# ch14 人工智能安全

沙之洲 2020012408

## 1 请解释人工智能算法的鲁棒性和可解释性

鲁棒性指的是算法模型对数据变化的容忍度。

可解释性是人们了解模型决策原因的程度。可解释性越好的模型，其作出的决策对人们而言就越透明，反之模型本身就越像黑盒。

## 2 请简述投毒攻击和对抗攻击的不同点，请利用“自动驾驶”为场景各举一例

投毒攻击指的是，考虑一个鱼和狗的分类器，攻击者可以直接接触到分类器的训练过程。投毒攻击是将大量精心设计的狗的类别的图片混入训练集，使分类器不能够正确识别鱼类的图片。

在自动驾驶场景中，攻击者可以通过篡改交通标志或道路标记的数据来实施投毒攻击。例如，攻击者可能修改交通标志的形状或颜色，或者在模型训练数据中添加具有误导性的标志。这样，当自动驾驶汽车在现实世界中遭遇被篡改的标志时，它可能会误解道路规则，导致危险的驾驶决策。

对抗攻击指的是在模型的测试阶段，攻击者找到与正常样本非常接近，却能够让模型产生错误的样本，这种在测试阶段使模型产生混淆的攻击方式被称为对抗攻击。

在自动驾驶场景中，对抗攻击可以是通过对相机传感器输入进行修改来迷惑自动驾驶汽车的视觉识别系统。攻击者可能会对图像进行微小的扰动，例如添加不可见的噪音或干扰，使自动驾驶汽车误判周围环境，如错误地识别障碍物或道路条件，从而引发危险的驾驶行为。

两者的区别在于 投毒攻击是通过修改训练数据来使机器学习模型学习错误模式，而对抗攻击是通过对输入数据进行微小的修改来欺骗模型。投毒攻击主要影响模型的训练阶段，而对抗攻击则主要影响模型在实时推断时的性能。