RealSafe

RealSafe链接: <https://www.baai.ac.cn/research/realsafe>



人工智能的广泛应用改善着人类的生产生活。然而，当前的人工智能算法主要依赖于数据驱动、缺乏可解释性，因此算法模型中的安全隐患也给全社会带来不容忽视的风险挑战。

攻击者可以利用抗样本攻击和模型后门攻击等技术，使算法模型产生错误的决策，可能导致人工智能系统被入侵、执行错误的指令，最终可能造成严重的安全后果。

产品功能:

1. 模块安全测评: 基于主流和独有的攻击算法在数字和物理世界中构造对抗样本进行模拟攻击测试，并给出被测AI模型的安全评分帮助评估系统风险，并提供可量化的优化标尺；针对模型后门攻击，基于业界首款黑盒模型后门检测方法检查模型是否存在后门，避免被植入后门的模型应用于生产实践当中。
2. 防御解决方案: 在机器学习生命周期不同阶段、不同层面实施安全方案，避免出现疏漏。例如在运行阶段集成对抗样本检测、输入数据重构技术打造AI安全网关，防止对抗样本直接进入业务流程，从而避免模型出错；在模型训练阶段利用对抗训练、模型后门修复等方式提升模型自身的鲁棒性。

产品优势:

1. 国际领先的AI对抗技术: 集成多项国际领先的AI对抗攻防算法，荣获多项世界AI安全大赛冠军，部分已开源成果作为代表性方法被业内广泛引用。
2. 科学的可视化评估方法: 综合多强度多维度对抗攻击下AI模型的性能指标变化情况，输出AI模型的安全评分及 图表化报告，AI系统的安全性一目了然。
3. 开箱即用的简易操作: 零编码、低门槛，用户无需具备专业的模型安全算法知识和编程开发经验，全界面化操作。
4. 业内首创的黑盒测评能力: 上传封装后的黑盒模型文件即可开展测评，测评过程不暴露模型结构和参数等详细信息，严格保护被测评方知识产权。

概念介绍:

对抗样本:

对抗样本是指能够使机器学习模型出现错误的一类合成样本。攻击者通过在输入数据上增加扰动，就可以使模型出错。

如在图像分类场景中，在熊猫图片上添加少量的扰动，就可以生成使模型误判为长臂猿的对抗样本。

模型后门:

模型后门攻击指通过在训练数据中添加少量投毒数据的方式在模型中植入后门，植入好的后门通过攻击者预先设定的触发器激发。在后门未被激发时，被攻击的模型具有和正常模型类似的表现；当模型中植入的后门被攻击者激活时，模型给出错误的输出。

如在交通标志识别场景中，在标志牌上粘贴黄色方块即可触发事先植入在识别模型中的后门，使模型将停车标志误识别为减速。

测评所支持的具体攻击方法:

目前RealSafe 人工智能安全平台社区版支持迁移攻击对抗样本攻击测评。

迁移攻击方法指攻击者事先训练与攻击目标类似的“替代模型”，对替代模型进行白盒攻击生成具有迁移性的对抗样本。再利用这些对抗样本对目标模型进行攻击。

迁移攻击方法不需要多次查询，攻击成功率稍低，但隐蔽性很强。

模型安全测评:

准备工作:

在测评前，您需要做好以下准备：

准备待测评的服务

测评前您需要按照RealSafe社区版服务封装规范，将待测服务封装成规定的格式后在系统中创建测评任务时上传即可。

准备测评数据

测评时需使用数据集，为保证测评结果的准确性，建议您上传真实业务场景中的数据进行测评。若准备数据较为繁琐，您也可以选择系统内置数据集进行测评。

如果您需要创建数据集，可以按照以下步骤进行准备：

准备图片

待测图片时需要符合以下要求：

格式：图片格式仅支持 JPG、JPEG、PNG

文件大小：图片文件大小不超过10M

数量：至少提供100对图片，最多不超过500张图片

人脸比对场景特殊要求：至少包含100个不同的人，且每人至少需要含有两张图片；并且为保证测试结果的严谨性，请确保每张图片上仅含一个人脸

准备人脸比对场景的 CSV 文件

用于人脸比对场景的数据集您还需要额外准备 CSV 文件以标识图片与人的对应关系，CSV文件的格式要求如下：

CSV 文件结构

序号 图片名称（带后缀）

1 test1.jpg

1 test2.jpg

2 test3.jpg

2 test4.jpg

... ...

CSV 文件必须按照上述格式创建，每行对应您要上传的一个图片，且每行至少包含2列：

第一列为连续的整数字，相同的数字代表同一个人；

第二列为带后缀的图片名称；

创建 CSV 时常见的错误:

请注意避免创建 CSV 文件过程中可能发生的常见错误，如：

CSV 各行之间出现空行

非 CSV 格式的文件

CSV 文件内的图片名称与图片文件没有关联或不带后缀

代表同一个人的数据列在 CSV 中处于不相邻：如第一列数字1和2分别代表两个人。数据从上往下按 1、1、2、2、1排列不符合要求，需要排列成1、1、1、2、2

开始测评

测评方法:

按任务流程指引创建完成测评任务后，平台将会首先接入您的服务，再利用数据集生成对抗样本模拟攻击进行测评，测试完成后，平台还将提供安全评分和完整的测评报告。

安全评分及分级

测评过程将综合被测服务在不同算法、扰动大小的攻击下效果的变化进行评分，安全评分满分为100分，评分越高则安全性越高。安全等级依据评分高低进行划分：小于60分为风险较低，介于60-80分之间为风险较高、80分以上为风险极高。

常见问题:

任意服务均可以用于测评吗?

测评对象为模型文件

目前平台只支持测试 TensorFlow、PyTorch 深度学习框架训练出的模型，并且场景仅限于人脸比对。

测评对象为API

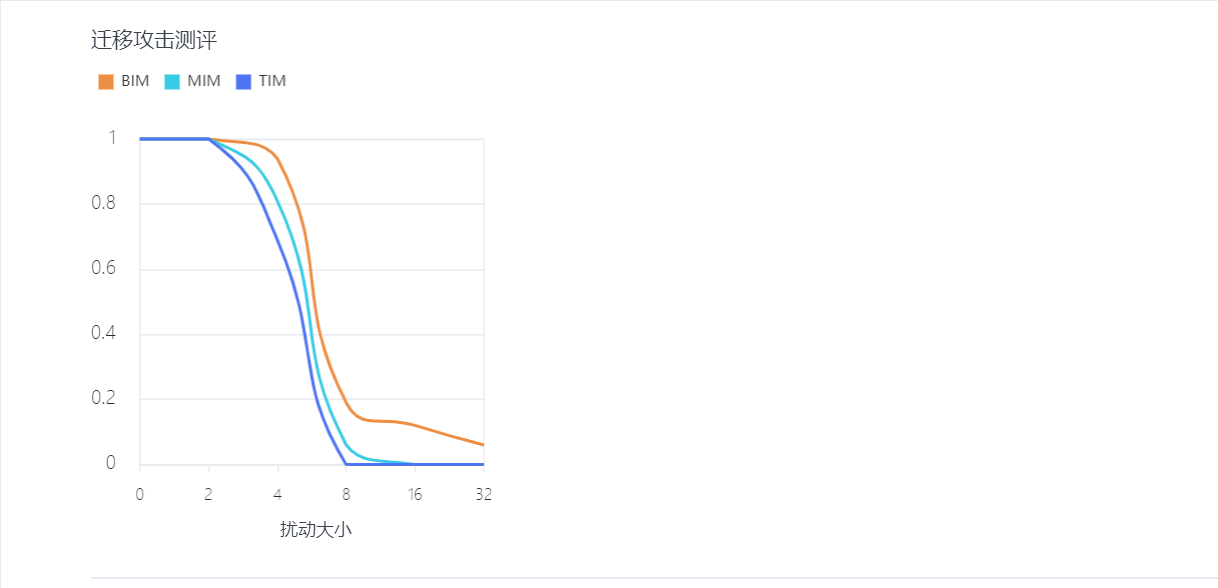
测评时将在约2小时内调用API服务不低于2000次，请保证您的服务能够满足以上要求，并且在测评过程中API服务需保持稳定，如出现异常情况如调用失败等将导致测评失败。

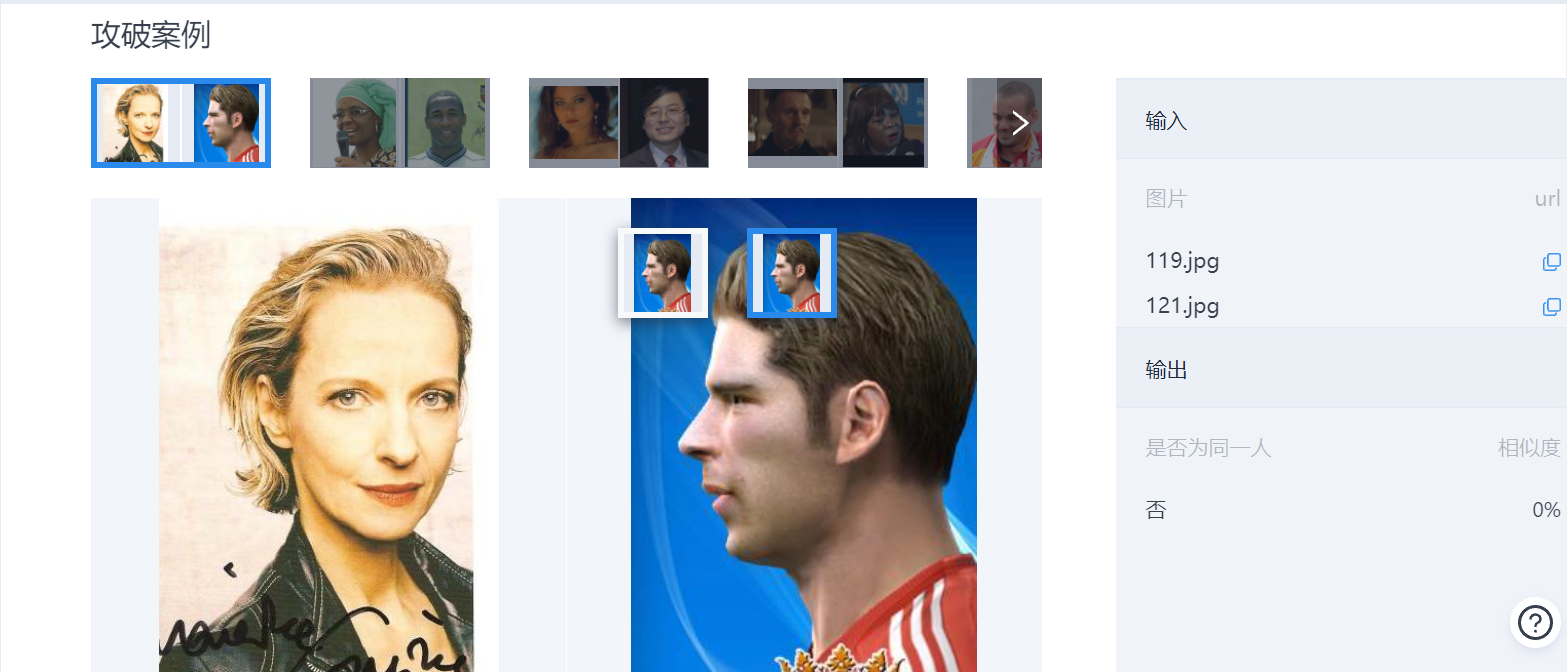
不准备数据可以测试吗?

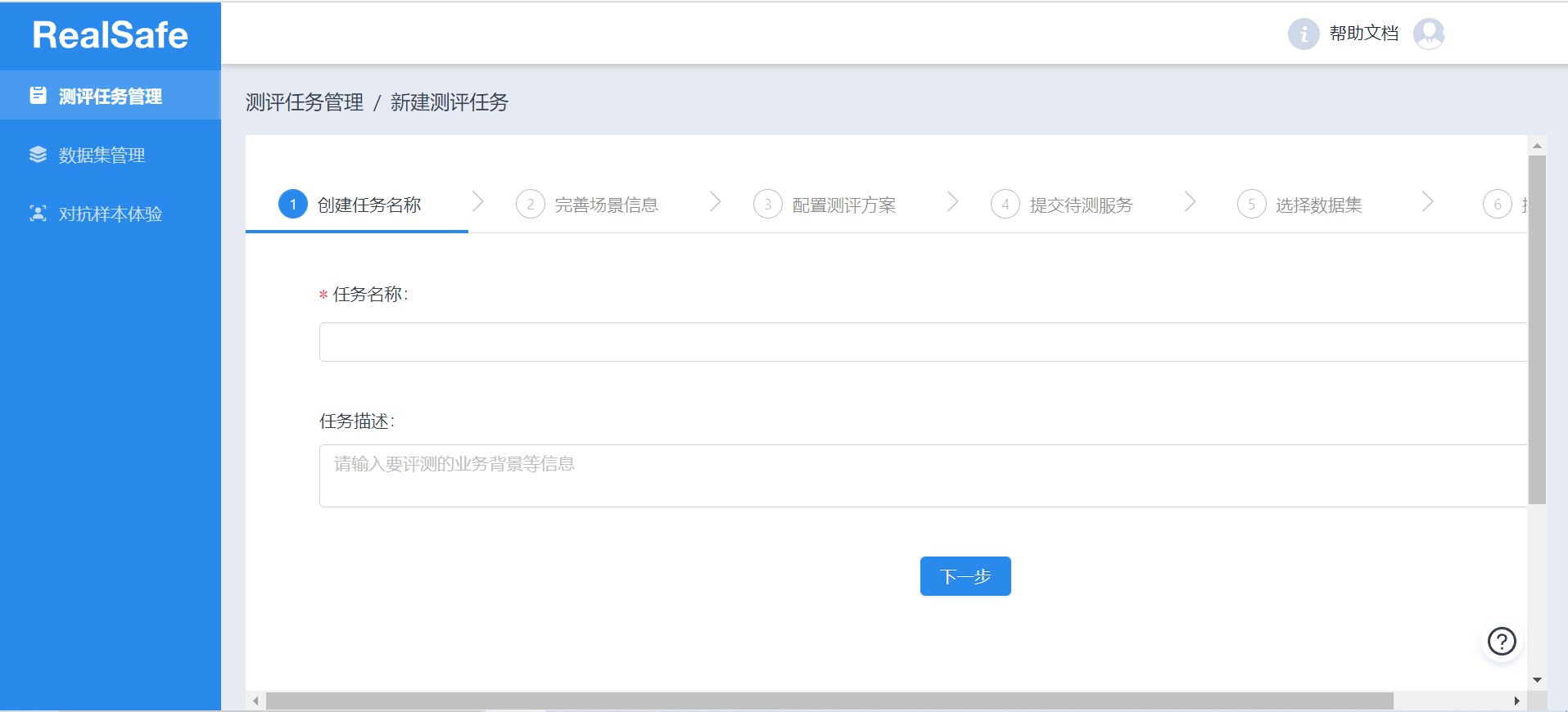
可以，您可以使用系统内置的数据集进行测评。但需要注意系统提供的数据集为通用数据集，可能无法与您的实际应用场景完全吻合。



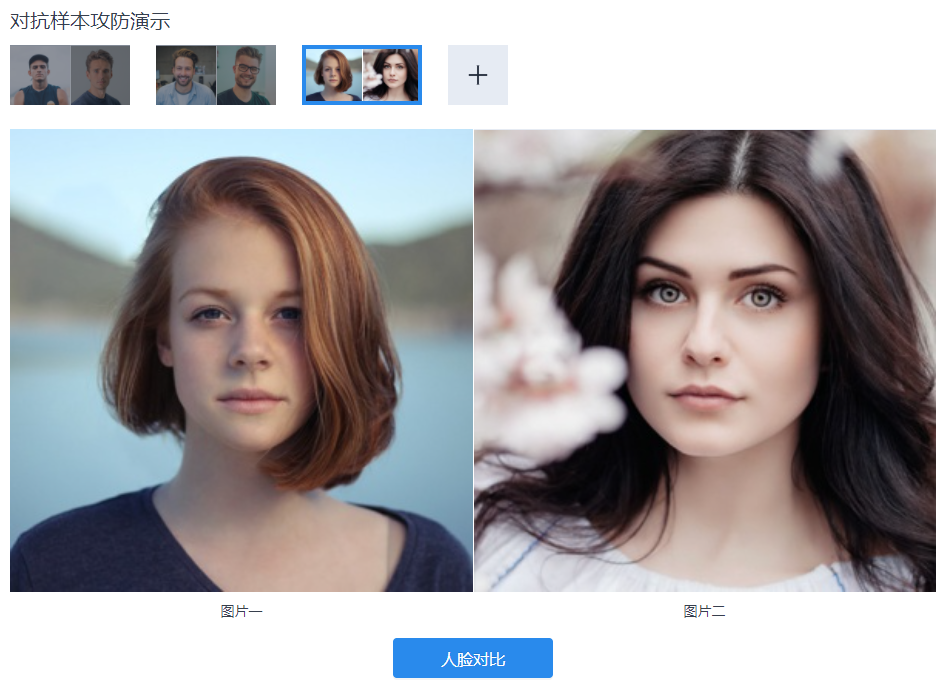


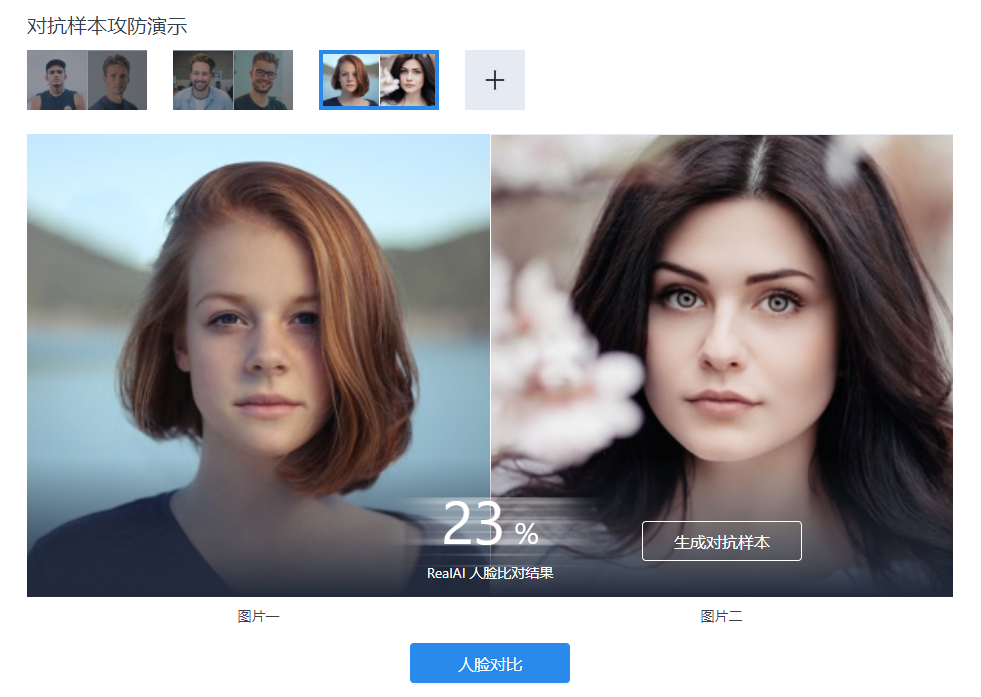


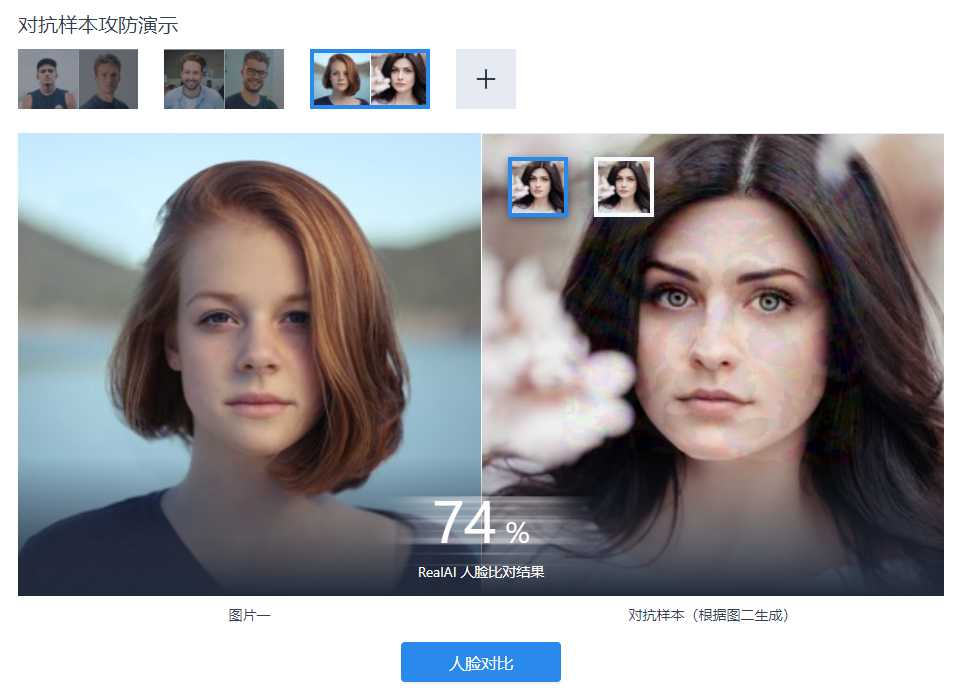












(参考

<https://baijiahao.baidu.com/s?id=1663308974057100039&wfr=spider&for=pc> )

一、从安全评测到防御升级，RealSafe让AI更加安全可控

为了解决以上痛点，近日，清华大学AI研究院孵化企业RealAI（瑞莱智慧）正式推出首个针对AI在极端和对抗环境下的算法安全性检测与加固的工具平台——RealSafe人工智能安全平台。

据了解，该平台内置领先的AI对抗攻防算法，提供从安全测评到防御加固整体解决方案，目前可用于发现包括人脸比对等在内的常用AI算法可能出错的极端情形，也能预防潜在的对抗攻击。

RealAI表示，就如网络安全时代，网络攻击的大规模渗透诞生出杀毒软件，发现计算机潜在病毒威胁，提供一键系统优化、清理垃圾跟漏洞修复等功能，RealSafe研发团队希望通过RealSafe平台打造出人工智能时代的“杀毒软件”，为构建人工智能系统防火墙提供支持，帮助企业有效应对人工智能时代下算法漏洞孕育出的“新型病毒”。

RealSafe平台目前主要支持两大功能模块：模型安全测评、防御解决方案。

其中，模型安全评测主要为用户提供AI模型安全性评测服务。用户只需接入所需测评模型的SDK或API接口，选择平台内置或者自行上传的数据集，平台将基于多种算法生成对抗样本模拟攻击，并综合在不同算法、迭代次数、扰动量大小的攻击下模型效果的变化，给出模型安全评分及详细的测评报告。目前已支持黑盒查询攻击方法与黑盒迁移攻击方法。

防御解决方案则是为用户提供模型安全性升级服务，目前RealSafe平台支持五种去除对抗噪声的通用防御方法，可实现对输入数据的自动去噪处理，破坏攻击者恶意添加的对抗噪声。根据上述的模型安全评测结果，用户可自行选择合适的防御方案，一键提升模型安全性。另外防御效果上，根据实测来看，部分第三方的人脸比对API通过使用RealSafe平台的防御方案加固后，安全性可提高40%以上。

随着模型攻击手段在不断复杂扩张的情况下，RealSafe平台还持续提供广泛且深入的AI防御手段，帮助用户获得实时且自动化的漏洞检测和修复能力。

二、“对抗样本”成“AI病毒”，国外主流人脸识别算法相继被“攻破”

站在人脸识别终端前，通过人脸识别摄像头完成身份校验，类似的人脸识别身份认证已经覆盖到刷脸支付、酒店入住登记、考试身份核验、人证比对等等生活场景中。

考虑到公众对于对抗样本这一概念可能比较模糊，RealSafe平台选取了公众最为熟知的人脸比对场景（人脸比对被广泛用于上述的身份认证场景中）提供在线体验。并且，为了深入研究“对抗样本”对人脸比对系统识别效果的影响，RealAI 团队基于此功能在国外主流 AI 平台的演示服务上进行了测试。

选取一组不同的人脸图片，通过RealSafe平台对其中一张图片生成对抗样本，但不影响肉眼判断，添加“对抗样本”前后分别输入到第三方人脸比对平台中查看相似度。

最终结果显示，添加“噪声”前，两张图片被 Azure、AWS 判定为不属于同一个人，但添加“噪声”后，以上两个平台的演示服务均给出了错误的结果，认为两张图片属于同一个人，甚至 Azure 平台的演示服务在添加“噪声”前后相似度变化的幅度高达70%以上。

为了探究结果的普适性，RealAI团队又选取了国内三家主流人脸比对平台进行测试，结果同样显示，添加扰动之后，原本判定为“不同人脸”的图片均被错误识别为“相同人脸”，前后相似度的变化幅度可达到20%以上。而通过RealSafe防火墙“去噪”过滤后，这几个人脸比对平台的识别“误差”获得不同程度的纠正，识别效果得到稳定提升。

RealAI团队已经将这种潜在风险以及相关防御方法反馈给上述企业，以帮助降低风险。

实测证明，“对抗样本”可以极大的干扰人脸比对系统的识别结果，据介绍，目前市面上很多中小型企业在落地人脸识别应用时大多会选择采用上文测试的这几家互联网公司开放的人脸比对SDK或者API接口，如果他们人脸比对技术存在明显的安全漏洞，意味着更广泛的应用场景将存在安全隐患。

除了人脸比对外，对抗样本攻击还可能出现在目标检测的应用场景中，延伸来看，这可能会危害到工业、安防等领域的安全风险检测。比如某电网的输电塔的监控系统，由于输电塔的高安全性防护要求，防止吊车、塔吊、烟火破坏输电线路，需要对输电塔内外进行全天候的实时监控，而这实时监控系统背后就是基于目标检测的AI算法来提供保障。

而RealAI研究团队发现，只要通过RealSafe对其中的目标检测算法进行一定的对抗样本攻击，就会造成监控系统失效，导致其无法识别非常明显的烟火情形，类似情形如果真实发生，将可能带来难以估计的损失。

事实上，像以上提到的这些AI安全风险由于都是AI底层算法存在技术缺陷而导致，往往比较隐蔽，但牵一发动全身，这些“难以预见”的风险漏洞最有可能成为被攻破的薄弱环节，而RealSafe平台同步推出的防御解决方案则可以有效增强各应用领域中AI算法的安全性。

三、“零编码”+“可量化”，两大优势高效应对算法威胁

据介绍，RealAI此次推出的算法模型安全检测平台，除了可以帮助企业高效应对算法威胁还具备以下两大优势：

组件化、零编码的在线测评：相较于ART、Foolbox等开源工具需要自行部署、编写代码，RealSafe平台采用组件化、零编码的功能设置，免去了重复造轮子的精力与时间消耗，用户只需提供相应的数据即可在线完成评估，极大降低了算法评测的技术难度，学习成本低，无需拥有专业算法能力也可以上手操作。比如上文中针对微软、亚马逊等第三方平台的测试，整个流程按照步骤提示完成，只需几分钟就可以查看到测评结果。

可视化、可量化的评测结果：为了帮助用户提高对模型安全性的概念，RealSafe平台采用可量化的形式对安全评测结果进行展示，根据模型在对抗样本攻击下的表现进行评分，评分越高则模型安全性越高。此外，RealSafe平台提供安全性变化展示，经过防御处理后的安全评分变化以及模型效果变化一目了然。

四、落地安全周边产品，为更多场景保驾护航

其实对抗样本原本是机器学习模型的一个有趣现象，但经过不断的升级演化，“对抗样本”已经演变成一种新型攻击手段，并从数字世界蔓延到物理世界：在路面上粘贴对抗样本贴纸模仿合并条带误导自动驾驶汽车拐进逆行车道、胸前张贴一张对抗样本贴纸在监控设备下实现隐身……

所以，除了针对数字世界的算法模型推出安全评测平台，RealAI团队也联合清华大学AI研究院围绕多年来积累的领先世界的研究成果落地了一系列AI攻防安全产品，旨在满足更多场景的AI安全需求。

比如，攻击技术方面，RealAI团队实现了世界首个通过“对抗样本”技术实现破解商用手机刷脸解锁，让手机将佩戴“特制眼镜”的黑客误识为机主。

通过在目标人服装上张贴特制花纹使AI监控无法检测到该人物，实现“隐身”，以及通过在车辆上涂装特殊花纹，躲避AI对车辆的检测。

在发现以上各种新型漏洞的同时，RealAI也推出相应的防御技术，支持对主流AI算法中的安全漏洞进行检测，并提供AI安全防火墙对攻击AI模型的行为进行有效拦截。

(参考

<https://baijiahao.baidu.com/s?id=1685597234302511367&wfr=spider&for=pc> )

据了解，RealSecure是业内首款隐私保护AI编译器，首创以底层数据流图的视角揭示机器学习算法与对应分布式隐私保护机器学习算法的联系，通过算子组合将机器学习生态与隐私保护机器学习生态一脉打通，解决企业搭建隐私保护生态面临着的性能差、易用性差、黑盒协议等诸多难题，实现两大生态的一体化。

与传统处理方式相比，RealSecure可实现训练速度提升30倍以上，白盒可验证的安全性以及便捷易用三大优势，实现数据源“高速公路”级别的节点链接，在支付、消费、税务、交通、司法、运营商等领域能实现大数据更快更安全的互联互通，助力反欺诈、小微普惠、AI医疗等智慧决策真正得以落地。

RealSafe则是世界首款企业级AI安全平台、业内首个针对人工智能模型的杀毒软件。该产品主要应用于人脸对比、目标检测、图像分类等场景，且具有算法领先、可量化、零编码等特点。RealSafe2.0版本在RealSafe1.0基础上，新增对目标检测、图像分类模型的支持，以及模型后门的检测能力，不仅拓宽了可评测的应用场景，实现了图像分类模型的后门自动化检测，还提供了包括样本去噪方法，对抗样本检测方法，对抗训练等解决方案。

目前，该产品已在工信部重大建设项目以及某电网公司落地应用，未来将集成更多安全风险检测能力，同时，瑞莱智慧RealAI 正积极探索无须人工干预、自动化训练的高安全性模型训练方案。

谈到关于本次两款新产品的开发理念，田天表示：“RealAI一直坚持推动底层技术变革，不同于行业中常态化的‘见一个解决一个’的问题处理办法，RealAI主张‘发现一个问题就看到一类问题’，并通过底层框架、平台和方法论的突破，助力产业的整体升级。本次新发布的两款产品就是这一理念的典型代表，无论是产品定位，还是功能价值都是RealAI首创。

同时，依托RealAI在构建模型安全测评防御及加固、内容安全检测、数据安全检测、伪造内容及模型安全攻防演练等方面的能力，RealAI与北京智源人工智能研究院联合设立了“安全人工智能创新中心”，并于本次会议上正式揭牌。该中心主要围绕基础数据、算法技术、应用产品三方面，落地人工智能安全性测评、防御以及伪造内容检测等解决方案，助力北京市切实发展安全、可控、高效的技术研究及管理体系。