Daniel Berry
HCC 6600
February 23, 2025

Project Report 3

1. How does AI alignment align with your project's goals?

   We need the application to stick to it's role of a conversational practice assistant, but we do not want it to make decisions that do not align with our planned interactions. It should provide responses that are not controversial or harmful to users or others.

2. Which 'Dreams and Nightmares' scenario is most relevant to your project's potential impacts?

   The chat bot can enhance human flourishing by giving users the opportunity to learn and improve on their foreign-language conversational skills. A nightmare scenario would be the amplification of biases and inequalities. The chat bot could encourage negative stereotypes or incorrect language within Italian culture.

3. Which Prometheus Principle is key to your project's ethical framework?

   The autonomy principle implemented in the project will give users the option to opt-in to data sharing, contribute corrections to the system, and decide what paths they should ultimately take within an interaction.

4. How might challenges in defining human values impact your AI project?

   I will need to hard code reasoning into the chat bot that is able to detect and prevent misaligned behavior. Using the RASA tooling system this must all be completed through modification of text files in the project or through practice conversations. It will be a very time consuming, but ultimately effective practice.

5. How could value alignment theory improve your project's adherence to human values?
   We do not want the chat bot to replace the emotional support that users should seek outside of the application. The tool should only ever be supplemental, but the best conversation practice will of course happen in the real world. A rule-based approach to certain sentiments should allow us to keep the chat bot on track.

6. Which value learning technique will you use in your project, and why?
   We will use reinforcement learning within the RASA framework to provide example good and bad conversational issues. The open source api, has all the tooling to train the model on new conversations.

7. How will you implement RLHF to ensure ethical AI outcomes in your project?
   The open source RASA project tooling has the option to interactively provide both sides of a conversation to train on. Using saved conversations I can improve the training of the model and teach it to avoid certain misaligned speech patterns.

8. Which online learning technique for alignment will you use to update your AI's objectives?
   Users will have an option on each response from the chat bot to provide a possible correction if they see an error? This will allow users to help shape the policy of the chat bot and guide it to a more accurate model.

9. What middleware for alignment strategy will you use to maintain ethical boundaries in your project?

   Monitoring responses from the chat bot and acting on them before they are returned to the user. Sentiment analysis would allow me to detect misaligned chat bot responses and correct them before they are returned.

10. How could international cooperation on AI alignment affect your project's impact?
    Sourcing feedback from native Italian speakers regarding cultural sensitivity, correctness and usefulness would help guide me to better aligning the training of the chat bot. Additionally, they could help train the model through reinforcement learning until it is able to more accurately model authentic Italian conversation.