# Put your name here: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Assignment 9

## Secure AI

## Instructions

Follow the steps. Fill in the boxes. It is OK to ask ChatGPT to help you answer these questions. However, you are responsible for understanding your answers. Similarly, short concise answers are preferred over long wordy ones.

Pay attention to the wording of each question. I will indicate roughly how many sentences I expect. If you go way over that limit, you will lose points on the assignment. I will not be impressed by overly wordy, convoluted answers.

## Overview

In 2016, Microsoft launched Tay, an AI chatbot designed to engage in playful conversation on Twitter. Within 24 hours, Tay was shut down after it began posting racist, sexist, and inflammatory content—learned directly from users.

This assignment invites you to explore what went wrong, how it could have been prevented, and what Tay’s story teaches us about designing secure, ethical AI systems.

1. (10 points) Summarize Microsoft’s goals and the technology behind Tay. (One short paragraph)

|  |
| --- |
|  |

1. (10 points) What made Tay different from other chatbots at the time? (One short paragraph)

|  |
| --- |
|  |

1. (10 points) Describe the specific unaligned behavior Tay exhibited.

|  |
| --- |
|  |

1. (10 points) How did users exploit the system?

|  |
| --- |
|  |

1. (10 points) Refer to the SecureAI lecture… what would the Tay incident be an example of?

|  |
| --- |
|  |

1. (10 points) How is this a good example of SecureAI?

|  |
| --- |
|  |

1. (10 points) How is this a bad / limited / poor example of SecureAI ?

|  |
| --- |
|  |

1. (30 points) Rather than “pull the plug” on Tay, what other actions could Microsoft have taken? (List three items in a bulleted list.)

|  |
| --- |
|  |