网络空间安全导论 · Ch14

计01　容逸朗　2020010869

Q1.

请解释人工智能算法的鲁棒性和可解释性。

- **鲁棒性：**即模型对数据变化的容忍度，鲁棒性越高的模型，其识别噪声和对抗样本的准确率越高。
- **可解释性：**让人类了解模型做出某一决策的深层原因。在现实生活中，如果能够更透明地了解模型的决策过程，可以加强人们对模型的信任程度。

Q2.

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。

- 对抗攻击通过修改输入样本来完成欺骗目标系统的攻击，而投毒攻击则通过混入特殊样本直接对模型进行修改，不需要修改测试数据。
- 以自动驾驶识别交通标志信息为例：
  - 投毒攻击：在训练模型时，我们可以通过加入一些精心设计的图片，使得模型把「停止」标志识别为速限 100 的标志，从而完成投毒的目标。
  - 对抗攻击：同样在训练模型时，我们可以在「禁止通行」的样本中加入杂讯，使得模型识别为另外的交通标志，这样便是对抗攻击。