AI 及 AI 安全

6 班 陈静雯

一 AI 定义

AI ，“Artificial Intelligence”，即人工智能，是指能够感知环境并采取行动以最大程度地成功实现目标的任何设备。也称机器智能，是由机器实现的智能，与人类显示的自然智能相反。它可以从大量历史数据中挖掘出隐含规律，并用于预测或者分类。从数学的角度来说，一个系统能够正确解释外部数据，从这些数据中学习并适应来实现特定目标和任务，可以看作是一个函数，输入样本数据，输出期望的结果。[1]

二 AI 安全

（一）利用 AI 实现安全[1]

利用机器学习，可以自动识别垃圾邮件和网络钓鱼邮件，主要依托类大脑的“神经网络”，将其引入到垃圾邮件过滤器后，可以通过分析大量计算机上的信息来学习识别垃圾邮件和钓鱼信息。

AI 也可以预测和鉴别诈骗电话，如今来电显示可以自动标注“广告推销”“房产中介”等等，有诈骗电话和短信也可以自动进行拦截。

（二）AI 带来的安全隐患（对抗样本）

1. 对抗样本攻击

Szegedy 等首先提出了对抗样本这一概念[2]，即对输入样本添加一些人无法察觉的细微干扰，导致模型以高置信度给出一个错误的输出。在很多情况训练集的不同子集上，训练得到不同结构的模型，都会对

相同的对抗样本实现误分，这意味着对抗样本成了训练算法的一个盲点。而 Nguyen 等发现面对一些人类无法识别的样本，深度学习模型会以高置信度将其分类。大致有图像对抗样本攻击和语音对抗样本攻击。

2. 防御方法[3]

主要有四类方法：第一类将对抗样本与原始样本一起训练，进行数据增强；第二类采用其他的神经网络来验证输入的图像是否为对抗样本；第三类采用传统的机器学习算法，对输入图像进行去噪或变换；第四类采用深度神经网络和机器学习结合，即第二类和第三类的混合。

三 参考文献

[1] 张国明. IoT 安全交流

[2] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R.Intriguing properties of neural networks[C].ICLR(Poster).2014.

[3] 钱申诚,文宇恒,马耀飞,毛鑫唯.基于深度神经网络的对抗样本攻击与防御方法研究[J].网络空间安全,2022,13(05):77-86.