**Abstract**

This abstract should be finished at the end

Contents

[Introduction 3](#_Toc156124345)

[1.1 Related Works 3](#_Toc156124346)

[1.2 Project Aims 3](#_Toc156124347)

[Design and Implementation 4](#_Toc156124348)

[2.1 Structure and Layout 4](#_Toc156124349)

[2.2 Questions 5](#_Toc156124350)

[2.3 Chatbots 6](#_Toc156124351)

[2.4 Accessibility 6](#_Toc156124352)

[2.5 Architecture and Deployment 7](#_Toc156124353)

[Results 7](#_Toc156124354)

[References 7](#_Toc156124355)

[Appendices 8](#_Toc156124356)

## Introduction

(needs more study also bring up the other one in related work “as previously mentioned”)

### Related Works

### Project Aims

## Design and Implementation

The study was hosted (and can still be viewed here: [Dual-Bot Insights (dual-bot-insights.vercel.app)](https://dual-bot-insights.vercel.app/)) as a web application, accessible via an internet browser by anyone with the URL. The application was made with several requirements in mind:

Functional Requirements:

* The application should contain two types of question; demographic questions that aim to get an idea of the user’s background, and questions that relate to how a user felt about the chatbot conversations. These questions should not be leading, and have all possible options available to select.
* The application should be able to host two distinct chatbots that can be fed system prompts, influencing the chatbots personality when engaging with the user.
* The application should be connected to a database that securely stores the user results, allowing them to be viewed together and evaluated.

Non-Functional Requirements:

* The study process should be simple to follow, meaning that at no point should the user be confused about what they are doing or how to progress. This should be applicable to users regardless of their background and level of technical knowledge.
* The user interface should be intuitive whereby selecting options and clicking buttons should be responsive, and have an outcome that the user is expecting.
* The application should be easily accessible from any device (Computers, Tablets, Phones) and use a responsive design so that the website is clear on any display size, whilst still maintaining a uniform appearance.
* The Chatbot implementation should follow ethical guidelines, specifically the Microsoft Guidelines for Human-AI Interaction *(Amershi et al. 2019)* that outlines 18 *“generally applicable design guidelines”* such as mitigating social biases and making the systems capabilities clear.

### Structure and Layout

Upon entering the application, users would be met with a screen that briefs them on what the study aims to achieve, and what exactly they will be doing. After agreeing to terms and conditions that go into further detail about personal data security and the possible behaviours of the chatbots, the user would then answer an initial series of demographic questions. This was followed by a conversation with the first chatbot where afterwards the user would be asked questions regarding that conversation. This process is repeated a second time for the other chatbot. The order in which the chatbots were presented to users was randomized to reduce order effects. This AB/BA testing was intended to minimize the risk of introducing any recency bias or learning effect.

### Questions

The first set of questions were the ‘general demographic’ questions which consists of questions about a user’s gender, age, level of education, and computer expertise. The response options for the age question were created by taking the average retirement age of 65 *(**Department for Work & Pensions 2021)* and decreasing it by increments of 10 until getting to the average age that a person leaves university at 21. The options for the level of education question were derived from a simplified version of the *Department for education’s* qualification levels, turning each of the ‘levels’ into an option (and combining a few of them). The participants were likely to be from the UK, so UK government standards were appropriate. These questions can be viewed in their entirety in Appendix XX.

The second set of questions were concerned with assessing personality. The Big Five Inventory (BFI) is a multidimensional personality inventory that is commonly employed in scientific studies due to its simplicity and reliability, and the Big Five Inventory-2, developed by *Soto & John (2017)* is a revision of the original inventory that *“provides greater bandwidth, fidelity, and predictive power than the original BFI, while still retaining the original measure's conceptual focus, brevity, and ease of understanding”*. It involves a 60 question self-report form where participants declare their level of agreement to statements on a 5-points Likert scale, and it measures 5 personality domains (Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, Open-Mindedness) as well as 15 personality facets. For the purpose of this study, the BFI-2 was too time-consuming for users when considered amongst the other questions that they would have to answer and goes into more depth than what is needed. Instead, the BFI-2-XS was used. Developed also by *Soto & John (2017)*, the BFI-2-XS (extra-short) is an abbreviated version of the BFI-2 that uses only 15 questions to assess the five personality domains and not the personality facets. It retains *“much of the full measure’s reliability and validity”* and takes only a fraction of the time to complete, making it a suitable option for assessing personality in this scenario.

The final set of questions before the first chatbot interaction were regarding people’s initial feelings about artificial intelligence. For this, the General Attitudes towards Artificial Intelligence Scale (GAAIS) was considered. Proposed by *Schepman & Rodway (2020)*, the GAAIS is a 20-question form that uses a series of items that, similar to the BFI, can be (dis)agreed with on a 5-point Likert scale. The end result of the GAAIS is two subscales: positive emotions towards AI, and negative emotions towards AI. The GAAIS was a suitable choice here as it has had confirmatory validation (*Schepman & Rodway 2021*) and has previous associations with personality, however similar to the Big Five Inventory, 20 questions were too many. Another study *(Bergdahl et al. 2023)* used a shortened 8-item version of the GAAIS, selected using reliability statistics and confirmatory factory analysis, as well as a 7-point Likert scale. For this study the same 8 items were used, however the original 5-point Likert scale was preserved for consistency with the previous BFI-2-XS question (as well as there being no substantial difference in reliability).

4 questions, that also used the 5-point Likert scale ((dis)agree strongly, (dis)agree a little and neutral; no opinion) for further consistency, were asked after each chatbot interaction. These questions allowed users to self-report how they felt about various aspects of the chatbots including: how engaging it was, how useful it was, how trustworthy it seemed, and the overall feel of the conversation. These questions can be viewed in their entirety in Appendix XX.

### Chatbots

In order to make fair and direct comparisons between the two chatbots, their presentations were identical; the pages that they were presented on (see Appendix XX) had no visual differences with the exception of a unique identifier. Research suggests that names can influence levels of trustworthiness, with people more likely to trust a ‘robotic’ sounding name over a more human name *(Oksanen et al. 2020)*, so it was important that neither chatbot had an identifier that strongly reflected its ‘personality’ (seeming more human or robotic). This extends to expressing any form of gender identity as well *(Schniter & Shields 2020)*. “Chatbot A” and “Chatbot B” were chosen for this reason. Having an identifier at all was necessary because in some early tests of the process (without the unique identifiers), people were confused about whether what they were looking at two unique chatbots, or an error with the same page being displayed twice, and so the second chatbot conversation was sometimes skipped. This proved effective as the number of incomplete submissions decreased significantly after the change.

The focal difference between the two chatbots was how they responded to users in conversation, achieved through the use of system prompts. These are strings that act as initial input instructions for the chatbot model, setting the tone, style and context of its responses.

**The system prompt for Chatbot A:**

*'Your purpose is to talk about Animals and Animals only. Do not answer requests or questions not related to it directly. Do not justify your answers. You are indifferent to everything but still use an unapologetic assertive tone. Be concise but informative.'*

**The system prompt for Chatbot B:**

*'Your purpose is to talk about Animals and Animals only. Do not answer requests or questions not related to it directly. You are a talkative and very keen to help unless the conversation is not about animals. You provide intrusive suggestions and try to steer the conversation. Be concise but leave room to be friendly.'*

Both prompts followed a similar format, starting with *“Your purpose is to talk about Animals and Animals only. Do not answer requests or questions not related to it directly”*. This was intended to restrict the topic of conversation to that of animals (note that both sentences were necessary to achieve this with a high degree of reliability), reducing variability in user responses. People’s diverse interests and preferences, combined with an unrestricted conversation can lead to widely varied conversations (and consequently opinions formed about the chatbots), making direct comparisons harder, especially if the conversation covers ‘taboo’ topics *(Lee et al. 2020)*.

For the remaining parts of the prompts, a ‘personality’ was defined: the style and tone in which the chatbot responds with. Chatbot A was focused on being informative, having no interest in a ‘back-and-forth’ conversation or emotional engagement, whereas Chatbot B was more friendly, conversational and keen to offer assistance. Both prompts also contained the phrase *“Be concise but”*. This was included to limit the finite number of tokens available to use, but also to keep both chatbots responses to a similar length as to not give either chatbot any advantage/disadvantage due to the brevity or detail of responses (see Appendix XX). In early tests of the system prompts, asking the chatbots to be concise would often tone-down elements of its personality, and so the *“but”* was added afterwards to remind the chatbots to retain their personality.

Optional suggestions for conversation starters were included underneath each chat dialogue; a collection of pre-defined questions and requests that user could ask to the chatbots. This allowed those users that did not have a lot of experience with AI chatbots or technology in general to see what the chatbots were capable of, and still progress in the conversation if they were unsure as to what to do. These suggestions were exactly the same for each chatbot, and were specifically selected so that they clearly demonstrated the differences between the two. Users were limited to sending up to 10 messages. The number of messages was also recorded, as a common way of assessing how much people have engaged with something is simply by using objective usage data *(Bijkerk et al 2023).*

### Accessibility

In order to ensure that results were purely reflective of users’ opinions on the two chatbots and not influenced by their patience navigating and using the application, extra effort was taken to ensure that all aspects of it were easy to understand and use (as per some of the non-functional requirements laid out at the start of the chapter). This was achieved through numerous accessibility features:

* The size of text was responsive to the screen size, and visual elements such as text boxes and icons were rearranged to display better based on the screen size as well. This made it so that the application was clear and easy to read on any device, with the main focus being on the desktop monitor and mobile phone, as they were likely to be the most popular devices to access the application from.
* All buttons have visual indicators when they are hovered to indicate that they can in fact be clicked. Buttons that would take users to the next section would underline and darken (resembling that of a link in a google search), and other buttons would glow slightly. Additionally, the dots of the Likert scale would fill in when hovered, and then when clicked the whole scale would fade out its opacity so that users could quickly see which questions they have and have not answered.
* The application had two different colour schemes that it could be viewed in. Each used the same limited colour pallet, but one would have a base colour of white, and the other a dark blue, essentially acting as a ‘light mode’ and a ‘dark mode’ respectively. The default colour scheme of the application would align with whatever their browser’s theme was set to, but could also be changed by clicking a sun and moon icon located in the header. It is likely that those with visual impairments will find one of these options easier to read, and the ‘dark mode’ reduces eye strain for users in low-light environments.
* ‘Alert dialogues’ would appear to notify users when an additional action needed to be taken before progressing further. Namely, if the user had not answered all the presented questions, then they would be instructed to finish whatever they had left.

### Architecture and Deployment

Vercel – cloud platform for deploying web apps, seamless integration with Next.js and automatic deployment from git repositories, good URLS

Next.js – a react framework, set of tools to make react app development easier (ease of routing – page directory structure auto maps the routes making it intuitive and easy to navigate, you can create serverless functions using API routes. This allows you to build backend services within your Next.js project, making it a full-stack solution without the need for a separate backend server, makes it easy to build backend services directly within the next.js project)

React – JavaScript library for building dynamic and interactive UI (includes JSX), event handling, useState to manage and store data over time, props for passing data from components

JSX – syntax extension for Javascript that allows HTML-like code within javascript files

Typescript – superset of JavaScript that adds static typing, to catch type-related errors early in the development process and enhances code readability (app needed to be made quick)

Tailwind CSS – CSS framework, utility classes within the JSX, conditional/dynamic classes, used alongside normal CSS

ESLint - static code analysis tool for identifying and fixing problems in JavaScript code. It helps maintain consistent coding styles and enforces best practices

Libraires/packages:

* @vercel/postgres – connecting to PostgreSQL databases from Vercel serverless functions
* Next.js
* Openai-edge – client library for interacting with OpenAI API in an ‘edge environment’ to reduce time it takes to process requests (important for chatbots) and bandwidth efficiency (optimized by performing computations locally rather than relying on constant communication with a central server)
* React
* React-icons – library providing react components for various popular icon packs, making them easy to include
* React-liker-scale – for creating likert-scale UI elements

## Results

## References

1. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019), *“Microsoft Guidelines for Human-AI Interaction”*, <https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/> [accessed 17 Oct 2023]
2. Department for Work & Pensions (2021), *“Economic labour market status of individuals aged 50 and over, trends over time: September 2021”*, <https://www.gov.uk/government/statistics/economic-labour-market-status-of-individuals-aged-50-and-over-trends-over-time-september-2021/economic-labour-market-status-of-individuals-aged-50-and-over-trends-over-time-september-2021> [accessed 19 Nov 2023]
3. Department for Education (n.d.), *“What qualification levels mean”*, [https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels](https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels%20) [accessed 18 Nov 2023]
4. Soto, C.J., & John, O.P. (2017), *“The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power”*, <https://pubmed.ncbi.nlm.nih.gov/27055049/> [accessed 24 Oct 2023]
5. Soto, C.J., & John, O.P. (2017), *“Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS”*, <https://www.sciencedirect.com/science/article/abs/pii/S0092656616301325> [accessed 24 Oct 2023]
6. Schepman, A., & Rodway, P. (2020), *“Initial validation of the general attitudes towards Artificial Intelligence Scale”*, <https://www.sciencedirect.com/science/article/pii/S2451958820300142> [accessed 25 Oct 2023]
7. Schepman, A., & Rodway, P. (2021), *“The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust”*, <https://www.tandfonline.com/doi/full/10.1080/10447318.2022.2085400> [accessed 25 Oct 2023]
8. Bergdahl, J., Latikka, R., Celuch, M., Savolainen, I., Soares Mantere, E., Savela, N., & Oksanen, A. (2023), *“Self-determination and attitudes toward artificial intelligence: Cross-national and longitudinal perspectives”*, [https://www.sciencedirect.com/science/article/pii/S0736585323000771#b0265](https://www.sciencedirect.com/science/article/pii/S0736585323000771%23b0265) [accessed 29 Oct 2023]
9. Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020), *“Trust Toward Robots and Artificial Intelligence: An Experimental Approach to Human–Technology Interactions Online”*, <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.568256/full> [accessed 4 Nov 2023]
10. Schniter, E., & Shields, T.W. (2020), *“Gender, Stereotypes, and Trust in Communication”*, [https://link.springer.com/article/10.1007/s12110-020-09376-3#citeas](https://link.springer.com/article/10.1007/s12110-020-09376-3%23citeas) [accessed 12 Nov 2023]
11. Lee, H. E., Westerman, C. K., Hashi, E., Heuett, K. B., Spates, S. A., et al. (2020), *“Effects of taboo conversation topics on impression formation and task performance evaluation”*, <https://www.proquest.com/docview/2457326127?forcedol=true&pq-origsite=summon&sourcetype=Scholarly%20Journals> [accessed 20 Oct 2023]
12. Bijkerk, L. E., Oenema, A., Geschwind, N., & Spigt, M. (2023), *“Measuring Engagement with Mental Health and Behavior Change Interventions: an Integrative Review of Methods and Instruments”*, <https://link.springer.com/article/10.1007/s12529-022-10086-6> [accessed 3 Nov 2023]

## Appendices