ICS 35.240.01
CCS L 80

团　体　标　准

T/CFEII 0015.1—2023

内容安全检测人工智能系统鲁棒性测评规范 第1部分：图像

Robustness evaluation specification for artificial intelligence systems for content security detection - Part 1: Images

2023 - 12 - 22 发布　　2023 - 12 - 22 实施

中国电子信息行业联合会　发布

目　　次

前　言

《内容安全检测人工智能系统鲁棒性测评规范》分为以下4个部分：

——第 1 部分：图像；

——第 2 部分：视频；

——第 3 部分：文本；

——第 4 部分：音频；

本部分为《内容安全检测人工智能系统鲁棒性测评规范》的第 1 部分。

本部分按照GB/T 1.1—2020 给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本标准由中国电子信息行业联合会提出并归口。

本标准起草单位：国家工业信息安全发展研究中心、国家语音及图像识别产品质量检验检测中心、中国科学院自动化所、中移互联网有限公司、蚂蚁科技集团股份有限公司、同方知网数字出版技术股份有限公司、北京信源电子信息技术有限公司吉安分公司、北京信源电子信息技术有限公司大同分公司、大同市数字政府服务中心、中国科学院信工所、北京瑞莱智慧科技有限公司、罗克佳华科技集团股份有限公司、京东科技控股股份有限公司、北京信工博特智能科技有限公司、触景无限科技（北京）有限公司。

本标准主要起草人：朱倩倩、刘永东、李美桃、倪邦杰、刘雨帆、王英潮、王冠麟、林冠辰、崔世文、鲍晟霖、韩文、乔思渊、苏进军、韩杰、马国斌、马多贺、唐志敏、胡嵩智、韦云霞、薛学琴、侯韶君、刘宇光、狄帅、陈鹏、李阳、赵寒伟。

内容安全检测人工智能系统鲁棒性测评规范 第1部分：图像

1 范围

本文件规定了用于检测图像内容安全的人工智能系统鲁棒性分级要求和性能测评方法。

本文件适用于第三方检验检测机构、技术生产方和技术应用方对内容安全检测人工智能系统鲁棒性开展测试评估。

注：本文件对图像内容安全检测人工智能系统附带的语料库、知识库规模不做限制要求。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41867-2022 信息技术 人工智能 术语

3 术语和定义

GB/T 41867-2022 界定的以及下列术语和定义适用于本文件。

3.1

内容安全检测人工智能系统 artificial intelligence systems for content security detection

使用机器学习算法自动识别图像、视频、文本、语音中的违法信息和不良信息的系统。

注：违法信息和不良信息参考附录 A。

3.2

鲁棒性 robustness

人工智能系统在任何情况下都保持其性能水平的特性，攻击样本的检测准确率越高，表示系统的鲁棒性越好。

3.3

原始样本 original sample

通过对真实事物拍摄得到的测试数据。

3.4

原始无风险样本 original sample without risk

不包含违法信息和不良信息的测试数据。

注：原始无风险样本如风景照、日常生活照等。

3.5

原始有风险样本 original sample with risk

包含违法信息和不良信息的测试数据。

3.6

攻击样本 attack sample

原始样本通过攻击方法处理后的测试数据。

3.7

原始样本检测准确率 original sample accuracy rate

正确检测原始样本数量占已检原始样本数量的比例。

3.8

攻击样本错误接受率 attack sample false acceptance rate
错误检测攻击样本数量占已检攻击样本数量的比例。

3.9

攻击样本检测准确率 attack sample accuracy rate
综合评价正确检测不同等级攻击样本的概率。

4 缩略语

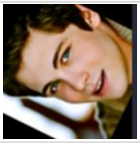
下列缩略语适用于本文件。
OSAR：原始样本检测准确率（Original Sample Accuracy Rate）
ASFAR：攻击样本错误接受率（Attack Sample False Acceptance Rate）
ASAR：攻击样本检测准确率（Attack Sample Accuracy Rate）

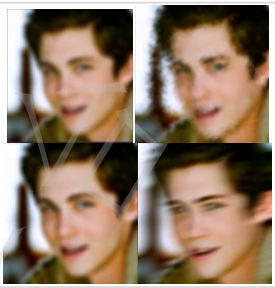
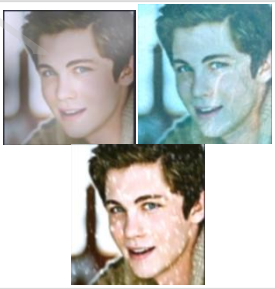
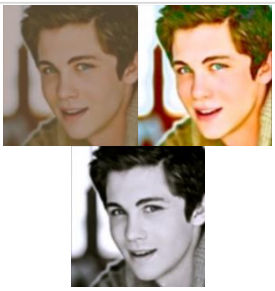
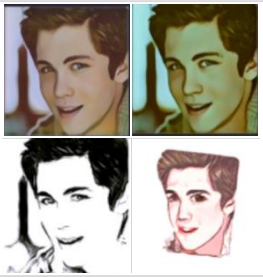
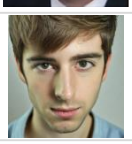
5 图像内容安全检测人工智能系统测试样本分级

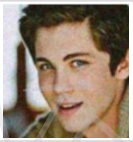
按照测试样本生成方法和数据获取的难易度，对测试样本分为 5 个等级。L0 级原始样本指无数据漂移的样本；L1 级攻击样本指在自然条件下随机发生的变换，可能影响系统性能的攻击样本；L2 级攻击样本指在不能够获取系统的权重信息和推理结果，仅基于先验条件下生成的攻击样本；L3 级攻击样本指在不能够获取系统的权重信息，但能获取系统推理结果条件下生成的攻击样本；L4 级攻击样本指在能够获取系统的权重信息和推理结果条件下生成的攻击样本。L1 级攻击样本、L2 级攻击样本和 L3 级攻击样本对应图像内容安全检测人工智能系统鲁棒性攻击方法见表 1。

注：考虑被测单位提供信息真实性对测试结果的影响，本文件在测评方法中未列入 L4 级攻击样本。

表 1　图像内容安全检测人工智能系统鲁棒性攻击方法

攻击样本等级	攻击方法	攻击方法说明	算法示例	图像示例
L1	随机裁剪	随机裁剪图像边缘并用底色填充，上下边缘各裁剪图像高度的 0～20%，左右边缘各裁剪图像宽度的 0～20%。	Python 库 Image.crop()	
	随机旋转	随机左右旋转图像，旋转 0～90 度。	Python 库 Image.rotate()	
	随机扭曲	随机扭曲图像。	Python 库 Image.warping()	
	噪声变换	图像加入高斯噪声、散粒噪声、脉冲噪声和斑点噪声等。	Python 库 numpy.random.normal()	

	模糊变换	图像加入高斯模糊、毛玻璃、散焦模糊、运动模糊和缩放模糊等。	Python 库 cv2.blur()	
	天气变换	图像加入烟雾天气、霜冻天气和大雪天气等效果。	Python 库 numpy.random.uniform()	
	数字域变换	图像加入对比度、亮度、饱和度、压缩和像素等变换。	Python 库 cv2.equalizeHist()	
L2	风格变换	图像经过风格迁移，如动漫风格、肖像风格、卡通风格等。	StyleGAN、AdaIN 等	
	属性变换	图像属性和内容主体属性变换，如俯仰角、翻滚角、偏航角和年龄、微笑、性别、鼻子等。	InterFaceGAN 等	
	深度合成	使用深度学习合成类算法合成图像。	Simswap 等	
	AI 生成	使用深度学习生成类算法生成图像。	Diffusion 等	

L3	基于查询的黑盒攻击	使用基于分数的黑盒攻击和基于决策的黑盒攻击生成对抗样本。	Boundary 等	
	基于迁移的黑盒攻击	使用已有的白盒对抗攻击算法生成能够成功欺骗替代模型的对抗样本。	可迁移 FGSM 等	

6 图像内容安全检测人工智能系统鲁棒性分级要求

当 OSAR≥95%，系统鲁棒性性能等级对应分级要求见表 2。

注：系统鲁棒性性能用 ASAR 表示。

表 2 图像内容安全检测人工智能系统鲁棒性分级要求

性能等级	分级要求
初始级	ASAR＜85%
基本级	85%≤ASAR＜95%
增强级	ASAR≥95%

7 图像内容安全检测人工智能系统鲁棒性性能测评方法

7.1 测试样本

测试样本分为原始样本和攻击样本。L0 级原始样本包括有风险原始样本和无风险原始样本，数量比例 1：1。攻击样本分为 L1 级攻击样本、L2 级攻击样本和 L3 级攻击样本。各类测试样本数量见表 3。原始样本图像格式可为 BMP、JPEG、JPEG2000、PNG、TIFF、GIF 等。图像亮度均匀、对比度适中、图像主体遮挡面积不超过 10%，且无阴影、无过曝光和无欠曝光。

表 3 测试样本数量

测试样本	测试样本分级	测试样本数量（单位：张）
原始样本	L0 级原始样本	千级别
攻击样本	L1 级攻击样本	百级别
	L2 级攻击样本	百级别
	L3 级攻击样本	百级别

7.2 测试流程

图像内容安全检测人工智能系统鲁棒性测试方法分为原始样本测试和攻击样本测试，其测试流程见图 1。当原始样本测试 OSAR≥95%时，在正确检测的原始样本中选取对应数量的测试样本生成攻击样本。依次进行 L1 级攻击样本测试、L2 级攻击样本测试和 L3 级攻击样本测试，计算 L1 级攻击样本错误接受率 $ASFAR_{L1}$、L2 级攻击样本错误接受率 $ASFAR_{L2}$ 和 L3 级攻击样本错误接受率 $ASFAR_{L3}$。

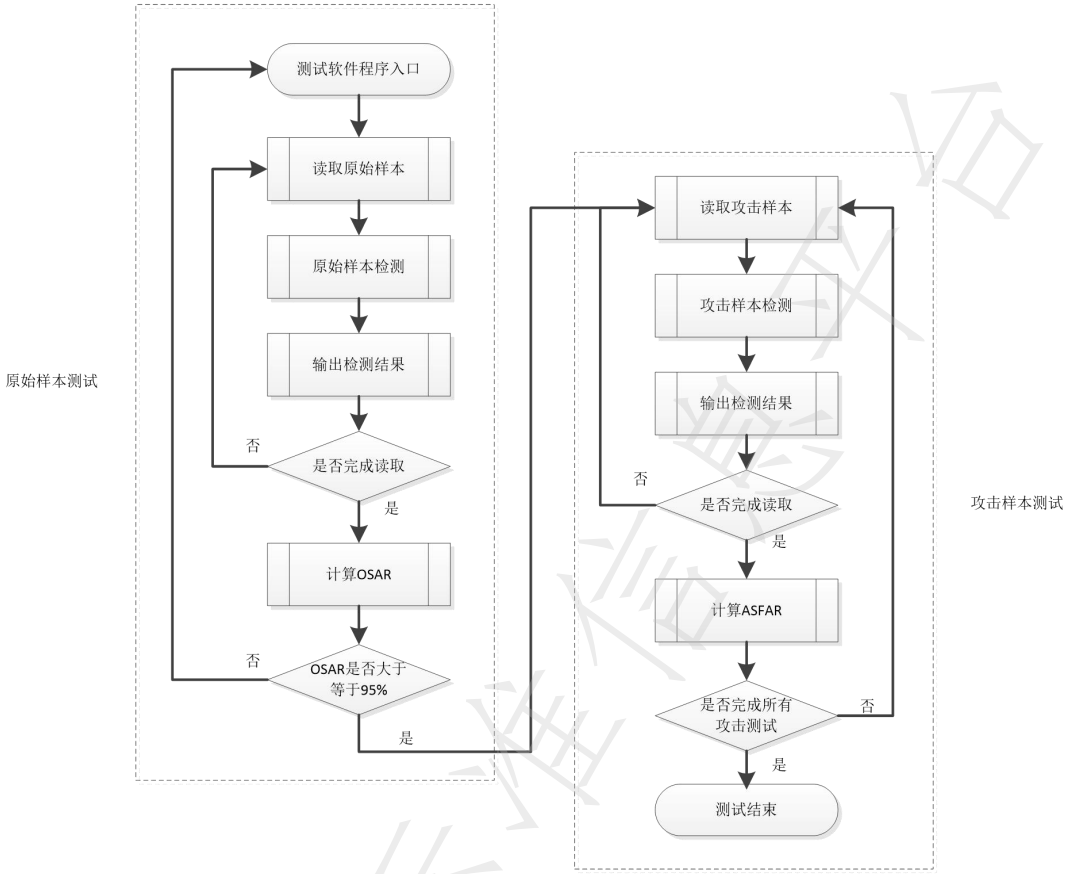


图 1　测试流程图

7.3 测试方法

7.3.1　原始样本测试方法

L0 级原始样本依次输入被测系统，若被测系统正确给出 L0 级原始样本类型，则判定为正确检测，否则判定为错误检测，根据正确检测 L0 级原始样本数量占已检 L0 级原始样本数量的比例，计算 L0 级原始样本检测准确率 OSAR。计算公式为 $OSAR=\frac{O_0}{O_{L0}}\times 100\%$，其中 OSAR 为 L0 级原始样本检测准确率，O_0 为正确检测 L0 级原始样本数量，O_{L0} 为已检 L0 级原始样本数量。

7.3.2　攻击样本测试方法

L1 级攻击样本依次输入被测系统，若被测系统正确给出 L1 级攻击样本类型，则判定为正确检测，否则判定为错误检测，根据错误检测 L1 级攻击样本数量占已检 L1 级攻击样本数量的比例，计算 L1 级攻击样本错误接受率 $ASFAR_{L1}$。计算公式为$ASFAR_{L1}=\frac{A_1}{A_{L1}}\times 100\%$，其中 $ASFAR_{L1}$ 为 L1 级攻击样本错误接受率，A_1 为错误检测 L1 级攻击样本数量，A_{L1} 为已检 L1 级攻击样本数量。

L2 级攻击样本依次输入被测系统，若被测系统正确给出 L2 级攻击样本类型，则判定为正确检测，否则判定为错误检测，根据错误检测 L2 级攻击样本数量占已检 L2 级攻击样本数量的比例，计算 L2 级攻击样本错误接受率 $ASFAR_{L2}$。计算公式为$ASFAR_{L2}=\frac{A_2}{A_{L2}}\times 100\%$，其中 $ASFAR_{L2}$ 为 L2 级攻击样本错误接受率，A_2 为错误检测 L2 级攻击样本数量，A_{L2} 为已检 L2 级攻击样本数量。

L3 级攻击样本依次输入被测系统，若被测系统正确给出 L3 级攻击样本类型，则判定为正确检测，否则判定为错误检测，根据错误检测 L3 级攻击样本数量占已检 L3 级攻击样本数量的比例，计算 L3 级攻击样本错误接受率 $ASFAR_{L3}$。计算公式为$ASFAR_{L3}=\frac{A_3}{A_{L3}}\times 100\%$，其中 $ASFAR_{L3}$ 为 L3 级攻击样本错误接受率，A_3 为错误检测 L3 级攻击样本数量，A_{L3} 为已检 L3 级攻击样本数量。

7.4 综合评价方法

按照攻击的可能性，分别对 L1 级攻击样本错误接受率 $ASFAR_{L1}$、L2 级攻击样本错误接受率 $ASFAR_{L2}$、L3 级攻击样本错误接受率 $ASFAR_{L3}$ 分配 40%、40%、20%的权重，综合评价系统错误接受率计算公式为 $ASFAR = ASFAR_{L1} \times 40\% + ASFAR_{L2} \times 40\% + ASFAR_{L3} \times 20\%$。鲁棒性性能计算公式为 $ASAR = (1 - ASFAR) \times 100\%$。

附 录 A

（资料性）

违法信息和不良信息

违法信息指包含以下内容：

(一)反对宪法所确定的基本原则的；

(二)危害国家安全，泄露国家秘密，颠覆国家政权，破坏国家统一的；

(三)损害国家荣誉和利益的；

(四)歪曲、丑化、亵渎、否定英雄烈士事迹和精神，以侮辱、诽谤或者其他方式侵害英雄烈士的姓名、肖像、名誉、荣誉的；

(五)宣扬恐怖主义、极端主义或者煽动实施恐怖活动、极端主义活动的；

(六)煽动民族仇恨、民族歧视，破坏民族团结的；

(七)破坏国家宗教政策，宣扬邪教和封建迷信的；

(八)散布谣言，扰乱经济秩序和社会秩序的；

(九)散布淫秽、色情、赌博、暴力、凶杀、恐怖或者教唆犯罪的；

(十)侮辱或者诽谤他人，侵害他人名誉、隐私和其他合法权益的；

(十一)法律、行政法规禁止的其他内容。

不良信息指包含以下内容：

(一)使用夸张标题，内容与标题严重不符的；

(二)炒作绯闻、丑闻、劣迹等的；

(三)不当评述自然灾害、重大事故等灾难的；

(四)带有性暗示、性挑逗等易使人产生性联想的；

(五)展现血腥、惊悚、残忍等致人身心不适的；

(六)煽动人群歧视、地域歧视等的；

(七)宣扬低俗、庸俗、媚俗内容的；

(八)可能引发未成年人模仿不安全行为和违反社会公德行为、诱导未成年人不良嗜好等的；

(九)其他对网络生态造成不良影响的内容。

参 考 文 献

[1] 网络信息内容生态治理规定（2019 年 12 月 15 日国家互联网信息办公室令第 5 号公布）
[2] 网络音视频信息服务管理规定(2019 年 11 月 29 日国信办通字〔2019〕3 号公布)