Meeting Agenda and Minutes

Location:	Mind Link Deck Level 3 - LD3-1
Date:	4/4/2023
Time:	9:40 a.m. - 10:40 a.m.

Participants and Absentees:

Present: Diong Chen Xi, Kang Hong Bo, Brandon Lau, Lana Malika Binti Hussayni, Shunnosuke Takei
Absent: None

Meeting Agenda:

Agenda	Expected Time Taken
Appoint a person in charge of taking down meeting minutes	2 Minutes
Discuss and Decide on the Team’s Ground Rules	10 Minutes
Conduct a personality test	20 Minutes
Discuss and decide expectations and roles based on the personality test results	10 Minutes
Decide on Assignment 2 topic:	10 Minutes

Agenda details:

I. Appoint a person in charge of taking down meeting minutes:

After having a majority vote, we decided to make Chen Xi the person in charge of taking down the meeting minutes.

II. Discuss and decide on the team's ground rules:

We had a discussion session about the team's ground rules and noted them down on the group contract file provided in Moodle.

III. Conduct a personality test:

We conducted a personality test using the recommended website, and obtained the following results:

Chen Xi - Type 1 The Perfectionist

Hong Bo - Type 9 The Peacemaker

Brandon - Type 7 The Enthusiast

Lana - Type 8 The Challenger

Shun - Type 3 The Achiever

IV. Discuss and decide expectations and roles based on the personality test results:

After a majority vote, we decided on the team leader to be Chen Xi and Shun to be the recorder.

We also had a discussion about how to split the remaining roles, and decided to let group members pick roles that they feel comfortable working as. The results are as follows:

Group Leader: Chen Xi

Recorder: Shun

Checker: Brandon

Innovator and Technologist: Hong Bo

Researcher: Lana

Chen Xi proposed that the roles of Researcher, Idea Generator, Checker and Technologist are to be shared within the group as everyone is responsible for their own by doing sufficient research to understand the assignment specifications and requirements, therefore the roles are only meant for that person to be in charge of ensuring everyone does their tasks for that phase. All of us agreed to this suggestion.

V. Decide on Assignment 2 topic:

After a small discussion, Only two of us finished the Assignment, where we obtained two possible topics of research:

Chen Xi: Misinformation and Disinformation using Chatbots (ChatGPT)

Hong Bo: DeepFakes

We tentatively decided that we are going to focus on the second topic, ‘Misinformation and Disinformation using Chatbots’. We came up with a possible solution of creating a website that serves as a forum for chatbot users to raise questions and problems, and for developers to provide assistance as well as acquire user feedback.

As most of us have not yet finished Assignment 1, we decided to have another discussion about the topic after all of us have finished working on the individual report. This ensures that we have a clear direction as to what to work on in Assignment 2a.

Action Items:

Member(s)	Action Item	Deadline	Status
Everyone	Finish Assignment 1 and make a short summary	30th April 2023	Completed

Meeting Agenda and Minutes

Location:	https://monash.zoom.us/j/89699280630?pwd=a3Q0RExzUXNONVFRb2syZUFVTFg1UT09
Date:	30th April 2023
Time:	15:00 – 16:15

Participants and Absentees:

Present: Diong Chen Xi, Kang Hong Bo, Lana Hussayni, Sia Yi Bin (New member), Shunnosuke Takei

Absent: Brandon Lau

Meeting Agenda:

Agenda	Expected Time Taken
New Member Introduction	2 Minutes
Share brief summary of A1 findings	30 Minutes
Re-decide on topic for A2:	20 Minutes
Discuss and decide expectations and roles based on the personality test results	10 Minutes
Decide on Assignment 2 topic:	10 Minutes

Agenda details:

I. New member introduction:

We welcomed a new member, Yi Bin to our team. He transferred to our team due to issues with his previous team, and all of us agreed on letting him join the team. Due to the addition of a new member, we decided to split one of the roles of Hong Bo to Yi Bin.

Our final role delegation is as follows:
Group Leader: Chen Xi
Recorder: Shun
Checker: Brandon
Innovator: Yi Bin
Technologist: Hong Bo
Researcher: Lana

II. Share brief summary of A1 findings:

Each member took a few minutes to share a summary of their findings from the assignment 1 report, which have been noted down below:

Deepfakes, Images generated without the consent of the subject person (Lana)
1. Common methods used to produce DeepFakes are Autoencoder Neural Networks and Generative Adversarial Network. Due to community efforts, these techniques are becoming more accessible to non tech people.
2. Issue at hand is that the methods used to make DeepFakes are not only accessible but are becoming less complicated. It is very difficult to regulate who can make what because of this.
3. Action of detection deepfake techniques to identify images that are fake to limit misinformation of the subject person of generated image

Influencing, disinformation, and fakes problems caused by Deep Fakes (Hong Bo)
1. Generative adversarial network (GAN) leads to the technique of DeepFake.
2. Disinformation and Misinformation created by deep fake technique.
3. Action of detection of deepfake techniques

Misinformation/disinformation from ChatGPT (Shun)
- ChatGPT is trained on large datasets with potentially outdated/inaccurate/biased data
- Generates misinformation/disinformation, its spread is further aggravated by its user base
- Its responses are not self-validated and must be fact-checked by the user themselves
- Potential to be used maliciously against organizations/businesses/individuals

Misinformation and Disinformation in ChatGPT (Chen Xi)
- ChatGPT is very efficient and widely accessible

- Trained using web data, and has no self-validation algorithm, prone to giving misinformation
- ChatGPT is frequently used for queries, and its responses are not always validated by the user.
- Used to spread disinformation due to its prowess to generate human-like texts, as well as mimic a certain person's writing style and speech tone.
- if used maliciously, ChatGPT may be abused to inject policy bypassing prompts to make false accusations, impersonate authorities etc.

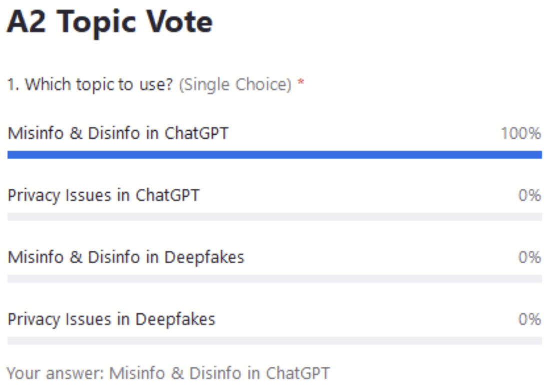
Decoding the algorithmic framework of ChatGPT: An ethical analysis of its language generation and decision making processes (Yi Bin)
- The vast amount of data used to train LLMs makes it difficult to filter out offensive or inaccurate content. -the lack of understanding of deep learning and neural networks raises ethical concerns, as it may lead to unintended consequences and make it challenging to hold the model accountable for its actions.
- ChatGPT employs Reinforcement Learning with Human Feedback (RLHF) to incorporate human feedback and generate more natural and relevant responses. While this approach may improve response quality, it also raises concerns about privacy and the use of human data.

III. Re-decide on topic for A2:

Since our last meeting, 2 new topics have been proposed, increasing the total count to 4. The new selection of main topics are:

1. Privacy Issues in ChatGPT
2. Misinformation/disinformation in ChatGPT
3. Misinformation/disinformation in DeepFake
4. Privacy Issues in DeepFake

After holding a majority vote, all members unanimously agreed to keep the topic ‘Misinformation and disinformation in ChatGPT’. Our solution idea remains unchanged.



A2 Topic Vote

1. Which topic to use? (Single Choice) *

Misinfo & Disinfo in ChatGPT	100%
Privacy Issues in ChatGPT	0%
Misinfo & Disinfo in Deepfakes	0%
Privacy Issues in Deepfakes	0%

Your answer: Misinfo & Disinfo in ChatGPT

IV. Delegation of tasks:

We decided on 3 factors to conduct research on regarding the topic through a vote, as well as 3 ACM codes ChatGPT has violated, with 2 members working on one topic/code. The grouping is as follows:

Factors:

1) **ChatGPT has no self-validation algorithms (HongBo/Brandon)**
2) **ChatGPT is trained on large datasets that are prone to inaccuracies/biases (Shun/Lana)**
3) **ChatGPT is prone to policy-bypassing prompt injection (YiBin/ChenXi)**
4) ChatGPT lacks contextual understanding and may generate responses that are not appropriate or accurate for the given context.
5) ChatGPT can be intentionally manipulated by individuals or organizations to spread misinformation or disinformation for their own purposes.

ACM codes violated:
1.2 Avoid Harm (Shun/ChenXi)
1.3 Be honest and trustworthy
1.4 Be fair and take action not to discriminate (Lana/Brandon)
1.6 Respect privacy
2.2 Maintain high standards of professional competence, conduct, and ethical practice
2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks. (YiBin/HongBo)
2.7 Foster public awareness and understanding of computing, related technologies, and their consequences.
3.7 Recognize and take special care of systems that become integrated
into the infrastructure of society.

For the next meeting, all group members were to come with the research/analysis completed.

Action Items:

Member(s)	Action Item	Deadline	Status
HongBo & Brandon	Research on Factor ChatGPT has no self-validation algorithms	6th May 2023	Completed
Shun & Lana	Research on Factor ChatGPT is trained on large datasets that are prone to inaccuracies/biases		Completed

Yi Bin & Chen Xi	Research on Factor ChatGPT is prone to policy-bypassing prompt injection		Completed
Shun & Chen Xi	Research on ACM Code 1.2		Completed
Lana & Brandon	Research on ACM Code 1.4		Completed
Yi Bin & Hong Bo	Research on ACM Code 2.5		Completed

Meeting Agenda and Minutes

Location:	https://monash.zoom.us/j/89225329442?pwd=a2U5UnFQUFlaMmw4ZDRnZkNmRnVEUT09
Date:	6th May 2023
Time:	15:00 – 17:00

Participants and Absentees:

Present: Brandon Lau, Diong Chen Xi, Kang Hong Bo, Lana Hussayni, Sia Yi Bin, Shunnosuke Takei

Absent: None

Meeting Agenda:

Agenda	Expected Time Taken
Share findings from tasks allocated prior (Factors and ACM Codes)	1 hour
Working on the report	1 hour
Delegation of Tasks	5 Minutes

Agenda details:

I. Share findings from tasks allocated prior:

Each member took a few minutes to share a summary of their findings from the tasks distributed during our last meeting:

Factors
ChatGPT has no self-validation algorithms (HongBo/Brandon)
ChatGPT is trained on large datasets that are prone to inaccuracies/biases (Shun/Lana)
ChatGPT is prone to policy-bypassing prompt injection (YiBin/ChenXi)

ACM Codes
1.2 Avoid Harm (Shun/ChenXi)
1.4 Be fair and take action not to discriminate (Lana/Brandon)
2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks. (YiBin/HongBo)

Factor: ChatGPT has no self-validation algorithms

1. Researchers
The validity of a research study refers to how well the results among the study participants represent true findings among similar individuals outside the study. This concept of validity applies to all types of clinical studies, including those about prevalence, associations, interventions, and diagnosis.

2. Pharmaceutical industry
The pharmaceutical industry is highly regulated, and ensuring the quality and safety of drugs is of utmost importance. One aspect of this process is the validation of computerized systems, which is necessary to ensure that these systems are reliable, secure, and accurate.

Factor: ChatGPT is trained on large datasets that are prone to inaccuracies/biases

1. OpenAI
Carrying out training on datasets that may be inaccurate/biased is harmful for the future of the development of their GPT model. Developing an incompetent AI model is the last thing the development team wants, and must take measures to weed out those imperfections in their training process. Leaving the issue unresolved could also result in OpenAI losing partnerships and sponsors (most notably Microsoft), which will be very detrimental to the organization as a whole.

2. Users of ChatGPT and society as a whole
ChatGPT being trained on inappropriate datasets means its output will be affected negatively, being more prone to generating misinformation/disinformation. Which means it will affect the users as well as they are being fed that information. It is very likely for users who are unaware to take information generated by the chatbot at face value and assume it as factual, which can be very detrimental for society as a whole. Furthermore, its users can aggravate the spread of misinformation further through human interaction, social media, etc. Given the fact that so much misinformation is already circulating around the internet, this prospect will only escalate the current situation further.

ChatGPT is prone to policy-bypassing prompt injection

1. Community users (Educational e.g. StackOverFlow, Gaming etc.)
ChatGPT can be prompted to give convincing but inaccurate answers, confusing users on the forum. ChatGPT's responses may also be used as proof which in reality isn't always true.

2. Human Population
ChatGPT can be prompted to write about almost all exploitative behaviors, as well as promote illegal activities such as self-harm, murder, substance abuse, rape, etc.

3. Policy makers
Controversies emerge on the topic of to what extent speech freedom can be given to a Generative AI model. Also if a user who does not have prior knowledge uses ChatGPT to learn about how to commit a crime and does it, who is to blame?

4. Celebrities
ChatGPT can be used to write realistic sounding fake news, which can harm a celebrity's reputation.

5. Programmers and IT Developers
ChatGPT can give ways of exploiting or hacking into systems, applications, devices etc., teaching users how to compromise security, this challenges developers to implement better security and exercise extra caution.

6. Government
ChatGPT can be used to write stories in a certain style, aiming to defame politicians and governmental figures, potentially sparking unnecessary conflicts.

7. OpenAI
ChatGPT's production team may be held responsible for misuses of the tool, i.e. sued by victims and demanded compensation. Competitors may abuse ChatGPT to gain unauthorized insights on the product, or the company as a whole.

ACM Code 1.2 Avoid Harm

1. Cause harm to systems, applications and devices. Users can prompt ChatGPT to write malware.
2. Cause harm to people's reputation. It is capable of writing convincing slanders, generating fake news reports and mimicking writing styles to write sexist and racist remarks. ChatGPT has also had instances where it generated fake quotes said by real people (and claiming them as true), which can be severely damaging to one’s reputation, especially in areas such as politics and finance.
3. Cause harm to individuals. It can write phishing emails by imposing officials to scam individuals of their private information and money. It could also be used to influence investments in the stock market by tricking individuals into doing so using misinformation generated by the AI. Individuals could capitalize on this for financial gain. ChatGPT has also shown signs of racism, sexism, and other controversial/biased views in its generated output. This can influence users to have those problematic views, which can indirectly result in harm and/or discrimination of certain groups.

ACM Code 1.4 Be fair and take action not to discriminate

1. Discriminates against individuals of different races, gender etc. ChatGPT has generated output that was very racially discriminative, specifically discriminating black from white people. This reflects the situation in real life, and is a disturbing insight into the AI’s decision-making model.

ACM Code 2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.

1. Incomplete or Inaccurate Analysis: ChatGPT may generate responses that lack comprehensive analysis of potential risks associated with computer systems. This could be due to limitations in its programming or training data, which may not cover all possible scenarios or risk factors.
2. Bias or Prejudice: ChatGPT's responses may reflect the biases or prejudices of its training data or the people who created it. This could lead to incomplete or inaccurate evaluations of computer systems and their impacts, particularly in areas such as fairness, privacy, and security.
3. Limited Scope: ChatGPT's training data and programming may be limited in scope, which could lead to incomplete or inaccurate evaluations of computer systems and their impacts. For example, it may not be able to account for the full range of potential risks associated with emerging technologies or complex systems.

II. Working on the report:

After the discussion, members started working on the report itself, mainly focusing on 2.1, 2.1.1 and 2.2.

III. Delegation of tasks:

All group members were to research and determine which ethical theories to utilize to write our report. In addition, members were to think about possible solutions to address the problem statement. Research results and ideas are to be shared at the next meeting.

Action Items:

Member(s)	Action Item	Deadline	Status
Everyone	Read the textbook on Ethical Theories	11th May 2023	Completed
Everyone	Think of ideas to possible solutions		Completed

Meeting Agenda and Minutes

Location:	https://monash.zoom.us/j/86173673971?pwd=M1VDaFBRRzNRTFI1UC80VFpBWmp1QT09
Date:	11th May 2023
Time:	20:00 – 23:30

Participants and Absentees:

Present: Brandon Lau, Diong Chen Xi, Kang Hong Bo, Lana Hussayni, Sia Yi Bin, Shunnosuke Takei

Absent: None

Meeting Agenda:

Agenda	Expected Time Taken
Share findings from tasks allocated prior (Virtue Ethics)	30 Minutes
Working on the report	2 hours

Agenda details:

I. Share findings from tasks allocated prior:

Group members shared their findings regarding the ethical principles that could be used as a guideline to proceed with our report. Our findings are compiled below as follows:

Virtue Ethics:

1. (Honesty) Generative AI models should only be trained using true data, and remind users to fact check in the case where it is unsure of the credibility of its response.
2. (Honesty and Responsibility) Open AI has the responsibility to show the right information. It shouldn't give information that is not updated
3. Honesty: OpenAI should maintain a level of transparency in their development process. This includes the sources of their training and testing datasets, fixes, etc.
4. (Security) ChatGPT should prioritize the security of user data and take appropriate measures to protect against unauthorized access or data breaches. This includes implementing encryption and other security measures to safeguard user data.

Utilitarianism:

1. When queried for illegal responses such as money laundering or computer hacking, request for user verification and send to authorities for review.
2. Making programs online (i.e. synchronous with real time) instead of having stagnant data. Outdated information mostly equals false information.
3. Flag responses that contain claims that have been debunked by fact-checking organizations or have been identified as false by reliable sources of information.

Deontology:

1. Generative AI models should refuse to give responses to prompts for illegal activities, no matter what the purpose is.
2. Generative AI can provide alternative responses whenever possible
3. Generative AI shouldn’t influence its users to perform morally wrong actions.
4. Generative AI shouldn't compile false responses to mislead users.
5. Generative AI should prioritize transparency and interpretability, which provide a clear and transparent communication to users about how the AI model works.
6. Generative AI should not lie, and if it detects a question that has not been learned, it should honestly tell the users that it cannot provide an accurate answer.

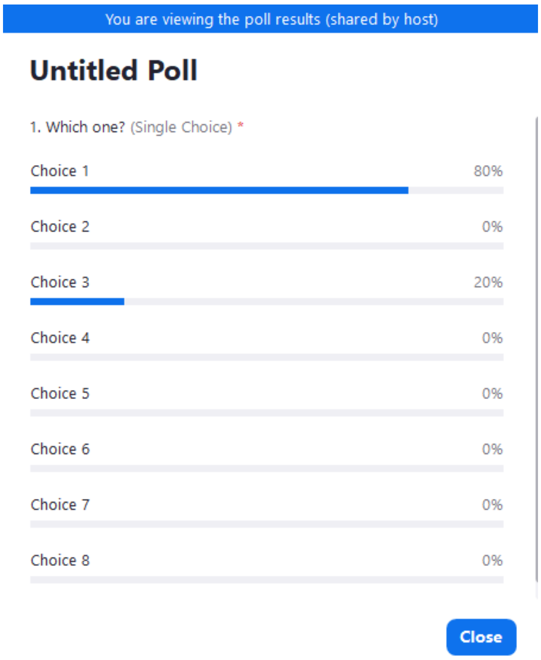
II. Working on the report:

After the discussion, we began working on completing our report, starting with the ideate step, where group members proposed multiple solutions to the problem. The proposed solutions are listed below:

1. Forum website
2. Online AI engine
3. App that rewards users in return for flagging issues
4. Data sanitiser
5. Live monitoring of user prompts
6. Implement paywalls
7. Fake news detector
8. Restriction of user input

The solutions are discussed in detail in the report. A final majority vote was held to confirm whether we were sticking to the original solution (solution 1) or changing to a different one.

Below is the result of the vote:



You are viewing the poll results (shared by host)

Untitled Poll

1. Which one? (Single Choice) *

Choice 1	80%
Choice 2	0%
Choice 3	20%
Choice 4	0%
Choice 5	0%
Choice 6	0%
Choice 7	0%
Choice 8	0%

With our choice of topic finalized, we began working on the remainder of the report, including the deliberation step, proposition of our solution, and conclusion.

Action Items:

None. This is our final meeting.