**Chatbot Conversations: A Comparative Study on User Interaction and Preference: Project Proposal (working title)**

**Introduction:**

In the digitized era we live in, our relationship with artificial intelligence has become increasingly intricate, presenting many unique challenges to navigate. Joseph Weizenbaum’s 1967 program, ELIZA, while simplistic by today’s standards, was known to have elicited some feelings of attachment in those that used it, so one can only imagine the plethora of emotions that a modern or future program could evoke.

With the speed at which artificial intelligence and natural language processing is developing, it is becoming more and more vital to establish a solid foundation, backed by research, upon which we base our developments. Within the scope of this project specifically: the creation of chatbots with Natural Language Processing.

Chatbots have the potential to revolutionize many aspects of our lives, whether through virtual assistants at work, mental health assistance, or even serving as substitutes for human companionship. Hence it is necessary that we implement them in the right way – thoughtfully and effectively.

This project aims to provide an understanding of what people want from their interactions with chatbots as well as establish any correlations between types of users and their preferences. The objectives that need to be completed in order to achieve this include:

* Developing (prompt engineering) two distinct chatbots for users to interact with
* Composing a study to collect data from a user after they have interacted with a chatbot
* Creating a simple machine learning program that, given some data about a user, could predict what type of chatbot the user would prefer

What some researchers refer to as HMC (human-machine communication), is an emerging yet underdeveloped field (especially when framed around AI), so any contribution to it is a step in the right direction.

We live in exciting times when it comes to Artificial Intelligence. While the concept of an “artificial intelligence” has been around for a long time, only in the past few years have we really seen an acceleration in its progress and accessibility, with the likes of NLP models such as ChatGPT 4.0 and stable diffusion image generation to name a few examples. I have a deep desire to help contribute to this new technological boom, with the hopes that I can, in any way, better the lives of people in their day-to-day.

Overall, this project has the ability to act as the perfect ending to my academic journey, incorporating all I’ve learnt over the years, whether that’s SQL databases in first-year, web-based applications in second-year, machine learning algorithms in third-year, and even phycology/research methods that I learnt at A-level.

**Literature review:**

**Central Texts:**

*GAAIS:*

* [Full article: The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust (tandfonline.com)](https://www.tandfonline.com/doi/full/10.1080/10447318.2022.2085400) – This study outlines the development of GAAIS, and shows correlations between aspects of personality and attitudes towards AI (for example, introverts are more positive towards AI)
* [Who Likes Artificial Intelligence? Personality Predictors of Attitudes toward Artificial Intelligence: The Journal of Psychology: Vol 156, No 1 (tandfonline.com)](https://www.tandfonline.com/doi/abs/10.1080/00223980.2021.2012109) – a variation of the GAAIS, linking it to other personality factors (innovativeness) using a different sample of 1530 south Korean adults
* [Initial validation of the general attitudes towards Artificial Intelligence Scale - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2451958820300142#appsec1) – a validation of the GAAIS, confirming and building on its methods

*Importance of researching human-AI interaction:*

* [Artificial intelligence and communication: A Human–Machine Communication research agenda - Andrea L Guzman, Seth C Lewis, 2020 (sagepub.com)](https://journals.sagepub.com/doi/full/10.1177/1461444819858691) – highlights the emergence of AI as a communicator (HMC) and how traditional communication theory no longer applies. Proposes an agenda for HMC research
* [Frontiers | Editorial: On the “Human” in Human-Artificial Intelligence Interaction (frontiersin.org)](https://www.frontiersin.org/articles/10.3389/fpsyg.2021.808995/full) – highlights the importance of understanding how humans interact with AI, identifying key areas for further investigation (Eg: Human Characteristics Influencing AI Interaction)

*Trust in AI:*

* [The effects of personality and locus of control on trust in humans versus artificial intelligence (cell.com)](https://www.cell.com/heliyon/pdf/S2405-8440(20)31416-X.pdf) – research that shows participants are more likely to trust and incorporate suggestions given by an ai over a human in a decision game
* [Frontiers | Trust Toward Robots and Artificial Intelligence: An Experimental Approach to Human–Technology Interactions Online (frontiersin.org)](https://www.frontiersin.org/articles/10.3389/fpsyg.2020.568256/full) – a money giving experiment that suggests participants are more likely to trust ‘jdrx894’ (a robot) over ‘Michael’

*Other:*

* [Guidelines for Human-AI Interaction - Microsoft Research](https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/) – outlines some guidelines when setting up AI programs at different stages (important points include “make clear what the system can do” and “mitigate social biases”)

**General Texts:**

*Psychology of trust:*

* [Gender, Stereotypes, and Trust in Communication | SpringerLink](https://link.springer.com/article/10.1007/s12110-020-09376-3) – explains and investigates how it is generally stereotyped that women are more trustworthy and trusting
* [Study of the Day: How Gender Affects People's Trust - The Good Men Project](https://goodmenproject.com/newsroom/study-of-the-day-how-gender-affects-peoples-trust/) – has results that show that perceived trust can be dependent on timing of speech
* [(PDF) Trust and Gender: An Examination of Behavior and Beliefs in the Investment Game (researchgate.net)](https://www.researchgate.net/publication/222329553_Trust_and_Gender_An_Examination_of_Behavior_and_Beliefs_in_the_Investment_Game) – finds, via ‘the investment game’, that men trust more than women, and women are more trustworthy than men

*General psychology:*

* [Effects of taboo conversation topics on - ProQuest](https://www.proquest.com/docview/2457326127?forcedol=true&pq-origsite=summon) – states that a conversation is not evaluated positively when covering taboo subjects such as ‘sex, money, religion and politics’
* [Anthropomorphism and object attachment - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S2352250X20301548) – shows how it is common to attach human traits to everyday objects
* [Measuring Engagement with Mental Health and Behavior Change Interventions: an Integrative Review of Methods and Instruments | International Journal of Behavioral Medicine (springer.com)](https://link.springer.com/article/10.1007/s12529-022-10086-6) – outlines ways of measuring engagement (here in reference to mental health treatment), and states that the most common way is through usage
* [Short assessment of the Big Five: robust across survey methods except telephone interviewing - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3098347/) – an evaluation of the effectiveness of using a short version of the Big Five Inventory (BFI-S)
* [The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power - PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/27055049/) – Paper on the BFI-2: “an important advance over the original BFI” that is said to provide greater predicative power
* [Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0092656616301325) – an evaluation of shorter versions of the BFI-2 (XS) that are found to retain reliability
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4413730/> - research on how the colour red can draw your attention more

*General AI Research:*

* [ELIZA—a computer program for the study of natural language communication between man and machine (acm.org)](https://dl.acm.org/doi/pdf/10.1145/365153.365168) – Shows an early ‘NLP’ model ELIZA
* [Medical students' attitude towards artificial intelligence: a multicentre survey | SpringerLink](https://link.springer.com/article/10.1007/s00330-018-5601-1) – survey results showing medical students’ attitudes towards AI (positive for the most part)

**Methodology:**

**Creating the two chatbots (prompt engineering):**

When considering how to implement the chatbots, the first step is choosing what model to use. A new model made specifically for this research does have its advantages, however the cost and complexity of training a new model capable of doing what we need (having detailed and varied conversations) would be too high.

Out of the existing models, ChatGpt-3.5 seems to be the best suited for this task. It is capable of handling a diverse range of conversation style and topics, has a large context window, and the use of system messages allows for further direction. Additionally, I already have some experience integrating it into a web-based application. Here is the transcript for some preliminary tests run on GPT-3.5:

1. ***User:*** *we are talking about Napoleon and Napoleon only; the user asks you: what year was Napoleon born?*
2. ***GPT:*** *Napoleon Bonaparte was born on August 15, 1769, on the island of Corsica, which was then a part of the Republic of Genoa (now a region of France).*
3. ***User:*** *we are talking about Napoleon and Napoleon only; the user asks you: how are clouds formed?*
4. ***GPT:*** *Clouds are formed through a process called condensation, during which water vapor in the atmosphere cools and changes from a gas to tiny water droplets or ice crystals, depending on the temperature. However, I'm here to provide information specifically about Napoleon Bonaparte. If you have questions related to Napoleon, feel free to ask!*
5. ***User:*** *we are talking about Napoleon and Napoleon only. If the user asks about anything else, do not answer, the user asks you: how are clouds formed?*
6. ***GPT:*** *I apologize, but we are focusing solely on Napoleon-related questions. If you have a question about Napoleon, please feel free to ask.*
7. ***User:*** *we are talking about Napoleon and Napoleon only. If the user asks about anything else, do not answer, the user asks you: what year was Napoleon born? (please answer with as much personality as possible, even if that means obscuring some information)*
8. ***GPT:*** *Napoleon Bonaparte graced the world with his presence on a particularly warm August day in the year 1769. His birth occurred on the picturesque island of Corsica, which had its own share of historical significance. Quite the start to an extraordinary life, don't you think?*

This conversation highlights ChatGPT’s ability to keep within the limits of a specific topic (given that you direct it to only answer questions on that topic) as well as its capabilities when it comes to changing its patterns of speech: It can become a lot more conversational, giving its own “opinion” or have more of an emphasis on detail and accuracy if needed.

When it comes to the bots needed for this project, both bots will be limited in what they can talk about, making it much easier to compare results (reduces variance – some users might be more engaged when trying to push the bot to its limits). Research suggests that ‘taboo’ subjects can leave a negative impression after the conversation, but equally the topic should be interesting as to not bore the user so *‘Natural Geography’* has been settled on as it strikes a good balance.

There are several investigations into Gender differences and trust that conclude that women are often perceived as being more trustworthy than men. However, a key consideration when designing the AI application was the Microsoft Guidelines for Human-AI Interaction which states: “ensure the AI systems, language and behaviours do not reinforce undesirable and unfair stereotypes and biases”. Therefore, both bots shall present themselves as genderless, as to not perpetuate any stereotypes.

**Designing the questionnaire:**

The questionnaire/survey that users will answer in order to provide data for the research will be split into two parts. The first part revolves around gathering general user info and will take place before speaking to any chatbots. The areas it covers are:

* General demographic information (age, gender, level of education, computer expertise)
* The BFI-2-XS (an abridged, 15 question version of the Big Five Inventory 2) for assessing personality. BFI-2 is used in similar studies (such as GAAIS studies), and research shows that the shortened version is a robust enough method be used in a situation such as this, despite its brevity
* The positive and negative GAAIS: two subscales that measure how people feel towards AI in general. Scoring a high value on the positive scale and negative scale shows good attitudes towards the positive aspects of AI and a forgiveness for the negative aspects respectively. It is also found to strongly correlate with ‘general trust’, so measuring that as an individual variable can be left out

The second part of the questionnaire occurs twice after the user has spoken to each of the chatbots. It will use a 1-5 scale (Disagree strongly, disagree a little, Neutral no opinion, agree a little, Agree strongly) similar to that of the BFI-2-XS and the GAAIS for simplicity and uniformity. It will cover the users’ thoughts and feelings about the bot/conversation, and will look something like this:

* I felt confident in the accuracy and reliability of the information provided by the chatbot
* The chatbot was a useful tool in learning about Natural Geography
* The conversation with the chatbot kept me engaged and interested
* I am satisfied with the quality of the conversation I had with the chatbot
* I, at any point, felt as though what the chatbot was telling me was untruthful or unexpected
* [if the user answered yes to the previous question, they can explain, in words, what the unexpected occurrence was]

How long the user talked to each of the chatbots, as well as how many messaged they sent will also be recorded to measure engagement, though the user will not be explicitly asked about this.

**Creating (and using) the application:**

Upon entering the application (which will be a web application), the user will be met with a disclaimer that details:

* What the project aims to do and how the website will aid it (specifically what their data will be used for)
* How their data will be kept secure and used only for the purpose of analysis within this project
* What exactly the user is going to partake in (as well as roughly how long it will take), explaining how they will answer a few general questions, speak to the chat bots, and then answer a few questions about the conversations
* A warning about how the chatbots may “hallucinate” and produce responses that are not-aligned with its intended purpose. As well as a reminder for the user to ‘take note’ of any of these occurrences to report later

If the user accepts the outlined terms, they will then be presented with the first part of the questionnaire (general questions) and then start talking to one of the two chatbots (which one they speak to first is randomized to try and reduce any order bias). The conversation will have a capped time limit, as to not deplete the tokens attached to the ChatGPT key, that will be somewhere in the 5-10 minutes range (will be determined by early testing). After the time limit is exceeded (or the use clicks an ‘end conversation’ button) the user will then be asked questions regarding the conversation they just had. This process is repeated: speaking to the second bot and answering the same questions.

The two pages that display the chatbots will have a similar layout and display (different colours are known to elicit different emotions so colour schemes should be the same) with a few differences between the two. Namely, the more ‘conversational’ chatbot will have a more human-sounding (gender neutral) name displayed whereas the other chatbot will have a more robotic name (maybe simply ‘chat-bot’). Research suggests that an agent that is implied to be a bot (via its name) is more likely to garner trust in people, so this will be a negative factor effecting the trust of the conversational bot. To balance this out, some other research has shown that getting an immediate response reduces trust in that response, so the conversational bot will have a pause (as well as an animated ellipses to imply “thinking”) before each message, whereas the other bot will reply immediately.

**Data Collection and Analysis:**

Data collected from participants through the web application will be stored in an SQL database to organize the responses. From here we can compute the means and standard deviations of the demographic data, add correlate any of that initial data against the responses after talking to one of the bots (differentiating between and comparing results of the two different chatbots), displaying most of the data in graph form.

**Plan:**

**Project Plan (and roadmap):**

*The steps needed to be taken to complete this project are, in order, as follows:*

*(To be completed as soon as possible, before Christmas)*

* *Write a project plan, that has been approved by staff*
* Develop a good prompt format for ChatGPT to produce one of two types of response (one response will be quick and informative. The other, is more conversational)
* Create a questionnaire, drawing inspiration from several similar studies, that gathers both relevant information about the user, as well as their reaction (engagement and trust) to the chatbot they spoke to

*(To be completed over the Christmas period)*

* Create a web application that brings the chatbots and the questionnaire together – allowing participants to speak to the bot and fill in the questionnaire (with their results being saved to a database)
* Start rolling out the web application to a diverse group of participants/users

*(To be completed once enough data has been collected – aiming for February/March)*

* Analyse the participants data and producing a series of graphs to highlight any correlations or trends that might appear
* Using an analysis of the data, come up with a conclusive answer as to people’s preference on chatbot (writing the bulk of the final report)

*(To be completed in the final few weeks, after everything else)*

* Creating a machine learning program that, given some data about a fictional (or new) participant, could predict what type of chatbot the user would prefer
* Evaluate any shortcomings with the completed project and comment on any future additions that could help improve future iterations of the research

*(note: the writing of the final report will be a continuous process that happens as relevant tasks are completed, with the bulk of the writing occurring once enough data has been collected)*

**Resource Plan:**

*Resources that are needed to complete the project include:*

* A valid **ChatGPT API key** (a temporary one can be generated for free, and this process can be repeated a few times for the duration of the project)
* A service capable of hosting a web-application that can handle node.js (for GPT API calls) as well as SQL databases (for storing the collected data), an example of this is **Amazons AWS EC2**
* A service like **Jupyterlab** for training a learning model (The University has granted me access to theirs through the LH Neural Computation course)
* A source of **users** to complete in the questionnaire (a lot of them can be sourced via networking: friends, family, university societies, as well as possibly using a paid service)
* **Graphing software** for displaying the data (software like word and excel, which is already available, may suffice)

**Contingency Plan:**

*Aspects of the project that may need to be changed:*

* If it becomes an issue surveying users and saving the data via an SQL database on the application, then a simple Google form could be used instead
* If the ChatGPT model doesn’t work as needed then an alternative model such as Google Bard could be used
* If there are not enough users contributing data for analysis then a few things could be done: the amount of time the survey is available for could be extended, and a push for more users (from more sources) could be made. Other options include limiting the scope of the research so that a smaller sample is feasible (Eg: the research could just look into how introverts interact with AI)
* If time is becoming an issue, the predictive machine learning programming can be omitted from the project as it is not a central part of the research
* If the questionnaire is producing highly varied results (lots of ‘anomalies’ that make it hard to draw any correlations), then it may be the case that the survey takes too long (so less questions need to be asked) or is not focused enough, and so will have to be rewritten