



# 政企应用中的人工智能 安全

吴鹤意

南京网络空间安全技术研究院副院长

## 目录

AI对抗样本简介

对抗样本安全案例

AI安全解决之道

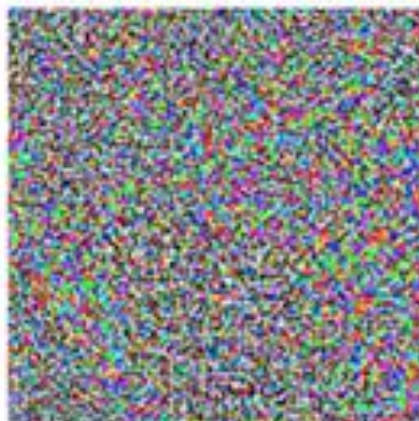


## Beginning (2014)



$x$   
“panda”  
57.7% confidence

$+ .007 \times$



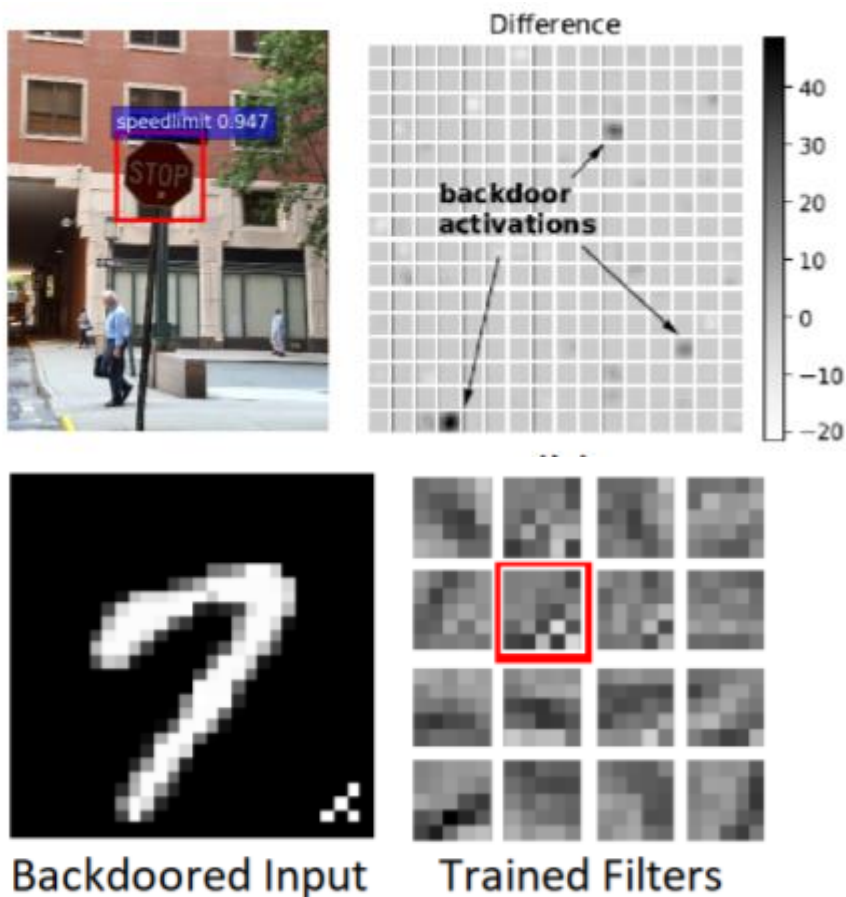
$\text{sign}(\nabla_x J(\theta, x, y))$   
“nematode”  
8.2% confidence

$=$

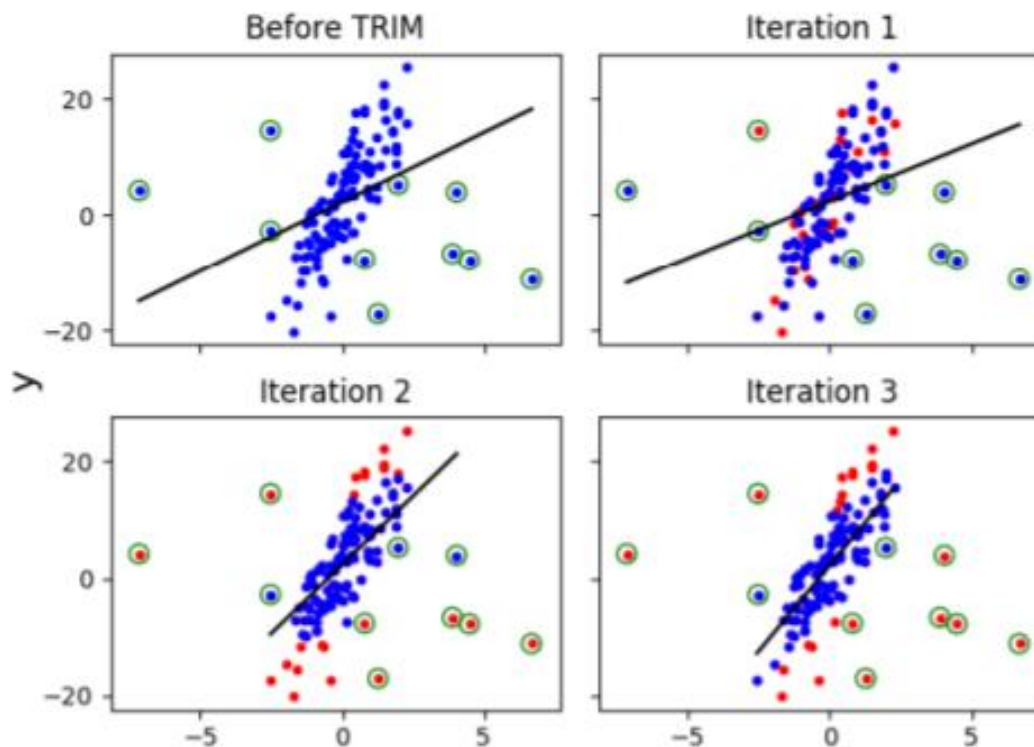


$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$   
“gibbon”  
99.3 % confidence

Now (2019)

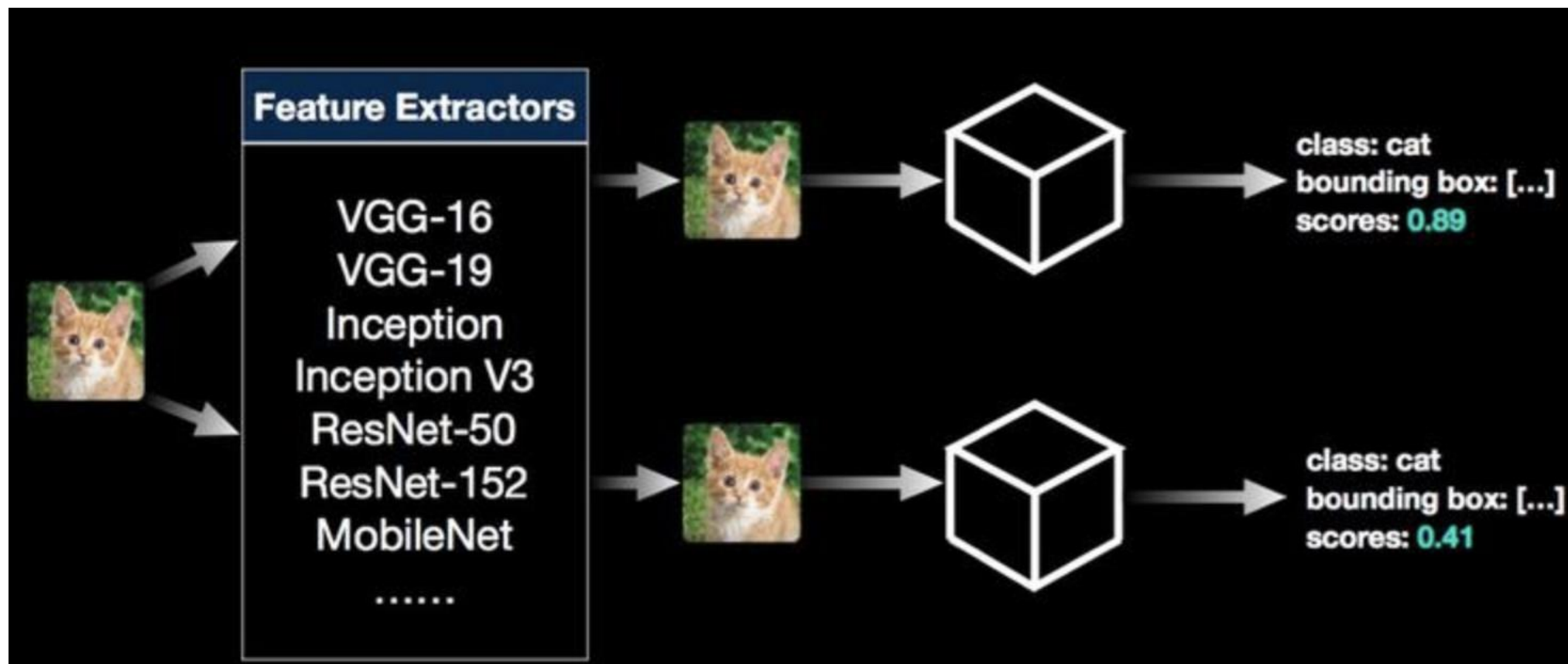


《 BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain 》



《 Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning 》

Now (2019)



百度《迁移攻击云端AI，一个被遗忘的战场》



## Now (2019)

	数据收集阶段	模型训练阶段	模型使用阶段
闪避攻击	对抗样本生成	网络蒸馏 对抗训练	对抗样本检测 输入重构 DNN模型验证
药饵攻击	训练数据过滤 回归分析	集成分析	
后门攻击		模型剪枝	输入预处理
窃取攻击	差分隐私	隐私聚合教师模型PATE 模型水印	

## 自助机人脸识别



## 政务AI机器人问答

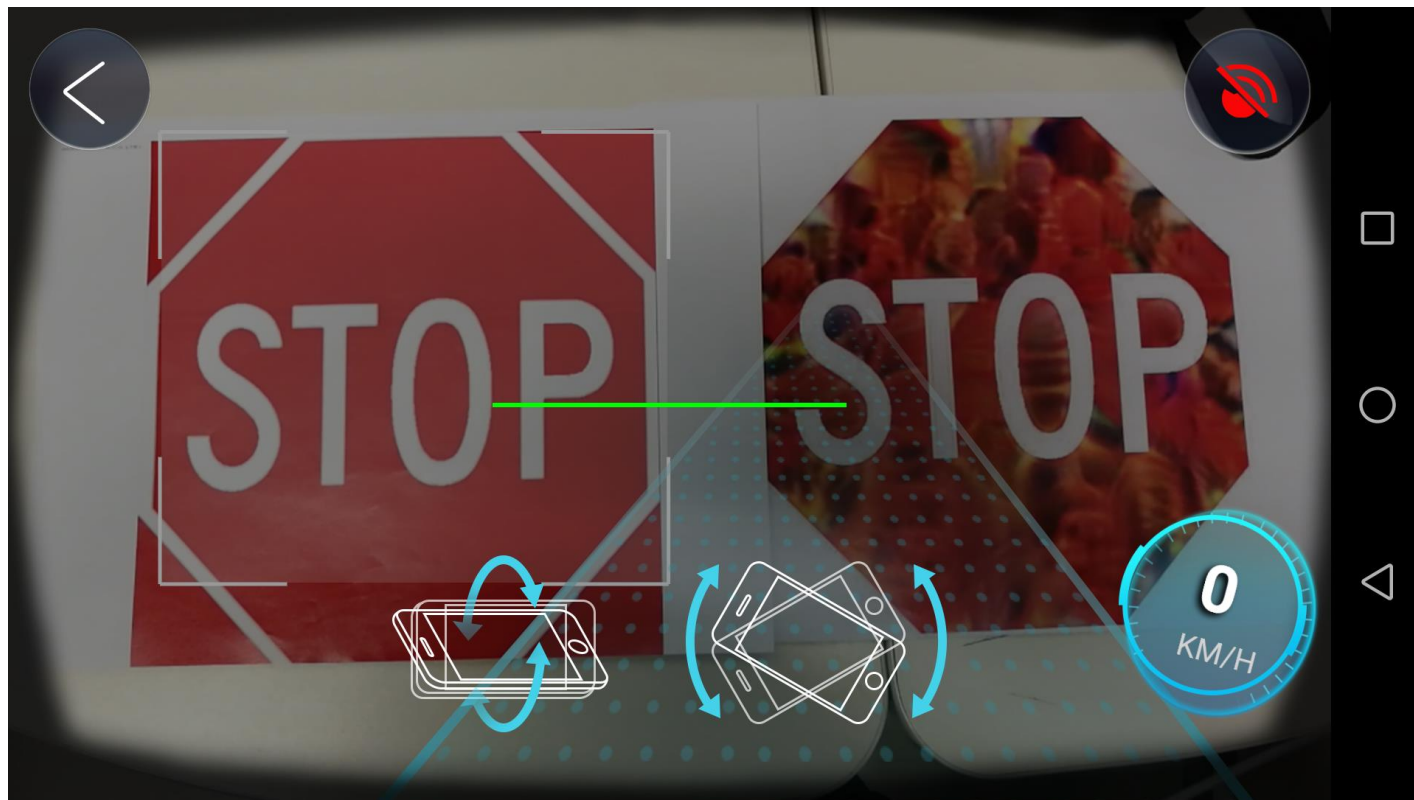




## 政务AI机器人问答



## 道路标志牌检测场景






## 人体检测场景

Demo by chenty use darknet yolo-tiny

关于



run time = 0.9318709373474121s

解压模型

分析

选择图片

10:46

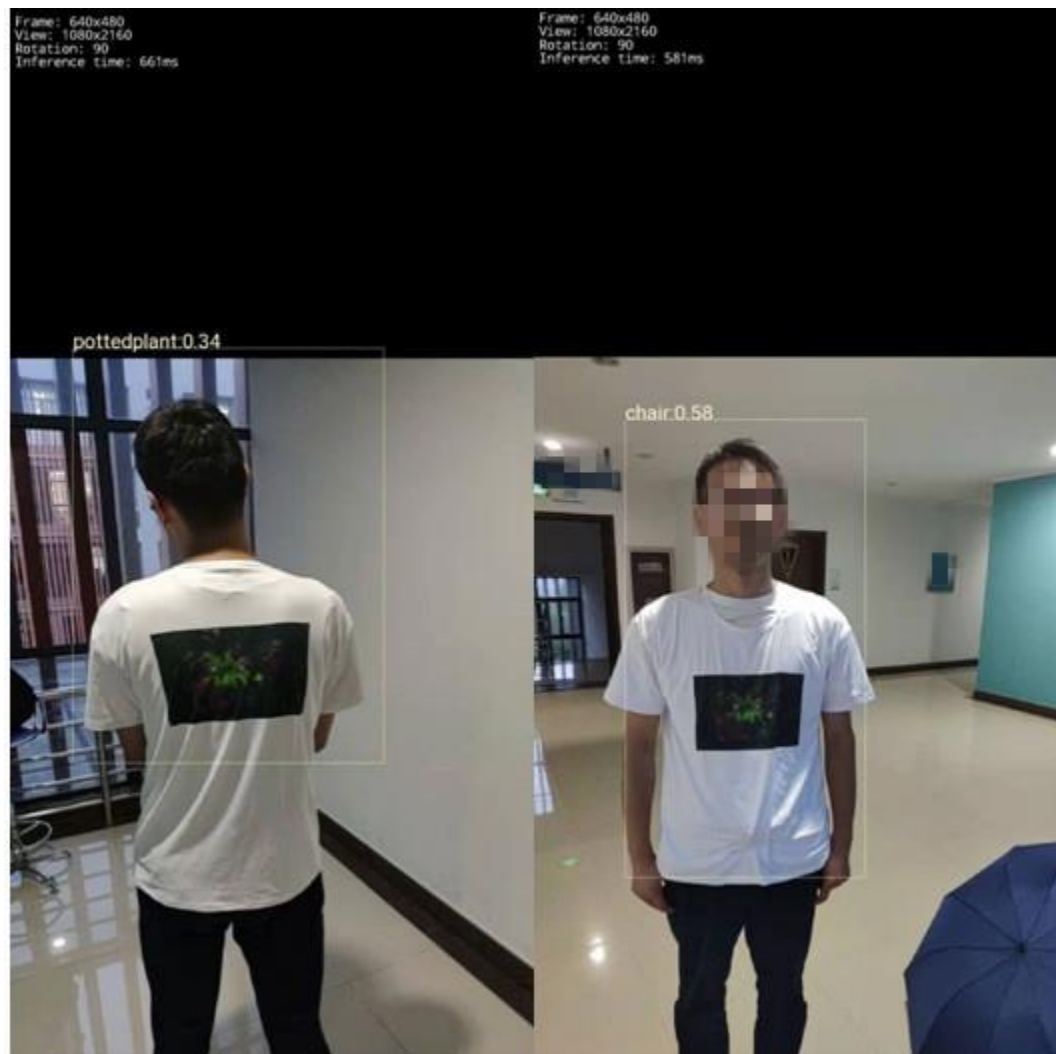
0.1K/s 4G+ 87



## 人体检测场景



## 人体检测场景



(1) AI与安全的结合，应该有更大的含义。

敏感字过滤，很多商用开放的AI平台上，还没有。安全中心也往往选择忽视。

(2) AI的接口，还是需要防御的。

点赞接口可以导致数据污染，知识库爬虫。

(3) 希望大家可以多交流。

AI骂脏话也会影响KPI，AI的落地交流，还不是很多。

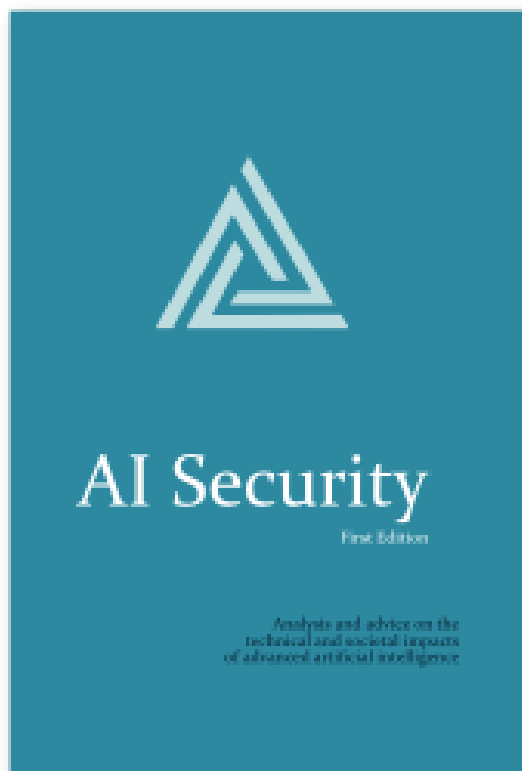
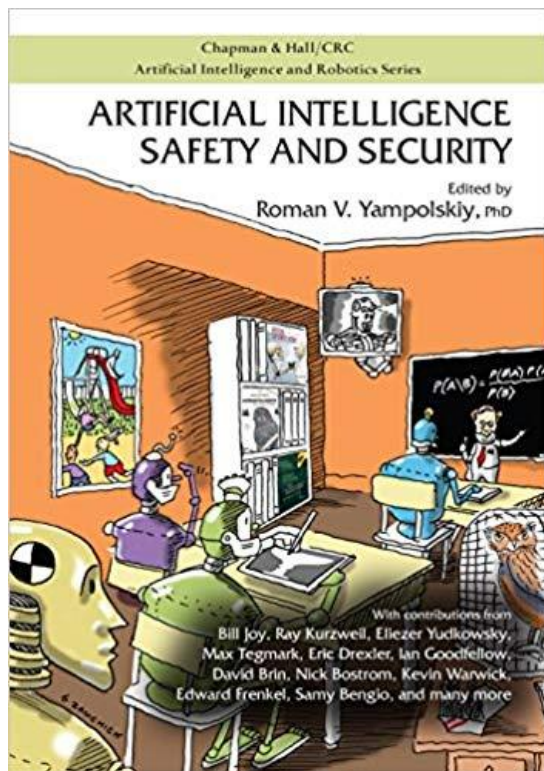
如今，距离 SIRI 发布已经七年了，它饱受诟病。根据分析师的早期推断，SIRI 是影响苹果最新产品性能的主要原因——一款售价 349 美元的 HOMEPOD 智能音箱。尽管这款音箱借其时髦设计和完美音质赢得了不少赞誉，但在测试之后，由于总是发出诸如“笨拙”、“烦人”和“尴尬”等词，这让苹果最终放弃了音箱的 SIRI 语音功能。

-- 《苹果和SIRI的七年之痒：SIRI的落寞之路》



找到并利用定量测量方法存在的漏洞，比按照研究者的期望努力达到预期结果要简单得多。


-- 《谷歌大脑工程师推荐：那些画风跑偏的AI之轶事合集》



AINOW







# THANKS

**2019 北京网络安全大会**  
2019 BEIJING CYBER SECURITY CONFERENCE