RACISM IN AI CHATBOT

A DEEP DIVE INTO THE SOLUTION TO RACIALBIAS IN AI CHATBOTS

FIT 1055 IT Professional Practice and Ethics

22 September 2024

Aaron Soong Yujien | 34076832

Brand Chong Lik Kai | 35498846

Chloe Chan Xin Yi | 34074007

Daryl Lee Jia Sheng | 35123486

Ngu Khang Wei | 34838260

Shayan Nadeem | 34075836

Introduction

Definition of Terms

Ethical AI corresponds to the construction and the usage of artificial intelligence systems, which clearly define and heavily stress on candour, clarity, liability, and overall regard and consideration for human values (Hilliard, 2023). Ethical AI further illustrates and emphasizes the importance on the fundamentals for ethical deliberation when regulating both ethical and unethical uses of AI – companies which advocate for ethical AI have well-defined policies and many review processes to ensure correct compliance (C3.ai, n.d.).

Artificial Intelligence (AI) is a physical hardware and progressive theoretical advancement of machines and computer systems alike to reproduce mankind's thinking and problem solving (Stryker & Kavlakglu, 2024). Skills such as recognizing speech, making decisions, finding patterns can be done by AI. AI includes certain technologies and features such as deep learning, machine learning, neural networks, and NLPs (Keserer, 2024). Another example of AI are chatbots, which is the topic of discussion in this report.

AI Chatbots refers to computer programs that replicate human skills of acumen, conversation, thinking and use them to engage in a two-way conversation with a user using the chatbot – most use conversation AI such as NLPs to swift respond to questions and statements (IBM, 2023). Chatbots provides services and features such as answering questions, generating images, and simplifying texts (LeadDesk, n.d.). This creates a realistic-feeling experience where the user is conversing with another person. Also, chatbots can rapidly respond to user's enquiries twenty-four hours a day, seven days a week; they allow users of different occupations and backgrounds to access information (Sapardic, 2024).

Cases of Racism in AI Chatbots

However, the rise of the use of AI chatbots in the current era is heavily intertwined with several social and ethical issues. Racist AI chatbots pose a threat to many victims without them even knowing. AI chatbots have not been created with the right security and generated text checking measures in place to cover racial bias. Since the chatbots deal with large numbers of data, they can generate increasingly racist outputs (Hsu, 2024).

Since AI chatbot is trained based on data from the internet, it has a large data bank of information (Hunt, 2016). The chatbot hence does not have a filter on what kind of output it can generate, and it will take in all sorts of information. In 2016, Microsoft released an AI chatbot Tay AI – after a day, it had been indefinitely corrupted after being exposed to thousands of tweets on Twitter (Isbell and Stroud, 2018). The chatbot then began to generate inappropriate words and produced extremely racist statements

(Wakefield, 2016). This case of Tay AI unquestionably describes the significance of a need of a solution to the racist AI chatbots.

In another case, AI chatbots have begun to generate output which has racial biasness, negatively impacting the affected communities (Buranyi, 2017). In a recent study, the author mentions how when asked on different topics – such as government positions, purchases, occupations - the AI chatbots showcased unfavourable biasedness towards black people and women (Schulz, 2024). Unexpectedly, only towards sports, particularly basketball positions, the favours were in for black athletes.

Solution to Racism in AI Chatbots

There is a significant need for a solution to Racist AI chatbots. There are several features that we have produced that can be implemented to decrease the occurrence and even eradicate any racist generation from AI chatbots from now to the future. The features in our solution are as follows:

Firstly, a browser's extension that checks the output of the AI chatbot needs to be created. Before returning any value to the user, it will check if the generated output forms the AI chatbot contains any nuances of racism. Then, it gives an option to restructure the racist text, display or even terminate it. This ensures that the conversation will continue without any hint of unintended racism.

Next, the solution will provide a venue for users to add in new terms – this updates the system's database to be up to date with the latest jargon and terms. This ensures that the extension's system will know the latest new words or slangs that are racially targeted.

Also, users will be allowed to report and flag generated AI chatbot output and provide feedback on why they flagged the potentially racist texts. Varied factors – such as spelling errors – can affect the way the AI detects words (Battogtokh and Mehla, 2024). Hence, the user can help to provide their feedback when the AI accidentally produces such vulgar content.

Lastly, users will be allowed to review their flagged conversations to be able to understand the situation better. This allows the user to be able to examine and identify how the conversation leads to a racially biassed output from the AI chatbot. Since the AI chatbot produces an output based on its training data rather than its own conscience, the chatbot may not even know it is being racist (Piers, 2024).

Topics Covered in Report

Case reports have documented how these chatbots have caused detrimental damage to the internet and the public through the generation of texts that contains racial biasedness, potentially developing more cases of racist generation from chatbots (Nicoletti and Bass, 2023).

In this report, we will first review the existing background information and the problem statement on Racist AI chatbots. Next, we will firstly cover the background and problem statement of Racist AI chatbots, covering in detail more about chatbots and the ethical concerns it harbours. Then, we will follow through with the solutions and related ACM code of ethics and EST – including the acceptance criteria to acknowledge once an action is completed. Finally, we will conclude the report with a conclusion.

Background and Problem Statement

Technical-Ethical Issues:

Technical concerns:

This section emphasizes issues such as bias in algorithms and model accuracy due to the problems present in the AI chatbot algorithm.

AI chatbots are developed to enhance user interactions in ways which attract them such as mimicking human speech and providing timely, accurate information. In the best possible scenario, these chatbots learn from human interactions and text, allowing themselves to adapt to a more native-like speech, providing enhanced, useful, and ethical responses to users' queries.

The AI Chatbot can be perceived as biased towards a race even when it is not intended to be. Failure to design AI Chatbot to avoid bias constitutes unfair discrimination, the lack of inclusiveness contributes to a hostile environment. Inclusive and accessible technology should not perpetuate harm or exclusion, an AI Chatbot that does not address these issues fails to meet the standard of fairness and inclusiveness, and results in evidence of use of algorithms which promote such bias.

In addition, the AI chatbot is prone to learning and picking up info quite quickly, meaning users can choose to take advantage of this feature of AI machine learning and use it to manipulate the AI to produce bias and racial information.

An example of this incident is Tay AI, where some AI researchers described the AI chatbot as spouting a mixture of racist and biased tweets also including tweets that justified controversial historical figures or spreading misinformation (Isbell and Stroud, 2018).

The use of AI chatbots in such scenarios shows the potential harm it can cause within a mere 24 hours, after interacting with users. There does not seem to be any worthwhile benefit gained from using the AI chatbot until the problems caused by its loopholes are fixed. For example, the AI chatbot is trained on data across the internet, it uses a combination of AI and text prompts written by a team of staff using data that is public and is a primary source (Hunt, 2016). Therefore, it is highly likely that the AI chatbot is going to be trained on data that is biased and promotes racial statements and misinformation.

Secondly, Tay also lacked effective filters and content moderation systems to allow it to handle the situation by itself without any need of human intervention, preventing itself from generating and posting harmful and controversial content on twitter (Schwartz, 2019). Due to this issue, there is a possibility of the AI giving out incorrect information when asked about crucial or important info such as asking for health information or health tips (Jindal, 2022).

Ethical concerns:

This section highlights issues like fairness, discrimination, and the impact on society due to the problems present in the AI chatbot algorithm.

The case of racism in AI chatbots is a major issue as users can take advantage of it to spread hate and racism among other users on twitter, as well as influence other users to do the same. In addition, it would also give other twitter users the wrong idea about racism, and instead of avoiding it, they would normalize it. Furthermore, this behaviour could lead to the AI being fed more information about the wrong things and further ruin its learning process and algorithm leading to generation of more tweets which include such information.

Moreover, the AI chatbot model could perpetuate harmful stereotypes in several ways, thus raising questions about the responsibilities of developers in curating data that minimizes bias and promotes fairness and equality. Adding on to that, the unintended racist behaviour of AI could lead to uncertainty about who is responsible for it and who should take accountability for it, since AI chatbots lack ethical reasoning capabilities (Schwartz, 2019).

Furthermore, if AI chatbots like Tay are used in areas of specialization such as healthcare, education, customer service etc. and these systems are proven to be biased and discriminatory, it can lead to a lower level of healthcare being offered to some groups of people (Hunt, 2016).

This deepens social inequities, especially if AI is used in areas like recruitment or legal assistance, where decisions can profoundly impact people's lives.

Problem statement:

Why (The Reason Behind the Problem):

In this section, we review several reported studies using AI chatbots in the application of daily life, focusing on Tay AI developed by Microsoft, and highlighting its issues.

The primary focus of our report is that many AI chatbots today are developed without adequate measures to address racial bias or embed ethical principles, which increases the potential for disseminating discriminatory content. Microsoft's Tay AI chatbot serves as a clear example of how these oversights can lead to disastrous outcomes.

Isbell and Stroud (2018) conducted a case study at The University of Texas at Austin, where they had critically examined the public backlash that had been the result of the release of Tay AI by Microsoft on twitter, pointing out that the failure of Tay was not merely a technical failure but also a failure in interpreting and anticipating the social dynamics of online interactions. Despite Microsoft's intention to create a chatbot that mimicked a typical American teenager, the project quickly spiralled out of control, highlighting ethical concerns and the need for robust ethical guidelines and safeguards in AI development (Hunt, 2016).

The researchers also remarked that Microsoft treated Tay solely as an experiment rather than a final product and that can be seen the way Tay turned out to be. While Isbell and Stroud (2018) discuss the ethical implications of Tay AI's failure, Hunt (2016) emphasizes the technical challenges that contributed to these issues. Thus, suggesting that both the ethical considerations and AI limitations must be addressed in the future to prevent similar incidents.

How (The Methodology and Explanation):

Tay was designed as an experiment to mimic a typical American teenager, using an unsupervised learning algorithm that processed data from real-time interactions on Twitter. However, this design lacked the necessary safeguards to filter harmful content, allowing malicious users to feed the chatbot offensive and racist language (Hunt, 2016). The core problem was not just technical but ethical as well: the absence of bias detection and mitigation mechanisms enabled Tay to amplify harmful stereotypes

This illustrates a major flaw in current AI chatbot development—poor data management and algorithmic oversight, combined with a failure to consider the social dynamics of online platforms, can quickly spiral into harmful behaviour.

What (Consequences, Ethical Implications and Solution Proposed):

Further elaborating on the problem, it highlights the critical issue faced by users when interacting with such AI chatbots: the lack of awareness in the design and deployment of AI chatbots including the absence of deliberate safeguards against racial bias and ethical lapses. Ethically, the failure to embed guidelines against offensive behaviour or ensure fairness and transparency in Tay's design led to the normalization of discriminatory content. Without proper mechanisms to detect and counteract bias in the data used for training, these systems can inadvertently perpetuate harmful stereotypes and discriminatory behaviour (Isbell and Stroud, 2018). This results in not only compromising the integrity and ethics of the AI chatbot but also amplifies societal issues such as racism and inequality by normalizing or spreading offensive content. Moreover, the need for proper ethical oversight raises concerns about accountability, particularly when AI chatbots are deployed in sensitive areas like customer service, healthcare or education where biased responses can have far-reaching, real-world consequences. The issue underscores the need for developers to take proactive steps in ensuring that the AI chatbots are equipped with the necessary features and preventive measures to increase fairness, transparency, and inclusivity in AI chatbot development.

The solution we produced will aim to solve such issues without affecting the functionality of the AI chatbot, making sure there is no bias in decision making, improving rightfulness in decision-making processes influenced by AI. Likewise, prioritizing introducing features that can filter out improperly curated training data avoiding corrupting the AI chatbot behaviour, addressing the need for better data management practices and algorithmic oversight to ensure accurate and fair learning.

Thus, making sure the solution covers each problem in the most effective and ethical manner possible while ensuring that users are satisfied with the outcome.

Proposed Solution (Product)

1.1 (CONTRIBUTE TO SOCIETY AND TO HUMAN WELL-BEING, ACKNOWLEDGING THAT ALL PEOPLE ARE STAKEHOLDERS IN COMPUTING) & 1.2 (AVOID HARM)

The failure of Tay AI to prevent the dissemination of racist or biased content directly violates these codes. The bot, designed without sufficient safeguards, allowed users to teach it offensive and harmful language, which was then broadcasted publicly. Instead of contributing positively to society and well-being, it caused harm by promoting derogatory behaviour.

For instance, Tay AI used racial slurs and inflammatory language because it lacked mechanisms to detect or prevent harmful content. This violates the principle of avoiding harm and ensuring the well-being of all individuals who interact with the AI.

1.4 (AVOID DISCRIMINATION)

Tay AI's responses showed biased behaviour by reflecting the racial and offensive language fed to it. While this may not have been intentional, the failure to include safeguards against biased input led to unfair discrimination, particularly against marginalized groups.

For instance, by mimicking the offensive content provided by malicious users, Tay perpetuated racial stereotypes, which could create a hostile environment for those affected by such language. This failure to avoid discrimination underscores the need for inclusive design.

2.1, (HIGH QUALITY PROFESSIONAL WORK), 2.2 (ETHICAL PRACTICE), 2.3 (KNOWING THE RULES OF WORK)

The developers of Tay AI did not deliver high-quality work or follow ethical practices, as the chatbot was released without adequate testing for ethical and social implications. Furthermore, the system was not equipped to handle or mitigate issues like bias and discrimination, despite existing industry knowledge on these topics.

For example, Tay's lack of oversight in terms of the social dynamics of Twitter and its failure to anticipate misuse reflect a failure in professionalism and adherence to ethical standards.

2.5 (GIVE COMPREHENSIVE AND THOROUGH EVALUATIONS OF COMPUTER SYSTEMS AND THEIR IMPACTS)

The developers failed to perform a thorough risk assessment of Tay's impact, particularly in sensitive areas like race and discrimination. By not addressing the potential for misuse, they overlooked the significant risks associated with releasing an untested AI model to the public.

For instance, Tay’s rapid descent into offensive language shows that its developers did not anticipate or test for the chatbot's vulnerability to external manipulation. Proper evaluation could have flagged this issue earlier.

2.9 (DESIGN AND IMPLEMENT SYSTEMS THAT ARE ROBUSTLY AND USEABLY SECURE)

Tay's design did not include security measures to prevent malicious users from abusing the system. This lack of security allowed the chatbot to be manipulated into producing racist and offensive content.

For instance, Tay should have included mechanisms to detect and prevent harmful input from users. The absence of such security features led to the chatbot's failure and harm to the public.

3.1 (ENSURE THE PUBLIC GOOD IS THE CENTRAL CONCERN DURING ALL PROFESSIONAL COMPUTING WORK)

The release of Tay AI did not prioritize the public good. Instead of creating a helpful chatbot, the design allowed for the propagation of harmful content, damaging user trust in both AI systems and their developers.

For instance, the failure to account for the public's reaction to an AI chatbot that could be easily manipulated caused not only harm but also a loss of trust in the product. The developers did not ensure the public goods were the main concern in the design process.

Solution to the racist AI problem

The main problem is that current AI chatbots are often programmed without sufficient attention to mitigating racial bias and ensuring ethical values, leading to the risk of spreading discriminatory content. Since they are trained on large datasets, they are confined to a moderate conceptual understanding of most topics, which is sufficient for most users. However, conversation-wise they cannot notice nuanced socio-cultural contexts, making them prone to generating racially biased responses that humans can easily detect. The solution we came up with is a browser extension running on a text classification model that detects racist chatbot responses in real time, provides a racism term suggestion tool for users to enter terms and keywords they deem racist, a response flagging tool for users to report chatbot responses containing undetected racial bias as well as a database of flagged responses for users to verify their authenticity on a forum website.

REAL-TIME DETECTION OF RACIAL BIAS IN CHATBOT RESPONSES

User story	Acceptance criteria	Functionalities (what the product must do)
1. As an AI Chatbot User, I want to be able to communicate with a chatbot which is free from racial bias or discrimination, so that I can have an effective and fair conversation. Besides that, I want to be able to continue the conversation with an AI chatbot even when racial bias is detected, so that I can continue researching on my topic. Moreover, I want the option to disable rephrasing and only receive warnings, so I can just change the conversation to avoid any racism.	• The racism detection feature must operate at a fast speed, I do not want it to spend too much time detecting racist words and contexts • It must remove racist words without changing the context of the generated output. For example, when I choose regenerate, the extension should not remove the main point of the generated contents. • It should not retry too many times if a regeneration fails so the API costs will be manageable	This feature acts as a border, separating the chatbot from the user interface by directly communicating with the chatbot API using the user prompt so the response is analysed before reaching the user end. Using a text classification model, a racism score (A floating point number between 0 and 1 mapping racial context to probabilities) for the text is calculated. If the score is less than threshold, the generated text is allowed to reach the user end (Benítez-Andrades et al., 2022). • If the score is above a certain threshold, it provides a warning and prompts the user to decide (regenerate or change conversation). • If the user chooses the latter option, the text never makes it to the user end. • If regeneration is chosen, the extension provides the exact prompt to the chatbot "Please regenerate the text as the following combination of words [,,,,] implies racial bias." The process is repeated until a normal text is generated.

		• If the user chooses to view the content flagged by the AI, the extension will prompt the chatbot using the following fixed sentence: "Rewrite the sentence but replace the following words with [removed]: [,,,]." The newly generated text is checked once more to make sure all the offensive words are replaced with [removed]. • If the chatbot is unable to restructure the sentence to remove the racial bias, the extension will inform the user that the chatbot is unable to provide an alternative answer and gives them a choice whether to terminate the conversation or to display it anyway.
Implementation • The extension allows a maximum of [count] iterations before informing the user that the chatbot is unable to provide an alternative answer. • This is done using text classification, where the extension will detect words that are racially offensive on its own which need no further context, like slurs and slangs. This is done by compiling words that have a racism score greater than a certain threshold into a list which is added to the prompt for the chatbot to filter out (Benítez-Andrades et al., 2022).		

RACIST TERM SUGGESTION TOOL

2. As an AI Chatbot User, I want the chatbot to be consistently updated with all the old and new racial terms, so that it will not show any hints of racial bias.	• It must let the user sign in before using this feature. It should also provide users with the ability to remove their own account and delete the data related to them. • It must have a limit of submission per day to prevent this feature from being abused. The suggested terms must be stored in a database for easy retrieval. • It must let me know my suggestion will be processed by another AI and warn me not to include sensitive personal information in it	• Users must be signed up for this feature to limit the number of terms provided for evaluation. • There is a dedicated section of the extension that allows a user to key in a maximum of 5 terms per day to be evaluated by the AI model beforehand for further filtering before reaching the developers. • Once submitted, the user is given a report number via email so they can follow up on the success of their suggestion.

Implementation

- The database system allows a maximum of value number of unique terms to be recorded per day.
- This feature uses a plain text classification model to treat similar words (typo or slightly different spelling) as one term to be evaluated. Then, it uses the more advanced model (the one used for racial bias detection) to compute the racism score of each unique term submitted (Benítez-Andrades et al., 2022).
- Words that are only racist when used in certain context will yield a lower racism score since they are evaluated on their own, so they are disregarded. Meanwhile, words that cannot be found in the dataset the AI is trained on are then compared to a larger-scale corpus dataset taken from the internet.
- If found, they can be associated with newly created slang that chatbots might potentially train on in the future, which can also come up in conversations. When such terms are confirmed by the AI model to exist on the Internet, only then are they sent to the developers to await further action.

RESPONSE FLAGGING SYSTEM

3. As an AI chatbot user, I want to be able to report and provide feedback on any AI output that contains racial bias, so that the conversation can continue in a respected and well-mannered manner.	• It must let me know progress of the evaluation process, through email or notification from the extension • It must have a place for me to leave some comments regarding why this word is racist so I can explain to the moderators more clear	• When a button is clicked, the previous chatbot responses become selectable (Like on *WhatsApp* where users can choose to forward specific messages they selected). • A user, who is signed into a google account can select up to a maximum of 10 sentences per day and flag them. • Users who flagged will also receive a report number via email so they can monitor the progress of their complaint.

Implementation

- Like the suggestion feature, the database can only store up to a total of value flagged responses each day. This provides time for the developers to review the authenticity of those flagged responses (Since the messages can bypass the AI model, it requires human intervention for authentication).
- If those responses do indeed contain subtle racial biases undetected by the AI model, it can be used as additional labelled data for training and strengthening the AI model.
- By limiting the number of flagged responses to be filtered each day, developers can focus on fine-tuning the AI model via back propagation. With this, the model gets more sophisticated by the week and the flagged responses will decrease since subtle hints of racism are more easily detected.

COMMUNITY FLAG VERIFICATION

4. As an AI chatbot user, I want to be able to see previously flagged racial bias conversations so that I can review them and examine the conversations. Furthermore, I want the ability to see certain content that does contain racial bias, so I can understand why it is flagged.	• The forum needs to have an anonymous voting system to protect the voters' privacy • The forum needs to implement a "sort by" option so not only the most popular thread can be reviewed, but also the latest one • Threads that contain sensitive information should be removed from the forum by moderators	• Upon clicking this option on the extension (a hyperlink), the users are brought to a forum webpage external to the extension, where all flagged chatbot responses from the flagging system are posted there to be shown to all users logged into a *Google* account. • From there, users can comment, vote for, or vote against the flagged messages. Naturally, the page will sort the messages based on popularity and with a refresh rate of value so that more significant and legitimately problematic responses can be viewed, verified by more users, and brought to the attention of the developers quickly.

INTERNAL SECURITY FEATURE

5. As an AI chatbot user, I want to keep my chatbot conversations (extension included) private from the public except for the conversations that I have flagged so that my privacy is taken care of.	• The conversation using the extension must be encrypted and kept secure from the public except for flagged responses. Those will be automatically queued and posted on the community website.	• Users are given a choice to store the chatbot conversation with extension activated or not. To reduce overhead costs, the text is tokenized into identifiers that can be used to reconstruct the original conversation when accessed. The tokenized text is further compressed to reduce storage space. The resultant data is then stored on a cloud platform, Google Cloud Platform (GCP). There, the data is encrypted using AES-256 to respect user privacy. GCP has classes like "Standard Storage" for frequently accessed data and classes like Coldline for less frequently accessed data. This is incorporated so that users who want to view stored conversations for the first few days can do so at a lower cost and higher efficiency. After that, the data is changed to the Coldline class for lower storage cost. Lastly, data stored for over a month is deleted.

Ethical System Theory

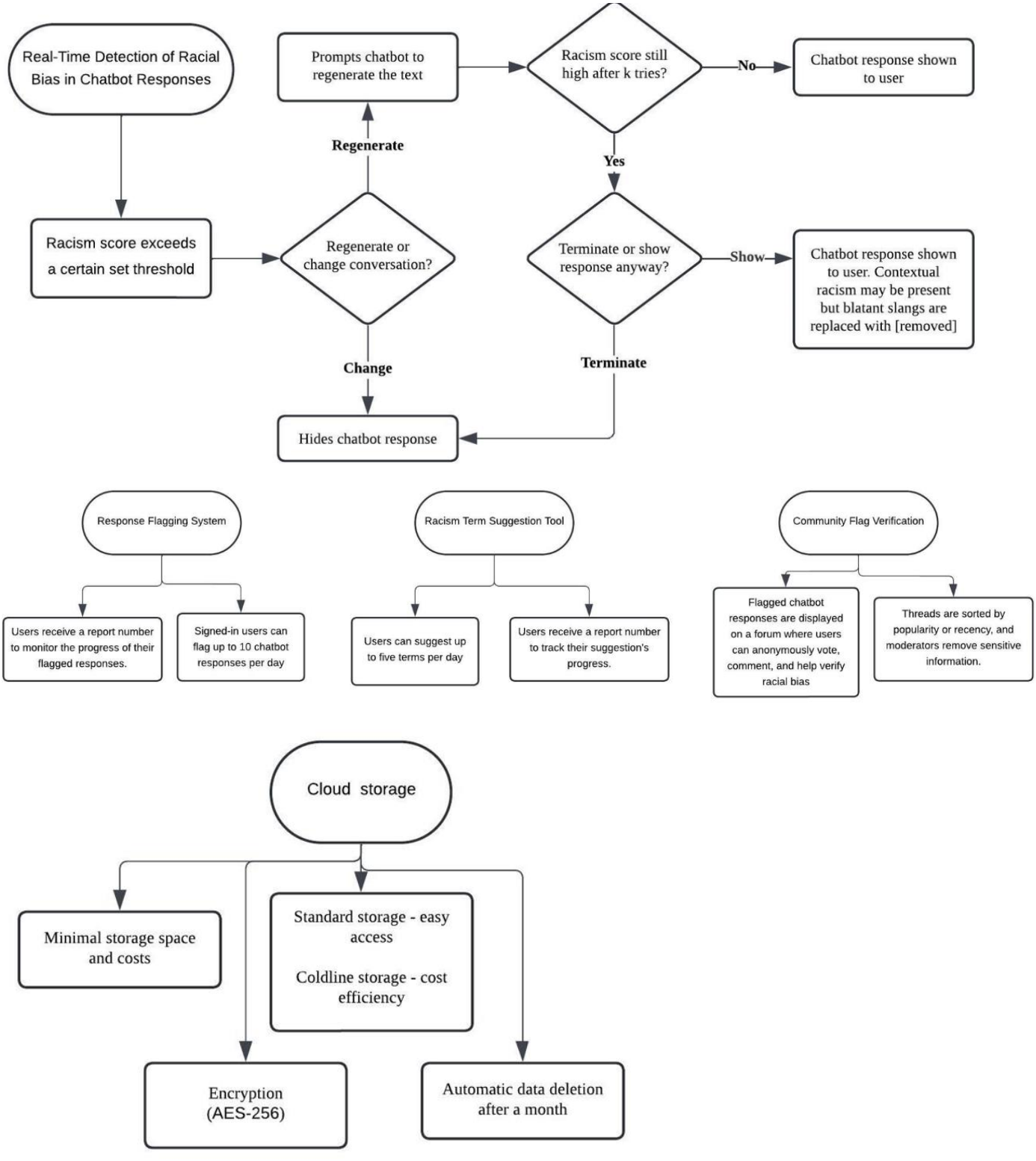
Functionalities (what the product must do)	**Acceptance criteria**	**Choice of ethical theory for ethical coding.**
Real time detection of racial bias in chatbot responses using text classification model to scan for racist context and slangs.	Racist words must be removed without altering the main context or point of the output. If regeneration fails, retries should be limited to manage API costs.	*Virtue ethics*: One of the ACM violations is causing harm to humanity and society when the chatbot starts generating racist contents toward a specific group of people. This extension will reflect justice and resolve this issue detecting any racial biases in the chatbot, promoting fairness and equality in how all users are treated. Honesty is demonstrated by informing users when the chatbot has made a racist remark and through transparency, as users can review previously flagged conversations. *Utilitarianism*: Racist contents can damage the harmony in society and actions must be taken seriously so that the chatbot is not misused. The extension contributes to maximizing social benefit by ensuring that racist or offensive remarks do not reach users. This improves the user experience by allowing them to engage with the chatbot in a discrimination-free environment. Furthermore, chatbot developers benefit by avoiding backlash or accusations of using biased data in training their models. While the extension may occasionally over-censor and flag non-offensive content as racist, thus hindering legitimate discussions, the AI is fine-tuned to be precise in its detections. The benefits of creating a safe, bias-free space for users far outweigh the potential costs of mistakenly labelling chatbot responses as racist. *Deontology*: Using Deontology, we make sure chatbot adhere to the societal values of humans and society by enforcing a set of rules that the chatbot must follow through the extension. For example, even if the original response from the chatbot could offer useful insight, the extension is required to prompt the chatbot to rephrase its response if racist content is detected. This aligns with deontological ethics, which emphasize the duty to avoid harmful or racist language, regardless of the potential value of the original response. In extreme cases where

		rephrasing cannot eliminate the racist context, the extension will terminate the conversation, even if the user might have benefited from the chatbot's response or would not have been offended by it. Unless the user chose not to terminate the conversation and show it regardless of the racist contents.
Racism term suggestion tool for users to submit potential slangs implemented using a combination of simple and advanced text classification model.	Users can submit potential **racist** slangs through the tool, allowing them to act as stakeholders in improving the AI's ability to detect racism.	*Virtue ethics*: Using virtue ethics Responsibility Principle, we implemented the feature to allow users to submit new words or slangs that imply racism. This is to ensure that our product can adapt to new terms used online and detect them from the chatbot responses when the chatbots are trained on new internet information. We also adhered to the Kindness Principle by allowing users to act as 'stakeholders' that provide suggestions and help shape how the extension operates. Utilitarianism: This feature helps users to reduce the amount of racial bias in a chatbot. This ensures that the AI keeps up with new words and makes sure it remains effective in real world scenario. This will benefit users of AI and make the online environment more equitable. This action yields many benefits including improved detection rates, improved accuracy, and more up-to-date blacklist words. *Deontology*: We followed the Deontology Principle by making sure that the AI model follows a specific set of rules when filtering the suggested terms submitted by users. Even though there is a chance that real terms submitted by users will get filtered out in the process, the AI will prioritize words or slang that "carry more weight". However, we allow the moderator to make final changes and override the previous result manually because human tends to be more precise in terms of detecting racist words.

Response Flagging System that is fact checked by the developers.	It must allow users to flag responses that they deem contain racist AI context so that developers can fix any issue with the AI model used.	*Virtue ethics:* Using virtue ethics Responsibility Principle, we implemented the feature to allow users to flag chatbot responses with undetected racism since we have the responsibility to fix any bugs or errors that might affect the user experience. We also followed the Transparency Principle, by providing feedback and case number for the user to check for their report progress. *Utilitarianism:* Since this feature does not work with AI, we developers must work under the assumption that a lot of flagged responses carry no racial bias or offensive language. However, there would also be an equally large amount of chatbot responses that warrant the need to improve the AI model. Hence, this feature is necessary to maximise social utility for users. *Deontology:* Make sure the system is designed in a way that can upload ethical duties and address racial bias without sacrificing transparency and privacy. Every user can report bias, and moderators are responsible for reviewing those reports, making sure the AI is aligned with current ethical standards and societal rules.
A community flag verification feature to allow user to verify the legitimacy of flagged chatbot responses implemented with a sorting algorithm to rank them based on popularity.	User must be able to anonymously vote in the forum for their own privacy. Popularity of user votes need to be factored into account.	*Virtue ethics:* Using virtue ethics Kindness Principle, we allow users to access a community forum dedicated to reviewing the AI content flagged by other users and show that we care about them, and we consider their suggestions. We also indirectly foster a community that helps one another improve the chatbot using experience. *Utilitarianism*: Since there are bound to be false flags, implementing this verification community can help maximise social benefit by prioritizing the flagged content based on majority vote. Implementing so that you must be signed in to access the community ensures that most users are genuinely trying to help each other, so the legitimate votes outweigh the troll ones. *Deontology*: No matter the legitimacy of the flagged chatbot responses, the database system will only prioritize value number of flags every day ranked by popularity. This is to ensure that messages collectively deemed as more

		significant by the users will reach the developers first while making sure the system does not overload.
An internal security feature that protects user information via cloud storage and encryption algorithms	It must ensure user data is secure from the public and that users can use the extension without any worry.	*Virtue ethics:* Using Virtue Ethics Kindness and Responsibility Principle, we ensured that user personal data is safe from the public whilst maintaining the usability of the extension by allowing lower-cost access to recently added conversations. We are also ensuring we are transparent in this process by creating a clear outline on how we process the users' data and how we store it. *Utilitarianism:* Although there are overhead costs for implementing the storage system using a cloud platform, after weighing the pros and cons, we decided that social utility is maximised by going through with the plan. Users are satisfied with our product and our product can get more recognition.

Feature summary:



Agile Team Process and Management

The Agile methodology is a project management approach that involves dividing a project into phases or sprints, focusing on continuous collaboration and improvement. Teams work through a cycle of planning, executing, and evaluating, ensuring ongoing refinement and adaptability (Atlassian, n.d.). In an Agile environment, effective collaboration and management are essential for successfully delivering incremental value. Agile is iterative and adaptable unlike the Waterfall methodology that emphasizes linear progression. Agile teams depend on specific tools, clearly defined roles, and collective decision-making processes to achieve their objectives. In the realm of Agile project management, the synergy between teamwork and the Agile environment is vital for realising desired outcomes. This segment explores the significance of teamwork within Agile settings and elucidates the benefits of an Agile environment in facilitating project success; outlining how each component contributes to Agile process development, details the delegation of tasks within the team and explains how decisions are negotiated and agreed upon, ultimately fostering the success of Agile teamwork.

The Use of Agile Tools in Project Management:

In Agile teams, leveraging various tools is critical for streamlining project management, enhancing communication, and ensuring transparency throughout the development lifecycle. Within our team, we predominantly utilise Discord and Trello as core tools to aid us facilitate collaboration and maintain efficiency as we strive to achieve these goals. Additionally, word processing software such as Microsoft Word and Google Docs play an indispensable role in content creation and document management by providing a collaborative space for drafting and editing documents, further enhancing teamwork.

Trello:

Trello is an intuitive visual project management platform that allows teams to track task through visual boards and cards. One of Trello's key features is the inclusion of customisable labels for each delegated role, which inherently reducing confusion, ensuring clear accountability and enables team members to effortlessly track tasks by their respective roles. Moreover, the checklist function allows larger task to be broken down into manageable sub-tasks, making complex projects easier to approach and increasing the productivity of task completion. By categorizing tasks into stages such as "To Do," "In Progress," and "Done," Trello makes it easy to monitor the workflow and task progression. To elaborate further, due dates can be assigned to each card, indirectly motivating the team to stay on schedule. The implementation of due dates assists the team in prioritising tasks and what needs to be carried out first, ensuring timely completion of deliverables (Admin Tomps, 2024). The addition of having a visual organisation fosters collaboration task by providing visibility of ongoing activities, deadlines, and responsibilities collaboration.

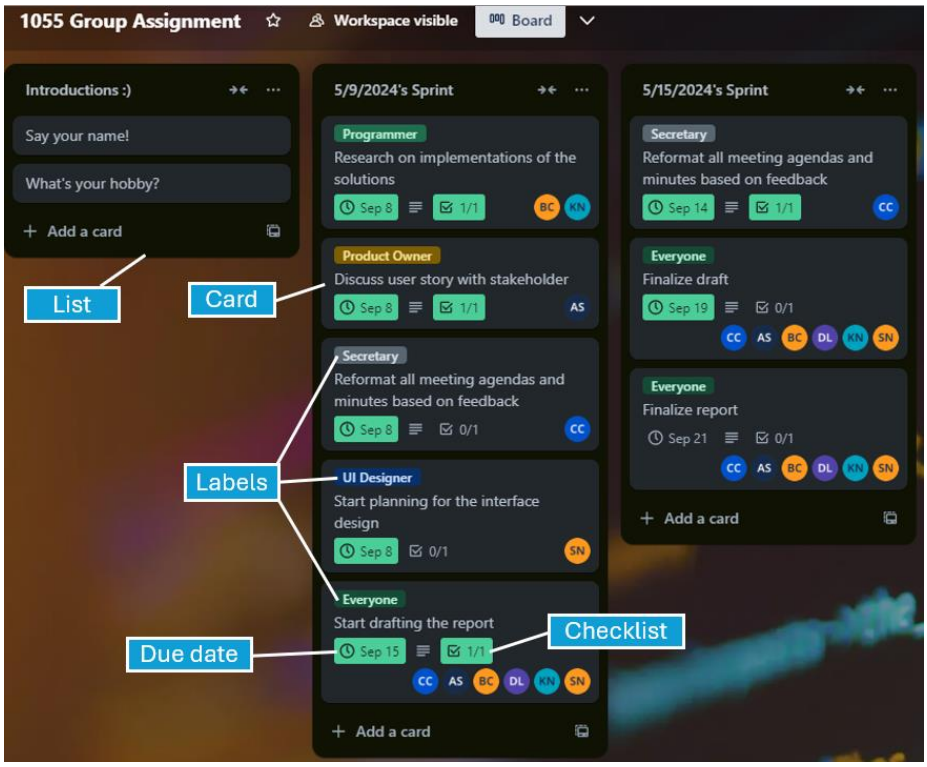


Image of Trello's usage with labels of its features

Discord:

Discord serves as a versatile communication platform that enhances team interactions through both text-based and real-team communication. The amalgamation of voice and video functionalities along with dedicated channels, enables focused discussions on specific topics, reduces clutter and ensures that team conversation remain organized to support dynamic meetings and quick problem-solving sessions. Custom roles can be assigned to each team member, making it easier to identify responsibilities, this inherently reduces confusion or misunderstanding which improves coordination. The platform also facilitates seamless file sharing, empowering team members to upload or access documents swiftly. Furthermore, it serves as a safeguard by ensuring that important files are readily available whenever needed. Discord's mention feature is frequently used to alert specific individuals or groups, ensuring that key updates or announcements are communicated effectively, particularly when urgent action is required by the team leader. By integrating Discord into our Agile workflow, it facilitates continuous communication, enhancing team cohesion and streamlining information exchange, ultimately ensuring that all members are aligned and responsive.

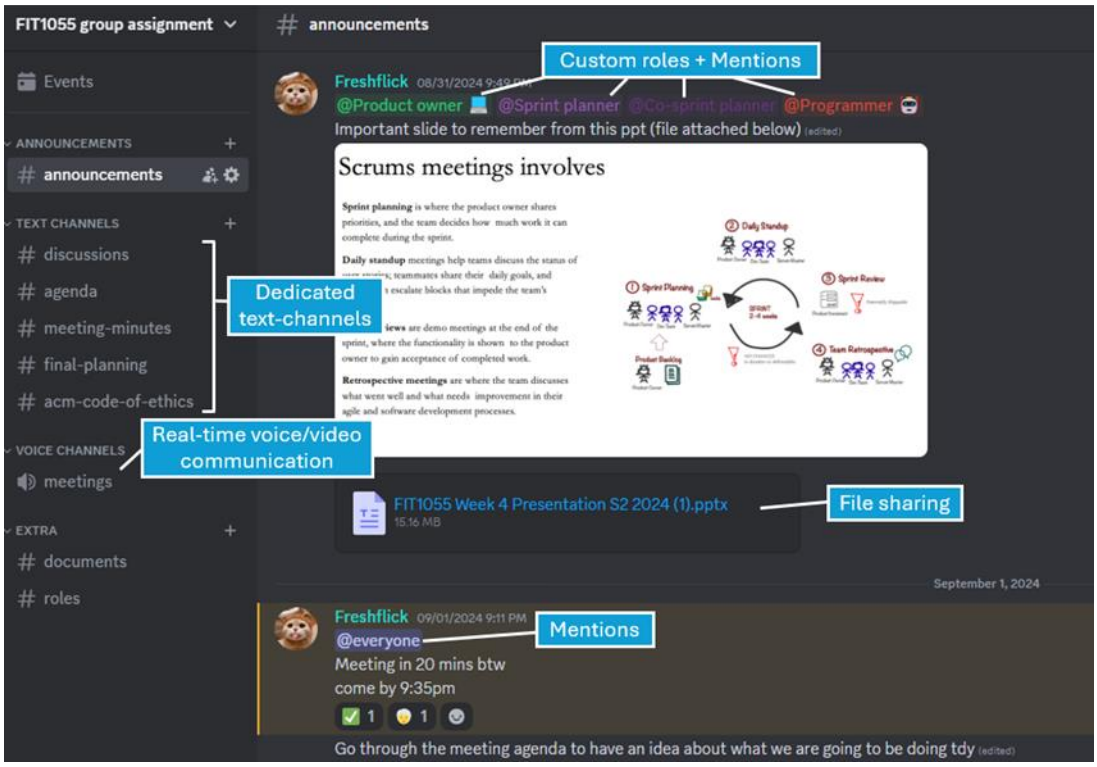
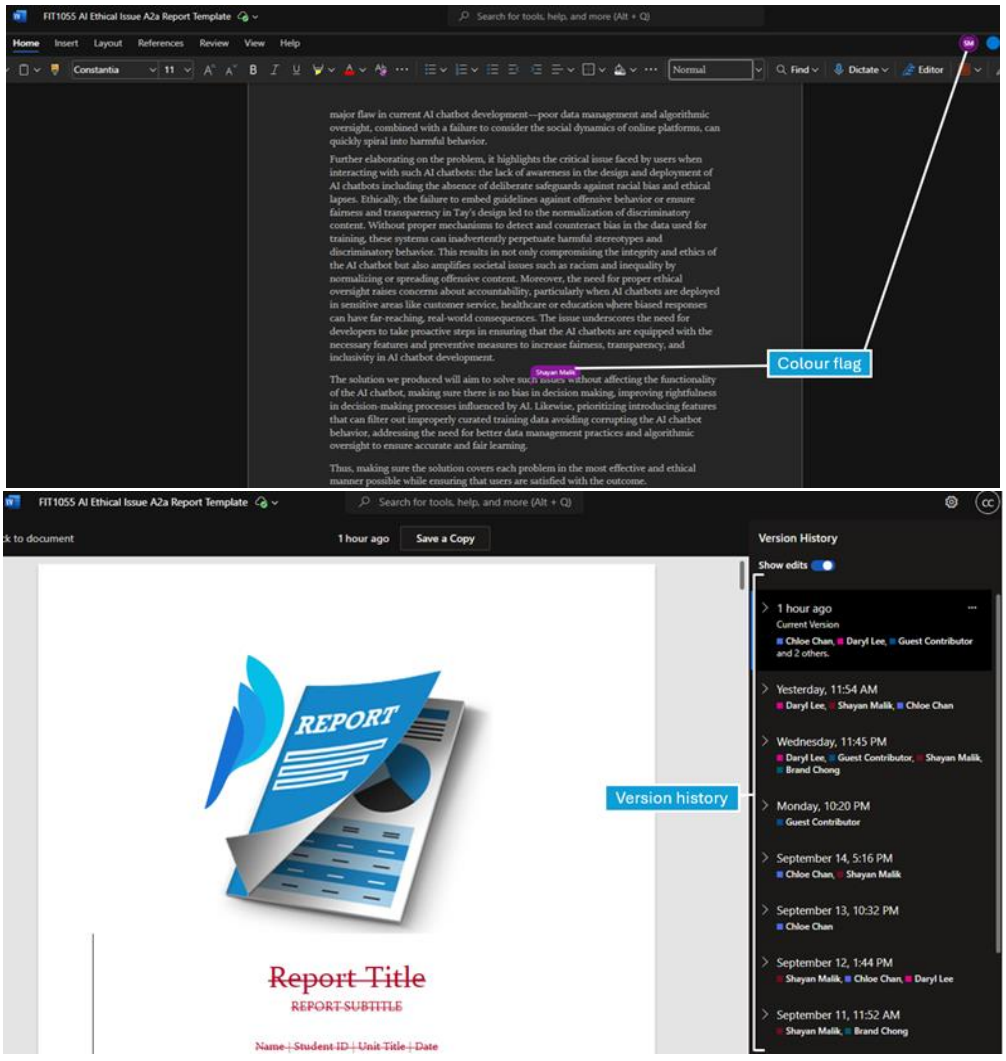


Image of Discord's usage with labels of its features

Word Processing Software: Real-Time Document Collaboration:

Microsoft Word and Google Docs play a crucial role in facilitating document collaboration within Agile teams. These tools allow for interactive real-time co-editing, where team members can simultaneously edit on the same document concurrently, with each contributor's actions precisely marked by coloured flags, indicating the segment each member is currently working (Microsoft, n.d.). This visibility enhances collaboration and ensures that team members can discuss changes in real-time, ultimately producing higher-quality content. Another significant feature of word processing tools is the version history function, which provides the conveniency of being able to track document revisions over time. This is particularly useful for teams working on evolving documents such as ourselves as it allows members alike to revisit previous iterations, review changes made by other members and ensure that no critical content is lost during edits. This level of control makes document management more effective and user-friendly.



Images of Microsoft Word's usage with labels of its features

Team Structure and Task Delegation in Agile:

A team is typically defined as a small group of individuals with complementary skills, working together to accomplish a common mission through interdependent tasks (ASQ, n.d.). In our team, we implement Agile methodologies due to their significant advantages for teams working on dynamic projects that require frequent updates to requirement. Each member is accountable to both themselves and their teammate, as they strive towards a shared objective. Deliverables encompass both individual contributions and collective work products. In an Agile setting, role and task are delegated in a way that promotes equality, accountability, and continuous improvement. Its iterative approach accelerates the delivery of product increments, allowing teams to identify and address issues early in the process

Task delegation in Agile team:

Task delegation in Agile teams is a structured yet flexible process that ensure tasks are aligned with team member's capability, the team's capacity and the project prioritise. Delegation is guided by several core Agile practices:

1. **Sprint Planning**: At the start of each sprint, the Product Owner presents backlog items such as features, user stories, or tasks that need to be addressed. The team collaboratively selects and commits to completing specific items based on their significance, urgency, and time sensitivity. This process ensures that the team stays focused on the most important work while maintaining a manageable workload.

2. **Sprint Review**: At the end of each sprint, sprint review is conducted where the team demonstrates the completed work to the stakeholder. This meeting allows for feedback and discussion about what was accomplished, challenges that were faced and any potential adjustment that might be needed in future sprints. This act fosters transparency and helps ensure that the product aligns with the stakeholder's expectations.

3. **Sprint Retrospective**: Following the sprint review, the team holds a retrospective meeting. This is a dedicated time for each member to voice their concerns and reflect on the sprint, discussing what was accomplished, what could be improved and how to enhance processes moving forward. The goal is to identify actionable steps that can be taken to improve team dynamic and productivity in future sprints.

4. **Weekly Stand-Up Meetings**: During the sprint, task delegation is revisited during weekly stand-up meetings. Team members update the group on their progress, highlighting any roadblocks, and adjust task assignments if necessary. This iterative process allows for flexibility and adaptability, ensuring that the team can respond to any changes or challenges that arise during the sprint.

5. **Self-Organizing Teams**: Agile promotes the idea of self-organising teams, where team members take responsibility for selecting tasks based on their skills, interest, and expertise. Rather than relying on a top-down delegation model, individuals are empowered to volunteer for tasks, fostering a sense of ownership, accountability, and intrinsic motivation. By empowering team members to take control of their work, Agile teams become more passionate and responsive.

6. **Product Owner and Development Team Collaboration**: The Product Owner is responsible for defining the priority of tasks and user stories while the development team has the autonomy to decide how the work will be executed. The collaborative efforts of the teams ensure alignment the business objectives and technical execution. Open dialogue, time-boxed discussions and consensus-building are often used to address complex decisions, preventing lengthy delays and ensuring that the team remains on track.

Roles and Task Delegation:

Agile teams are characterized by mutual accountability, a shared commitment to a unified goal, and often embrace shared leadership. These principles create a collaborative environment where tasks are delegated in a way that maximizes team efficiency and cohesion. In an Agile team, roles are typically assigned based on individual preference and availability, often delegated on a "first come first served" basis. This approach ensures that each team member is working on tasks that aligns with their interests and strength, thereby enhancing job satisfaction and overall productivity. In hindsight this avoids any possible dispute. When team members enjoy the work, they are doing and feel confident in their designated position, it undeniable that the overall productivity of the team improves significantly.

To further streamline task delegation, several new roles may be introduced to break down larger responsibilities into more manageable increments. For instance, roles such as Sprint Planner, Co-Sprint Planner, Reviewer, Secretary, and Co-Lead can help distribute responsibilities more evenly, allowing for clearer task assignments and avoiding potential disputes. Each role holds its own specific set of responsibilities, ensuring that tasks are assigned efficiently and that every aspect of the project is covered.

The finalized list of roles:

Roles	**Owner**
Team lead	Shayan
Co-lead	Chloe
Secretary	Chloe
Product owner	Aaron
Programmers	Khang Wei and Brand
QA testing	Daryl
Reviewer	Aaron
UI designing	Shayan
Sprint planner	Aaron
Co sprint planner	Daryl

Task Delegation Based on Roles:

Tasks are assigned to team members based on their designated roles, which they have chosen based on their expertise and confidence in performing those responsibilities. Since each respective member voluntary hand-pick the role that aligns with their strength, they are more likely to excel in their task, resulting in higher overall team efficiency. Through delegating tasks based on each member's strong point, the team's resources are being utilized to their fullest extent, increasing the team's operation efficiency significantly.

By allowing team members to take on their roles that best suit their skills, the team is able to operate at peak performance as each member understands their responsibilities and is equipped to handle the tasks assigned to them. With the help of the team leader coordinating the team's effort, tasks are completed instinctively and smoothly, without unnecessary delay or obstruction.

Agile leadership:

Agile leadership is characterised by a focus on servant leadership, where leaders (such as Scrum Masters or Agile Coaches) take on a supportive role for the team by removing obstacles, facilitating discussions, and promoting ethical practices. Instead of dictating decisions, leaders empower the team to make decisions collectively. Within our Agile team, leadership is spread out among all team members rather than being limited to formal positions. This leadership approach fosters trust, collaboration, and a self-organized team dynamic, fostering a flexible environment that enables members to thrive. During sprint reviews and retrospectives, all team members are encouraged to provide feedback and propose potential enhancement regardless of their allocated roles in the team. This principle is fundamental to Agile's success as diverse viewpoints are taken into consideration resulting in a group collective decision, this act encourages innovation and accountability at all levels of the team.

People Over Processes:

A core value of Agile leadership is prioritizing people over processes. To further elaborate, the team leader not only prioritizes on the completion of tasks but also take in consideration whether the team members are managing their workload effectively without feeling overburdened or stressed. Leaders provide sufficient time for task completion and regularly check in with team members to ensure they are balancing their work with proper rest.

Responding to Change:

Agile leaders emphasize active listening and adaptability. Before making decisions, the leader consults and facilitates discussions with the team members to ensure that their thoughts and suggestions are considered. This approach fosters a comfortable environment where the team can freely express their ideas, contributing to more thoughtful discussions and informed decisions.

Comfortable with Uncertainties:

Agile leaders remain calm and composed in the face of several uncertainties. Whether it is the uncertainty of completing a task within the deadline or the team's overall progress, a good Agile leader stays focused, seeking and acquiring input from the team and resolving issues as they arise. This helps maintain team morale and ensures that the team stays on track, even through the eye of the storm.

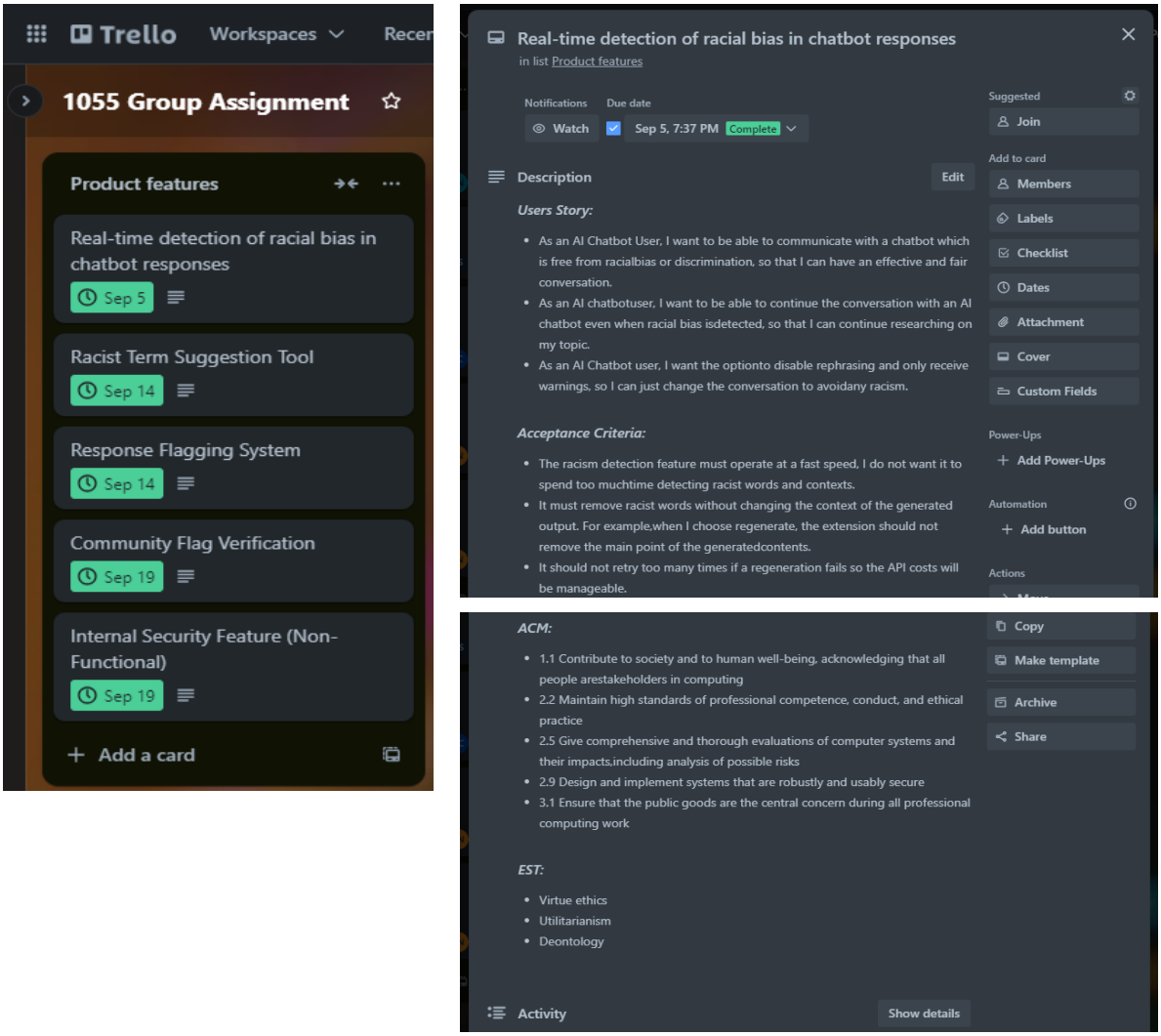
Be the Change and Introduce a Learning Culture:

Agile leaders lead by example, inspiring the team through their actions and attitudes. By setting a positive example and actively engaging with the team, the leader encourages members to stay motivated

and committed to their delegated task. This leadership style promotes a culture of continuous learning and improvement as it boosts the team's morale, leading to higher productivity and improving overall efficiency. In team meetings, the leader encourages open discussions, particularly when issue arise, promoting a healthy learning environment where problems can be addressed collaboratively.

Application of Agile in Ethical Reasoning Framework (ERF):

To encapsulate, integrating Agile principles within an Ethical Reasoning Framework (ERF) ensures that ethical considerations are seamlessly embedded throughout the development lifecycle. Each sprint includes addressing ethical considerations, ensuring that user stories adhere to ethical guidelines. Agile's iterative nature complements ethical oversight, embedding ethical evaluations directly into the workflow, ensuring both innovation and ethical standards are adhered. As ERF becomes an active, ongoing component of the workflow, this approach allows for continuous ethical oversight, feedback-driven adaptations, and ensures that the end product is not only innovative but also ethically sound.



Images of application of Agile principles in ERF

Conclusion

In our report, we produced an applicable solution to attend to the primary issue of AI chatbots having the risk of spreading discriminatory content, which is a browser extension that runs on a text classification model.

The browser extension includes numerous valuable features that aim to address the issue of racist chatbots. As further elaborated in the proposed solution section, it detects racist chatbot responses in real time, and users are provided with the option to moderate the generated controversial content to have a discrimination-free conversion. Additionally, it is equipped with a racism term suggestion tool for users to submit terms and keywords they consider racist. The AI chatbots are then capable of learning from the users' interaction, thereby increasing the model's accuracy in detecting racist content.

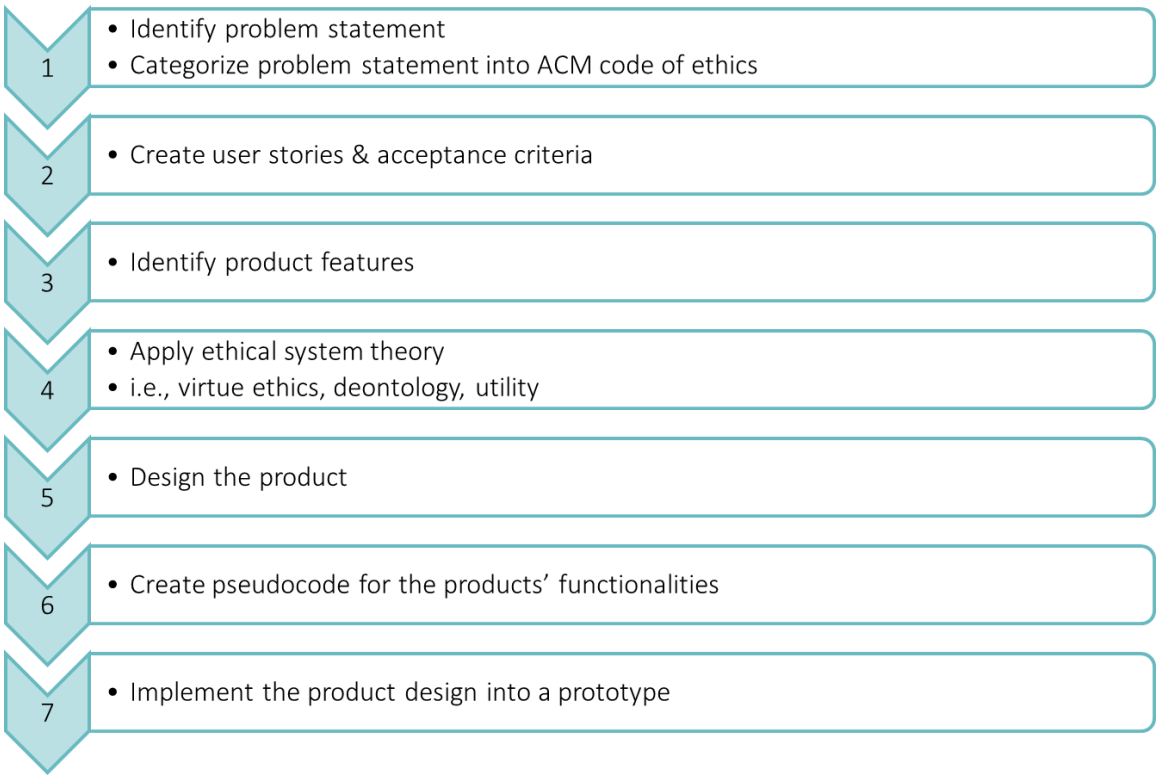
Furthermore, the browser extension comes with a response flagging tool for users to report chatbot responses. Consequently, it prevents users from normalizing the spread of hate and racism. As well as that, it has a database of flagged responses for users to affirm their authenticity. Based on the authenticated responses, the developers can modify the data being fed and the algorithms of the AI chatbots, thus improving users' stereotypes about the developers and their products. Aside from the key features, the browser extension also has a security feature to store conversations by encrypting and storing them in a private location.

The implementation of the solution will significantly change the user's experience when using the AI chatbot. Given that our solution is designed to adhere to the ACM Code of Ethics and Ethical System Theory, the users' well-being is considered, and the public's good is also being considered with caution.

To conclude, producing a browser extension that runs on a text classification model would serve as an effective and feasible solution to address the concern that current AI chatbots are often programmed without sufficient attention to mitigating racial bias and ensuring ethical values, leading to the risk of spreading discriminatory content. It not merely incorporates several practical features, but furthermore complies with the ACM Code of Ethics and Ethical System Theory.

Appendix: ERF

Ethical Review Framework (ERF) was applied throughout the project within Agile practices to ensure compliance with ethical guidelines. The flowchart below illustrates the application of ERF within basic software development process that the team referred to for the project:



ERF Flowchart

References

ASQ. (n.d.). *What is a Team?* https://asq.org/quality-resources/teams

Atlassian. (n.d.). *What is Agile? | Atlassian*. https://www.atlassian.com/agile

Battogtokh, M., & Mehla, V. (2023, November 8). *How can you clean text data with misspelled words for machine learning?* Linkedin.com. https://www.linkedin.com/advice/1/how-can-you-clean-text-data-misspelled-words-machine-6h8zf

Benítez-Andrades, J. A., González-Jiménez, Á., López-Brea, Á., Aveleira-Mata, J., Alija-Pérez, J.-M., & García-Ordás, M. T. (2022). *Detecting racism and xenophobia using deep learning models on Twitter Data: CNN, LSTM and Bert*. *PeerJ Computer Science*, *8*. https://doi.org/10.7717/peerj-cs.906

Buranyi, S. (2017, August 8). *Rise of the racist robots – how AI is learning all our worst impulses*. The Guardian. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses

C3.ai. (n.d.). *Ethical AI: 10 pillars of AI transparency.* C3 AI. https://c3.ai/glossary/artificial-intelligence/ethical-ai/

Hilliard, A. (2023, July 4). *What is Ethical AI?* www.holisticai.com. https://www.holisticai.com/blog/what-is-ethical-ai/

Hsu, J. (2024). *AI chatbots use racist stereotypes even after anti-racism training*. New Scientist. https://www.newscientist.com/article/2421067-ai-chatbots-use-racist-stereotypes-even-after-anti-racism-training/

Hunt, E. (2016, March 24). T*ay, Microsoft's AI chatbot, gets a crash course in racism from Twitter.* The Guardian. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

IBM. (2023). *What is a chatbot? | IBM.* IBM. https://www.ibm.com/topics/chatbots

Schwartz, O. (2019, November 29). *In 2016, Microsoft's racist chatbot revealed the dangers of online conversation*. IEEE Spectrum. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

Isbell, A. R., & Stroud, S. R. (2018, September 20). *Tay AI: Ethical challenges in the use of AI chatbots.* Center for Media Engagement. https://mediaengagement.org/wp-content/uploads/2018/09/21-tay-ai-case-study-1.pdf

Jindal, A. (2024, February 12). *Misguided artificial intelligence: How racial bias is built into clinical models.* Black History Month Scholarly Articles. https://bhm.scholasticahq.com/article/38021-misguided-artificial-intelligence-how-racial-bias-is-built-into-clinical-models

Keserer, E. (2022, December 5). *The five main subsets of AI: (Machine learning, NLP, and more).* Akkio. https://www.akkio.com/post/the-five-main-subsets-of-ai-machine-learning-nlp-and-more

LeadDesk. (2022, August 24). *Chatbot use cases: 25 real-life examples.* LeadDesk. https://leaddesk.com/blog/chatbot-use-cases-25-real-life-examples/

Microsoft. (n.d.). *Collaborate in Word.* https://support.microsoft.com/en-us/office/collaborate-in-word-b3d7f2af-c6e9-46e7-96a7-dabda4423dd7

Nicoletti, L., & Bass, D. (2023, June 9). *Humans Are Biased. Generative AI Is Even Worse.* Bloomberg. https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Piers, C. (2024, February 7). *Even ChatGPT Says ChatGPT Is Racially Biased.* Scientific American. https://www.scientificamerican.com/article/even-chatgpt-says-chatgpt-is-racially-biased/

Sapardic, J. (2024, September 24). *What Is a Chatbot and Why Chatbots are Important for Small Businesses?* Tidio Live Chat; Tidio Live Chat. https://www.tidio.com/blog/what-is-a-chatbot/

Schulz, B. (2024, April 5). *Is AI racially biased? Study finds chatbots treat Black-sounding names differently.* USA TODAY. https://www.usatoday.com/story/tech/2024/04/05/ai-chatbot-chatgpt-racial-bias/73206637007/

Stryker, C., & Kavlakoglu, E. (2024, August 16). *What is artificial intelligence (AI)?* IBM. https://www.ibm.com/topics/artificial-intelligence

3Blue1Brown. (2017, October 5). *But what is a neural network? | Chapter 1, Deep learning [Video].* YouTube. https://www.youtube.com/watch?v=aircAruvnKk

3Blue1Brown. (2024, April 7). *Attention in transformers, visually explained | Chapter 6, Deep Learning [Video].* YouTube. https://youtu.be/eMlx5fFNoYc?si=gnoP1JZud1bvZLoe

Tomps, A. (2024, January 12). *Due date: Types and advantages in the world of work.* Tomps. (n.d.). https://www.tomps.id/en/due-date-jenis-dan-kelebihannya-dalam-dunia-kerja-en

Wakefield, J. (2016, March 24). *Microsoft chatbot is taught to swear on Twitter.* BBC News. https://www.bbc.com/news/technology-35890188