保国华人寿、中国平安财产保险、平安健康保险、阳光财产保险、泰康人寿保险等；国内民营企业如华为、荣耀、vivo、立白集团、沃支付、哈啰出行、元气森林、金龙鱼、旺旺、农夫山泉、格力等；国际企业如德国博世集团、荷兰飞利浦集团等；国际金融机构如孟加拉国 bKash、泰国 TrueMoney、菲律宾 GCash 等），提供了金融安全保障，产生了超过 10 亿元销售与服务经济效益。

四、客观评价

1. 对本项目技术水平与成果的验收意见及第三方评测结论

(1) 2022 年 9 月 7 日, OPPO 广东移动通信有限公司通过对西安交通大学的技术委托开发项目《可信 AI 技术研究》的验收工作。公司对该研究成果评价:“有效地支撑了我公司相关人工智能产品的安全可信研发、部署及应用”。(附件 D-1)

(2) 2024 年 4 月 24 日, 科技部高技术中心在北京组织了新一代人工智能国家科技重大专项“人工智能安全理论及验证平台”项目综合绩效评价。专家组评价:“构建了一体化自主安全防御体系, 研制了人工智能安全验证平台, 支持百余种攻防算法、多种开发框架、多维立体安全评估”, “具有较高的实用价值和广阔的应用前景, 已在多家企业开展示范性落地应用, 为人工智能安全发展做出了重要贡献”。(附件 D-2)

(3) 2024 年 4 月 26 日, 人民网股份有限公司在北京市组织专家对国家重点研发计划“基于可信与共治的全媒体内容社会众创服务平台研发与运营示范”项目所属子课题进行了绩效评价。专家组评价:“为全媒体内容社会众创服务平台及其相关业务系统的研建, 提供了基础支撑”。(附件 D-3)

(4) 2024 年 7 月 5 日, 工业和信息化部办公厅关于印发人工智能产业创新任务揭榜挂帅优胜单位名单的通知指出:“经过两年揭榜挂帅攻关培育, 遴选出 91 家揭榜优胜单位”, 西安交通大学承研的“面向金融场景的人工智能算法安全测评与防护平台”作为胜选单位之一, “鼓励其进一步加强人工智能关键核心技术攻关, 打造自主可控的人工智能技术体系”, “推动人工智能赋能应用”。(附件 D-4)

(5) 2024 年 7 月 31 日, 陕西省科技厅组织有关专家对西安交通大学承担的陕西省重点研发计划项目“面向大规模多模态检索的对抗样本关键技术与系统实现”进行了验收。验收专家组评价:“项目开展了……对抗样本关键技术与系统实现, 研究了……高效对抗样本的生成算法与检测算法, 研究并分析了……3 种模态的特征表示机理方法, 并研发了……对抗样本检测系统”。(附件 D-5)

2. 对本项目的自主知识产权与技术发明创新性的评价

(1) 在人工智能模型后门检测方面: 针对机器学习后门检测, IEEE Fellow、IEEE Commun. Surv. Tutorials 期刊主编、南洋理工大学 Dusit Niyato 教授指出项目组提出的方法可以有效阻止恶意后门节点参与到机器学习训练过程中。针对智能系统运行的安全性和稳定性, IEEE Fellow、IEEE Trans. Ind. Electron. 期刊共同主编、加拿大维多利亚大学 Yang Shi 教授指出项目组提出的性能指标的优化问题在智能系统后门检测有重要作用。针对模型鲁棒性, IEEE Fellow、意大利锡耶纳大学 Mauro Barni 教授将项目组的成果作为该领域的代表性成果进行介绍。(附件 E-1~E-3)

(2) 在数字世界模型漏洞检测方面: 针对人工智能系统对抗性安全问题, 加拿大工程院与研究院院士、ASME/CSME Fellow、加拿大达尔豪斯大学 Ya-jun Pan 教授将项目组在非线性信息物理系统方面的成果作为多模态混合随机系统的代表性工作方法进行介绍。针对数字智能系统运行的安全性, 韩国科学与技术研究院院士、韩国岭南大学首席教授 Ju H. Park 指出项目组关于人机智能系统容错控制设计的工作是非常有必要的。针对人工智能系统联邦学习安全, IEEE Fellow、南加州大学 Salman

Avestimehr 教授将项目组成果作为该领域的代表性成果进行介绍并说明该成果解决领域内挑战的意义。针对智能系统安全性，IEEE Fellow、香港科技大学 Shing-Chi Cheung 教授将项目组相关成果作为代表进行介绍并进行实验对比。(附件 E-4~E-7)

(3) 在物理世界对抗攻击漏洞检测方面：针对信物融合的鲁棒性，墨西哥科学院院士、Journal of The Franklin Institute 期刊共同主编、墨西哥新莱昂州自治大学 Michael V. Basin 教授引述项目组在信物融合鲁棒机理分析方面的成果，认为时延、执行器故障和非线性动态等因素不容忽视。针对智能物理系统的稳定性，欧洲科学院院士、IEEE Fellow、东南大学 Jinde Cao 教授将项目组在信物融合稳定性机理分析方面的成果进行引述和比较。针对图像处理智能系统的安全性，IEEE Fellow、意大利锡耶纳大学 Mauro Barni 教授将项目组在人工智能模型输入缩放攻击的成果作为该领域的代表性成果进行介绍。(附件 E-8~E-10)

(4) 在深度伪造生成式内容检测方面：针对生成式内容安全域计算机视觉应用的安全性，IEEE Fellow、IEEE TIFS 期刊主编 Mauro Conti 教授指出项目组提出的图像缩放攻击会对计算机视觉应用构成影响。针对语音识别系统的安全性，IEEE Fellow、ACM 杰出科学家、腾讯 AI Lab 副主任 Dong Yu 表示受项目组在语音对抗样本检测成果的启发，提出了一种防御语音对抗攻击的方法。(附件 E-11、E-12)

3. 对本项目应用效果的评价

(1) 蚂蚁胜信(上海)信息技术有限公司应用证明指出：“广泛应用于蚂蚁集团保险业务……与全国超过 90 家保险机构合作共同服务超 6 亿保民，有效地对保险风控相关风险进行精确识别与防控”，“为蚂蚁保险带来了近 1 亿的保费增长，为合作机构带来了 2.7 亿的赔付减损以及亿级的利润增长”。(附件 F-1)

(2) 华为技术有限公司应用证明指出：“面向华为人工智能相关业务产品的上线测试，自部署应用至今，共开展了对 1495 例智能系统及模型的检测与修复”，“对人工智能相关业务或系统的上线测试提供了有力支撑，效果显著”。(附件 F-2)

(3) 淘宝(中国)软件有限公司应用证明指出：“相关技术成果被集成于阿里云业务系统的安全保护引擎之中，提升了其在数据安全以及身份信息确认方面的保护能力”，“广泛应用于智慧交通、城市大脑等众多其他信息系统的安全防护环节”，“带动了相关业务场景的经济社会效益的有效增长”。(附件 F-3)

(4) 国家医疗保障局规财法规司应用证明指出：“坚持安全至上，助力守护百姓‘钱袋子’”，“构建健全的风险防控体系，确保医保用户资金使用安全”。(附件 F-4)

(5) 公安部防范电信网络诈骗信息监控中心感谢信指出：“积极参与精准宣防、预警反制、新技术运用和科技创新研究，取得了明显工作成效”。(附件 F-5)

(6) 网商银行应用证明指出：“全面支撑了我公司在小微企业风险管理中信用风险评估、供应链金融、伪造信息诈骗等相关业务”，“有效降低了风控安全隐患，增强了整体风险防范能力”，“带来新增授信超过 915 亿元”。(附件 F-6)

(7) 支付宝(中国)网络技术有限公司应用证明指出：“有效地识别伪造信息，精准定位黑灰产商家，为支付宝大型运营活动保驾护航”，“自该技术应用以来，帮助了上万中小商家纾困解难，累计为支付宝商家节省运营成本过亿元”。(附件 F-7)

五、应用情况和效益

1. 应用情况

在核心技术自主发明创新的基础上，本项目研制出具有自主知识产权、从陕西省发展至全国的“多环节人工智能模型后门检测系统”、“多维度数字世界模型漏洞检测系统”、“多类型物理世界对抗攻击漏洞检测系统”、“多模态生成式内容检测系统”等系列产品，已在蚂蚁胜信（上海）信息技术有限公司、网商银行、支付宝网络技术有限公司、淘宝（中国）软件有限公司、华为技术有限公司、OPPO 广东移动通信有限公司等单位产业化推广应用，服务于金融机构、科技公司、民营企业等多家单位，授权国家发明专利 59 项，发布国际/国家/行业标准 36 项，出版专著/章节 5 部，发表论文 86 篇，获得最佳论文奖 9 项，销售与服务收入超 10 亿元，经济和社会效益显著。

研制出金融人工智能算法安全测评与防护平台，集成了后门检测、数字世界攻击防御、物理世界攻击防御、深度伪造音视频检测、智能算法可解释性分析等模块，应用于蚂蚁集团的智能金融风险分析平台、支付宝 AI 平台、保险风控平台等，实现了产业化推广应用。

相关成果及应用服务于欧美、日韩、东南亚等 25 个国家（地区）以及我国 34 个省市自治区共计 12 亿用户（我国 8 亿用户），为中信银行、网商银行、中国人民保险集团股份有限公司、TrueMoney（泰国）、GCash（菲律宾）等全球 20 家金融保险机构，华为、荣耀、vivo、博世（德国）、飞利浦（荷兰）等全球科技公司，立白集团、金龙鱼、旺旺、农夫山泉、格力等民营消费企业，以及 1000 万多家中小微企业的超过 2 万亿资产，提供了金融安全保障（如图 6 所示，相关列表见链接：

<https://bksupplychain-assets.mybank.cn/>、<https://www.mayibao.cn/>），产生了超过 10 亿元（涵盖新增销售额以及降低赔付/运营成本）销售与服务经济效益。

金融机构	 中信银行 CHINA CITIC BANK	 网商银行	 PICC 中国人民保险	 true money wallet	 GCash
科技公司	 HUAWEI	HONOR	vivo	BOSCH	PHILIPS 飞利浦
消费品企业	 liby 立白	 金龙鱼	 旺旺	 农夫山泉 NONGFU SPRING	 GREE 格力

图 1. 部分技术服务单位

主要推广应用情况如下：

（1）在人工智能模型后门检测领域。该技术应用于蚂蚁集团的智能金融风险分析平台，有效提升了平台在金融业务中的鲁棒性和可信度；同时应用于华为技术有限公司，自 2020 年 1 月起对其 1495 例智能系统及模型开展自动化检测与修复，问题检测准确率达 100%，修复成功率为 97.33%，平均准确率提升 47.08%；此外还应用于国家医疗保障局医保码风控体系，截至 2023 年底保障超 5 亿激活用户、覆盖全国超 80 万家医疗机构的资金使用安全。

（2）在数字世界人工智能模型漏洞检测领域。相关技术应用于蚂蚁集团金融风险分析平台的模型漏洞与对抗攻击漏洞检测，增强了整体风险防范能力；为 OPPO 公

司可信人工智能安全等级评估提供核心支撑，保障其人工智能产品研发部署；同时支撑国家医疗保障局构建覆盖全国所有省份、超 80 万家医疗机构的医保码风控系统，显著提升医保服务效率和安全性。

（3）在物理世界金融对抗攻击漏洞检测领域。该技术应用于网商银行小微企业信用风险评估与供应链金融业务，为超过 1000 万家中小微企业提供风险保障；深度支撑 OPPO 公司实现人工智能全生命周期可信管理；同时应用于国家医保局全国各级医疗机构医保码线下结算场景，2023 年底前实现 34 个省市自治区全覆盖，保障超 5 亿用户就医购药安全。

（4）在金融多模态生成式内容检测领域。相关成果应用于支付宝拉新促活项目，实时检测伪造、篡改或合成的图像与语音内容，有效防止金融诈骗和身份盗用；为 OPPO 公司人工智能产品安全部署提供关键技术支撑；同时保障国家医保码系统全年超 5 亿用户、80 万家医疗机构的医保结算安全，显著降低欺诈风险。该平台已服务全球 25 个国家（地区）12 亿用户（含中国 8 亿用户），为 20 家全球金融保险机构、多家科技巨头及消费企业提供安全保障，保护超 2 万亿资产，产生超 10 亿元经济效益。

表 2. 主要应用单位情况表

序号	单位名称	应用的技术	应用对象及规模	应用起止时间	单位联系人/电话
1	蚂蚁胜信（上海）信息技术有限公司	主要技术发明 1, 2, 3, 4	应用于保险风控业务场景中健康险、意外险核保及理赔、运费险车险差异化定价、黑灰产数据识别等业务；与全国超过 90 家保险机构合作共同服务超 6 亿保民	2022. 01-2023. 12	秦小波
2	网商银行	主要技术发明 2, 3, 4	应用于网商贷、供应链金融多项金融服务中，服务规模超五千万小微企业客户	2022. 01-2023. 12	何慧梅
3	支付宝（中国）网络技术有限公司	主要技术发明 1, 2, 3, 4	应用于支付宝拉新促活项目，支撑超 100 亿资金规模的中小商家及生态运营工作	2018. 11-至今	顾进杰
4	OPPO 广东移动通信有限公司	主要技术发明 1, 2, 3, 4	应用于 OPPO 公司可信人工智能相关业务及人工智能相关产品	2022. 12-至今	杨明慧
5	华为技术有限公司	主要技术发明 1, 2, 3	应用于华为公司人工智能相关业务产品	2020. 12-至今	曹辉
6	国家医疗保障局法规司	主要技术发明 2, 3	应用于支付宝医保码的全场景、全流程中，累计激活人数超 5 亿，超过 80 万家医疗机构支持患者通过医保码完成结算服务	2019. 11-至今	刘瑞彤

2. 经济效益和社会效益

2.1 经济效益

在金融风控领域，项目中研发的金融人工智能算法安全测评与防护平台应用于蚂蚁集团的智能金融风险分析平台、支付宝 AI 平台、保险风控平台等，实现了产业化推广应用，服务于欧美、日韩、东南亚地区等 25 个国家（地区）以及我国 34 个省市自治区共计 12 亿用户（我国 8 亿用户），为全球 20 家金融机构企业（机构）、1000 万多家中小微企业的超过 2 万亿资产提供了金融安全保障。其中，该应用支撑了网商银行公司在小微企业风险管理中信用风险评估、供应链金融、伪造信息诈骗等相关业务，带来新增授信超过 915 亿元；在蚂蚁保险业务中的应用带来了近 1 亿的保费增长，为合作机构带来了 2.7 亿的赔付减损以及亿级的利润增长；在支付宝拉新促活项目中，该技术的应用帮助了上万中小商家纾困解难，累计为支付宝商家节省运营成本过亿元。

根据完成单位和应用单位提供的经济效益证明合计，共计产生了超过 10 亿元（涵盖新增销售额以及降低赔付/运营成本）销售与服务经济效益。

2.2 社会效益

本项目成果以人工智能安全技术为核心，涵盖了金融人工智能算法训练、测试、部署的全生命周期，保障了人工智能算法在反洗钱、反欺诈、保险风控、身份核验等金融场景下的可靠运行和规模化应用；实现了人工智能系统的安全自动增强，对人工智能相关业务或系统的上线测试提供了有力支撑；建立了可信 AI 安全等级的评估标准，推动了 AI 产品的安全可信研发、部署与应用；构建了一套健全的风险防控体系，为医疗等敏感领域提供了安全支持，显著提升了医保服务的覆盖和效率。

在促进行业进步方面，该项目研制出具有自主知识产权的金融人工智能算法安全测评与防护平台，实现了产业化推广应用，可以有效保障金融决策的公正性与安全性，提升平台在金融业务中的鲁棒性和可信度，增强智能金融风险分析平台的安全性，为企业评估模型在面对复杂企业数据时提供决策支持，增强企业整体风险防范能力。相关成果已经应用于欧美、日韩、东南亚等 25 个国家（地区）以及我国 34 个省市自治区共计 12 亿用户（我国 8 亿用户）。该项目还与全国超过 90 家保险机构合作，服务了超过 6 亿保民。

在金融方面以外，其他领域也能受益于该系统。项目中构建的一体化自主安全防御体系具有较高的实用价值和广阔的应用前景，已在多家企业开展示范性落地应用，为人工智能安全发展做出了重要贡献，相关成果被继承于阿里云等业务系统的安全保护引擎中，有效提升了其在数据安全以及身份信息确认方面的保护能力，同时被广泛应用于智慧交通、城市大脑等众多其他信息系统的安全防护环节，带动了相关业务场景的经济社会效益的有效增长。

在领域标准制定方面，该项目建立了可信 AI 安全等级的评估标准，为相关公司企业在可信人工智能等相关技术的研究提供了建设性的意见和启示，有效地支撑了相关人工智能产品的安全可信研发、部署及应用。项目中对人工智能系统框架问题缺陷的自动化检测、优化和修复技术已经应用于在华为等企业面向人工智能相关业务产品

的上线测试中，自部署应用至今已开展了上千例智能系统及模型的检测与修复，问题检测准确率为 100%，修复成功率为 97.33%，平均准确率提升为 47.08%，对人工智能相关业务或系统的上线测试提供了有力支撑，效果显著。

在带动相关产业发展方面，应用于支付宝医保码的 AI 算法安全测评与防护，累计激活人数超 5 亿，超过 80 万家医疗机构支持患者通过医保码结算，显著提升了医保服务的覆盖面和效率。此外，该系统还为基于可信与共治的全媒体内容社会众创服务平台及其相关服务业务系统的研建提供了基础支撑；在防范电信网络诈骗方面，本项目积极参与精准宣防、预警反制、新技术运用和科技创新研究，取得了明显工作成效。

六、主要知识产权证明目录（限 10 条）

序号	知识产权类别	知识产权具体名称	国家（地区）	授权号	授权日期	证书编号	权利人	发明人
1	发明专利	决策系统的公平性修复方法、系统、设备及存储介质	中国	ZL202111095511.2	2023-04-14	第5878779号	西安交通大学	沈超, 郜炫齐, 蔺琛皓, 王骞, 李琦
2	发明专利	机器学习框架漏洞检测方法、系统、设备及可读存储介质	中国	ZL202010996845.6	2023-04-07	第5860566号	西安交通大学	沈超, 张笑宇, 蔺琛皓, 管晓宏
3	发明专利	机器学习框架漏洞 API 参数定位方法、系统、设备及介质	中国	ZL202010997867.4	2023-03-21	第5801008号	西安交通大学	沈超, 张笑宇, 蔺琛皓, 管晓宏
4	发明专利	深度神经网络模型公平性测试方法、系统、设备及介质	中国	ZL202011403188.6	2022-04-22	第5097493号	西安交通大学	沈超, 降伟鹏, 蔺琛皓, 王骞, 李琦
5	发明专利	深度伪造检测模型的评测方法、系统及设备	中国	ZL202110963494.3	2023-09-19	第6337678号	西安交通大学	蔺琛皓, 邓静怡, 沈超, 胡鹏斌, 王骞, 李琦
6	发明专利	深度学习模型对抗样本生成方法、系统、设备及存储介质	中国	ZL202110049467.5	2023-05-02	第5933353号	西安交通大学	蔺琛皓, 朱炯历, 沈超, 管晓宏
7	发明专利	一种提供完整性验证的可审计外包机器学习服务方法	中国	ZL202011439129.4	2023-03-24	第5809711号	武汉大学	王骞, 田楚, 赵令辰, 王聪, 李琦, 沈超
8	发明专利	基于模型剪枝和逆向工程的深度学习后门防御方法	中国	ZL202110386155.3	2022-08-30	第5414633号	武汉大学	王骞, 龚雪鸾, 孔维翰, 王子瑶
9	发明专利	对抗样本的生成方法和装置	中国	ZL202010725498.3	2021-2-23	第4263851号	支付宝（杭州）信息技术有限公司	傅驰林, 黄启印, 周俊, 张晓露
10	发明专利	PRIVACY PROTECTION BASED TRAINING SAMPLE GENERATION METH	美国	US 10,878,125B2	2020-12-29	US10878125B2	支付宝（杭州）信息技术有限公司	王力, 赵沛霖, 周俊, 李小龙

		OD AND DEVICE					司	
--	--	---------------	--	--	--	--	---	--

七、主要完成人情况表

姓 名	沈超	排 名	1
行政职务	高层次人才办公室副主任		
技术职称	教授		
工作单位	西安交通大学		
完成单位	西安交通大学		
对本项目主要学术贡献： 提出金融人工智能算法安全测评与防护平台的整体规划方案与技术路线；分析提炼金融人工智能算法安全测评与防护平台的平台架构、关键技术、关键业务；指导系统整体集成与测试；和企业合作建立创新性的产学研合作关系，开展成果的推广应用。对主要技术发明 1、2、3、4 均有主要贡献。			

姓 名	蔺琛皓	排 名	2
行政职务	无		
技术职称	教授		
工作单位	西安交通大学		
完成单位	西安交通大学		
对本项目主要学术贡献： 具体制定金融人工智能安全检测与服务关键技术整体规划方案与技术路线；发明了数字世界模型漏洞检测物理世界对抗攻击漏洞检测方法，发明了多模态生成式内容检测方法；作为校企合作联系人，直接参与了金融人工智能安全检测与服务关键技术的研发工作。对主要技术发明 2、3、4 有直接贡献。			

姓 名	周俊	排 名	3
行政职务	总经理		
技术职称	高级工程师		
工作单位	支付宝（杭州）信息技术有限公司		
完成单位	支付宝（杭州）信息技术有限公司		

对本项目主要学术贡献：

具体制定金融人工智能安全检测与服务关键技术整体规划方案与技术路线；发明了跨量化位宽的黑盒对抗攻击方法，发明了基于代理模型微调的量化金融模型漏洞挖掘方法与基于困难样本挖掘的金融模型防御增强方法；参与了建立金融人工智能算法安全测评与防护平台。对主要技术发明 2 有直接贡献。

姓 名	王 骞	排 名	4
行政职务	执行院长		
技术职称	教授		
工作单位	武汉大学		
完成单位	武汉大学		
对本项目主要学术贡献： 具体制定金融人工智能安全检测与服务关键技术整体规划方案与技术路线；发明了基于三维对抗伪装的攻击检测方法，发明了基于延迟优化的攻击检测框架，发明了金融场景的风险管理与决策支持系统的攻击方法；参与研制金融场景下多类型物理世界对抗攻击漏洞检测系统。对主要技术发明 3 有直接贡献。			

姓 名	李前	排 名	5
行政职务	无		
技术职称	副教授		
工作单位	西安交通大学		
完成单位	西安交通大学		
对本项目主要学术贡献： 具体实施金融人工智能安全检测与服务关键技术的研发工作；发明了基于局部差异分析生成式图像检测方法，参与发明了基于语义弱化与关键帧抽取的生成式视频检测方法；协助合作企业技术人员完成相应技术在金融人工智能算法安全测评与防护平台的实现；对主要技术发明 4 有直接贡献。			

姓 名	张志强	排 名	6
行政职务	技术总监		

技术职称	无
工作单位	支付宝（杭州）信息技术有限公司
完成单位	支付宝（杭州）信息技术有限公司
<p>对本项目主要学术贡献：</p> <p>具体实施金融人工智能安全检测与服务关键技术的研发工作；发明了基于困难样本挖掘的金融模型防御增强方法与智能决策引擎技术，参与发明了基于预警防御的多维度数字世界模型漏洞检测方法；参与金融人工智能算法安全测评与防护平台的研制工作；负责相应技术在金融人工智能算法安全测评与防护平台的实现。对主要技术发明 2 有直接贡献。</p>	

八、主要完成单位情况表

单位名称	西安交通大学
<p>对本项目主要学术贡献：</p> <ol style="list-style-type: none"> 1. 与支付宝（杭州）信息技术有限公司、武汉大学共同研制出金融人工智能算法安全测评与防护平台，集成了后门检测、数字世界攻击防御、物理世界攻击防御、深度伪造音视频检测等模块。 2. 发明了基于“参数迭代+聚类轨迹”的多环节人工智能模型后门检测方法；发明了基于“漏洞挖掘+预警防御”的多维度数字世界模型漏洞检测方法；发明了基于“对抗伪装+延迟优化”的多类型物理世界对抗攻击漏洞检测方法；发明了“特征提取+真伪预测”的多模态生成式内容检测方法。构建了金融人工智能算法安全测评与防护平台。授权国家发明专利 29 项，发布国际/国家/行业/团体标准 10 项，发表论文 50 篇，出版专著/章节 3 部。 3. 与支付宝（杭州）信息技术有限公司、武汉大学建立产学研合作关系，应用于蚂蚁集团的智能金融风险分析平台、支付宝 AI 平台、保险风控平台等，实现了产业化推广应用，服务于欧美、日韩、东南亚等 25 个国家（地区）以及我国 34 个省市自治区共计 12 亿用户（我国 8 亿用户），为中信银行、中国人民保险集团股份有限公司等全球 20 家金融企业（机构）、1 千万家中小微企业的超过 2 万亿资产，提供了金融安全保障，产生了超过 10 亿元销售与服务经济效益。 	

单位名称	支付宝（杭州）信息技术有限公司
<p>对本项目主要学术贡献：</p> <ol style="list-style-type: none"> 1. 与西安交通大学共同提出了多模态数字世界模型漏洞检测技术的整体规划方案与技术路线。 2. 参与发明了基于敏感区域定位与多粒度组合测试的多模态数字世界黑盒攻击方法、基于关键神经元响应的可解释性特征挖掘方法、基于异常分布投影和随机替换编码的多模态对抗防御的技术发明与攻关工作，进行系统总体集成与测试，完成了金融场景下多模态数字世界模型漏洞检测技术的系统实现。授权国际/国家发明专利 14 项，发布国际/国家/行业/团体标准 26 项，发表论文 16 篇。 3. 与西安交通大学、武汉大学建立产学研合作关系，研究成果应用于蚂蚁集团的智能金融风险分析平台、支付宝 AI 平台、保险风控平台等，实现了产业化推广应用，服务于欧美、日韩、东南亚等 25 个国家（地区）以及我国 34 个省市自治区共计 12 亿用户，为超过 2 万亿资产提供了金融安全保障，产生了良好的经济效益和社会效益。 	

单位名称	武汉大学
<p>对本项目主要学术贡献：</p> <ol style="list-style-type: none"> 1. 与西安交通大学共同提出了多类型物理世界对抗攻击漏洞检测技术的整体规划方案与技术路线。 2. 参与发明了基于替代模型增强的语音对抗样本生成方法、基于局部平滑和基于降采样的语音对抗防御方法、基于扰动增强的图像对抗样本生成方法、基于特征压缩的图像对抗样本检测方法的技术发明与攻关工作，进行系统总体集成与测试，完成了金融场景下多类型物理世界对抗攻击漏洞检测技术的系统实现。授权国家发明专利 16 项，发表论文 26 篇，出版专著/章节 2 部。 3. 与西安交通大学、支付宝（杭州）信息技术有限公司建立产学研合作关系并开展了产业化推广应用，应用于包括用户身份核验、商户注册反欺诈等业务，创造了良好的经济效率和社会效益。 	

完成人合作关系说明

本项目完成单位为西安交通大学、支付宝（杭州）信息技术有限公司、武汉大学。三方共同组建了金融人工智能算法安全测评与防护平台研究团队，展开了长期紧密合作，取得了一系列研究成果。

本项目完成单位西安交通大学组建了以项目完成人沈超教授为带头人的研究团队。项目完成人沈超、蔺琛皓、李前均属于西安交通大学，以上完成人发挥学科与平台优势，共同承担关键技术的理论研究和技術发明，并配合合作单位实现核心技术对接、系统集成、测试与推广应用。在项目研究过程中，上述完成人联合承担项目、合作发表论文、联合申请专利与技术标准。

本项目完成单位支付宝（杭州）信息技术有限公司，是领先的数字支付平台和金融科技企业，始终与西安交通大学、武汉大学保持紧密合作关系。双方自项目启动即开展深入合作，支付宝公司组建了由周俊、张志强等技术负责人组成的研发与应用团队，共同开展了技术攻关、平台研制、产业应用与规模化推广，实现了多项创新技术的落地应用。

本项目完成单位武汉大学组建了以王骞教授为负责人的研究团队，重点参与物理世界对抗攻击漏洞检测技术的研发与系统实现工作。在上述项目合作过程中，西安交通大学、支付宝（杭州）信息技术有限公司与武汉大学三方共同开展关键技术研究、系统集成、应用示范与成果转化，建立了稳定的产学研合作机制，取得了显著的经济与社会效益。

完成人合作关系情况汇总表

序号	合作方式	合作者/ 项目排名	合作起始时间	合作完成时间	合作成果	证明材料
1	共同知识产权	沈超(1), 蔺琛皓(2), 王 骞(4)	2020 年 12 月	2022 年 4 月	发明专利-深度神经网络模型公平性测试方法、系统、设备及介质	A-1
2	共同知识产权	沈超(1), 蔺琛皓(2)	2022 年 9 月	2023 年 4 月	发明专利-机器学习框架漏洞检测方法、系统、设备及可读存储介质	A-2
3	共同知识产权	蔺琛皓(2), 沈超(1)	2021 年 1 月	2023 年 5 月	发明专利-深度学习模型对抗样本生成方法、系统、设备及存储介质	A-4
4	共同知识产权	沈超(1), 蔺琛皓(2)	2022 年 9 月	2023 年 3 月	发明专利-机器学习框架漏洞 API 参数定位方法、系统、设备及介质	A-6
5	共同立项	沈超(1), 蔺琛皓(2), 李前(5), 周俊(3), 张志强(6)	2024 年 1 月	2025 年 12 月	国家重点研发计划战略性新兴产业科技创新合作项目-面向金融场景的人工智能算法安全测评与隐私增强技术	B-1
6	共同立项	沈超(1), 蔺琛皓(2), 李前(5), 周俊(3), 王骞(4), 张志强(6)	2022 年 1 月	2023 年 12 月	工业和信息化部人工智能产业创新任务揭榜挂帅项目-面向金融场景的人工智能算法安全测评与防护平台	B-2