



Lo-Fi Prototype Report

Ahana Banerjee, Clarise Liu, Iris Liu, Edgar Angeles, Alice Wen
Team D2

Table of Contents

01

Executive Summary

Examining the undertaken activities, key insights obtained, and outlining future steps.

02

Describe the Test and Prototype

Analyzing the specifics of the prototype implementation.

03

Describing Success and Failure

Presenting our honest signals, what we learned about our assumptions and the risks we identified

04

Outlined Proposed Changes

We reflect on design changes to make and the rationale and reasoning behind them.

05

Describe Next Steps

How would we move forward? What would the next prototype look like in terms of the risk it is testing, metrics of measurement, and dimensions?



Executive Summary

Executive Summary

For our lo-fi prototype, our group made a Figma file that adds two functionalities to the typical chat-based generative AI interface: a community forum for users to post about harmful AI bias, and a button in the bottom right corner to chat with a developer of the application. We focused on prototyping the desired path a user would take of going to the community tab, making a post, and interacting with a software developer. Our riskiest assumption was that users would be willing to engage with online strangers while using generative AI.

We tested our prototype on 5 college-age individuals, each with various exposure to generative AI, during scheduled semi-structured interviews. Through our sessions, we were able to understand how our approach to community-based interactions realistically influences individual motivations to report and acknowledge harmful AI behaviors. 80% of interview participants observed positive benefits from our prototype, and that it effectively improves the digital community and quality of AI reports.

After our interviews, our team synthesized our results and discussed remaining design goals. Our discussion revealed several changes we will make for our future prototypes, all of which aim to customize the experience to each user's needs. By making a more robust prototype that incorporates a search button, moderated posts and sorting, we aim to enhance user engagement and provide a more tailored experience. Increasing the fidelity of our prototype will also allow for more organic engagement, thus leading to more valuable insights.



Describing the Test and Prototype

Describing Test & Prototype

WHO: We tested our prototype with 5 individuals. Our participants were all college-educated and had various online experiences and motivations with generative AI.

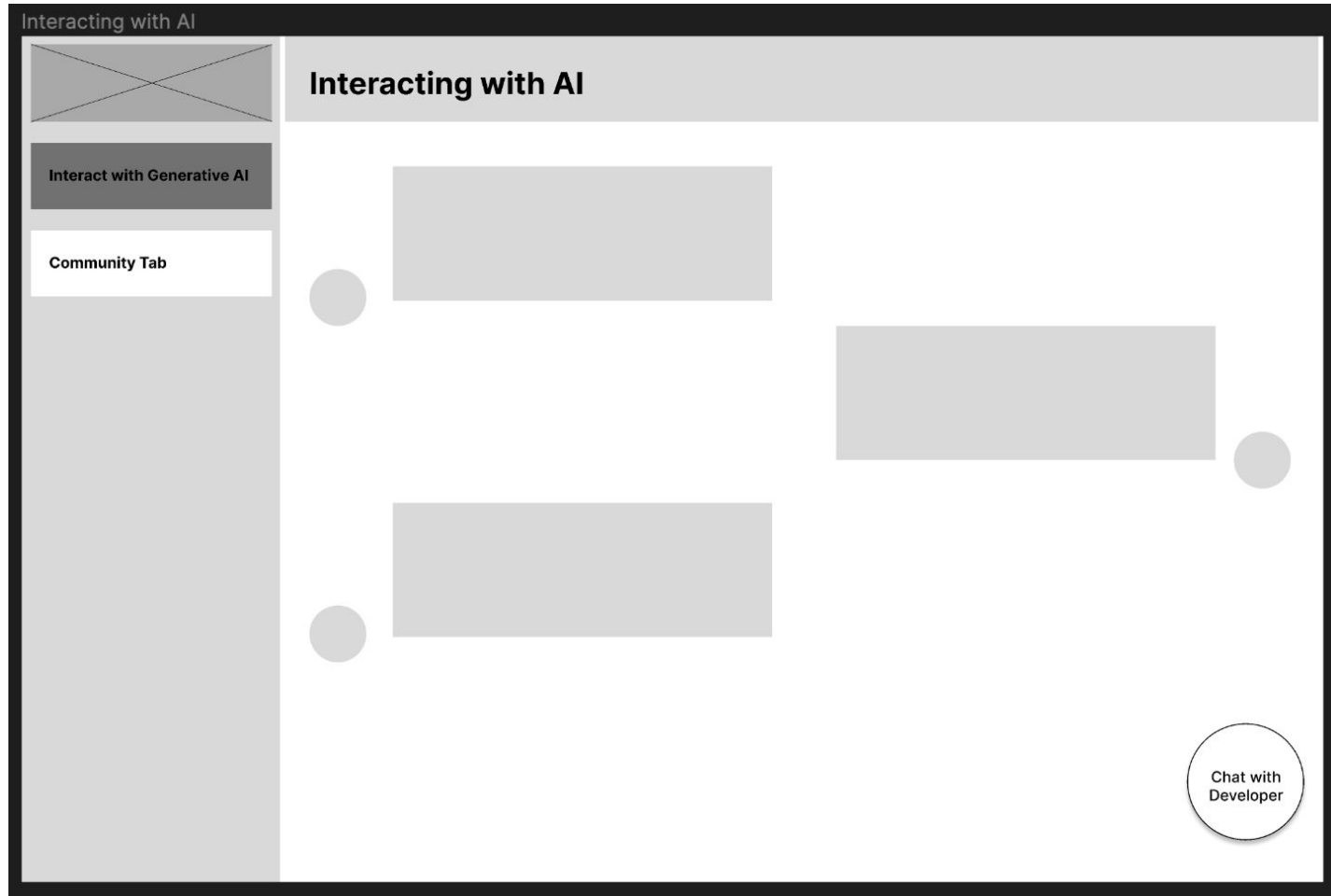
WHAT: Our lo-fi prototype includes a sidebar to navigate between AI activity and a community tab for discussions about genAI bias. Users can post threads, flag harmful biases, and chat with developers. From our research, we anticipate users seeking reassurance and impact when reporting biases and gaining insights into identifying harmful and unhelpful biases within the community. While conducting interviews, we also had our risk assumptions in mind and hoped for users to help us gain insight on what is working well with the prototype and what can be improved.

WHERE: These participants were gathered individually in quiet spaces for our interviews.

HOW: Each team member schedule a time with someone they know who has used ChatGPT in the past. During the interview, they were then presented with our low fidelity prototype. Then they were asked to perform various tasks to complete the goal of creating a thread about bias. With each step, interviewees were asked questions about their experience and thoughts.

WHY: We opted for this in person, one-on-one approach because we wanted to observe and uncover what people's thoughts were as they went through our prototype. This would also allow us to see if people seem confused on how to traverse our prototype and at the end we could ask for improvements to the current iteration.

Prototype Step 1: Navigate to Community Forum




Prototype
Step 2:
Click on the
“Create a
Thread”
button



Prototype Step 3: Fill Out Necessary Information

Creating a Thread



Interact with Generative AI

Community Tab

Community Tab

Create Thread

Title

Type of Bias ▼

Description

Post

Prototype Step 4: Post Thread

Confirmation

Interact with Generative AI

Community Tab

Community Tab

Bias

Category

Category

Create Thread

Congrats! Your Thread Has Been Posted!

Would you also like to speak with a AI developer representative?

Yes

No

Harmful

Unharmful

Chat with Developer



Describing Success and Failure

Honest Signals

Assumption: Users will know how to report + how they want to report

- ▶ Success: Participants know whether they would want to post on the forum or speak to a developer, when given the choice
 - ▷ They had different reasoning, but all knew where they would go and why (for example, they might not talk to a dev because they're worried that there's too many people trying to talk to them, but would be okay posting their thoughts on a forum)
 - ▷ We want to keep options open so users can choose the method most intuitive for them

Assumption: Communities will incentivize users to report biases

- ▶ Success: Those who didn't necessarily want to talk to a representative would have been okay with posting in a community
 - ▷ Users described the forum with language such as “validating” and felt as if posting allowed them to express their thoughts and be seen by others
 - ▷ Also felt more impactful as they knew others could see and interact
- ▶ Failure: Some users did not see a need to engage with community and still saw their interactions with AI as very individual

Overall, most of our assumptions were validated. However, there are still some failures, as well as metrics and risks we were not able to test with this prototype. We also don't know how well our assumptions would hold up in a more natural setting, which is where we want to aim for in the end. This would require a more refined prototype and a setting in which users are already using GenAI for unrelated reasons.



Outlining Proposed Changes

Proposed Changes

Moderation: Ensure the threads are vigilantly monitored by a dedicated moderator from our generative AI developer team. This oversight guarantees the integrity and quality of discussions on the platform.

Search Functionality: Include a search bar, empowering users to effortlessly discover specific types of bias that resonate with their interests and preferences.

Sorting Options: Offer users a spectrum of sorting alternatives tailored to their needs. From sorting threads by popularity and recency to assessing their level of harm or lack thereof, users can personalize their content consumption experience.

Filtering and Saving Topics: Grant users the ability to bookmark and save topics of personal interest. This feature fosters ongoing engagement and encourages users to participate actively in discussions that matter most to them.

Thread Examples: Enhance user guidance by implementing standardized filler text examples when users initiate threads. This ensures consistency and elevates the overall quality of conversations across the platform.

Visual Content Integration: Enrich user experience by allowing seamless integration of images into posts. This not only facilitates quicker content absorption but also alleviates text overload, enhancing overall engagement and satisfaction.

Communication with Developers: Optimize communication channels by transitioning from traditional calls to text-based interactions with developers for issue reporting. Additionally, offer users the flexibility to schedule calls when necessary.



Next Steps

Risk, Dimensions, Metrics

After synthesizing the interview data from our first prototype, our team realized that users need an inherent motivation. In particular, we received a lot of feedback on how some features are invasive on day-to-day interactions with AI. From our testing we have a lot of feedback related to how some features seem to be invasive for them to use in their day to day. They also mentioned that they do not seem motivated to report in general.

So here we want to create a new prototype that wishes to address this risk. Here visually we would like to improve the fidelity to make the tasks laid out less obvious. This is to show that this prototype is a natural addition to the browser. We want to have a pop up that shows after a generative result that asks if you found bias. We want to make it feel like a real popup through maybe some other interactive program. We think that this will be a part of the whole final design we want to thoroughly test. We would like the user only be able to either use this new feature or decide not to report any bias by closing the popup and then they go to our old prototype. We think that this would provide more value to the user since it would allow them to report easier and be easy to pick up and use.

To test the success of the prototype we would like to see if they use this pop up to report if given a biased prompt. Or see that they would close the popup and just decide to continue using the application. We then would want to see if they did see bias, but decide not to report and ask them why they decided to not report.



Appendix

Ahana's Notes

Ahana x Sarah | Interview Consent Form

Sarah Ou <smou@andrew.cmu.edu>
To: Ahana Banerjee <ahanasb@andrew.cmu.edu>

Tue, Apr 23, 2024 at 11:31 PM

Consent Form for Participation in Research

Study Title: Lo-Fi Prototyping on Auditing Generative AI

Contact Information:

UCRE Team D2
5032 Forbes Avenue, Pittsburgh, PA, 15213
Email: ahanasb@andrew.cmu.edu risulu@andrew.cmu.edu garageles@andrew.cmu.edu crliu@andrew.cmu.edu alicewen@andrew.cmu.edu

Purpose of this Study

This study is part of a course in the Carnegie-Mellon University, Human-Computer Interaction program. Students need to learn about your experience as part of their course project. The purpose of the study is to gather feedback on the lo-fi prototype that has been designed by our team.

Summary

We will showcase our lo-fi prototype and have you interact with it to complete a certain task. We may then ask you follow-up questions depending on how you completed the task.

Procedures

First, you will be sat down with the interviewer who will introduce you to the study and give you a consent form. Then you will be asked some preliminary questions on your background and familiarity with reporting bias. After those questions, you will be shown a lo-fi prototype and be asked to complete a certain task while interacting with it.

We would like to record the session for our record. However, this recording will have no identifiers for you and will be securely stored in a Google Drive. We will analyze the video and then delete the video after analysis. The only people that will have access to this video are the involved researchers as noted at the beginning of this form.

We expect this study to take no longer than 30 to 40 minutes. The study will be conducted remotely or in a quiet room located on the CMU campus.

Participant Requirements

To be eligible for this study the participant must be at least 18 years old. They must also have some experience with using generative AI. They must also be able to meet with a researcher in person or remotely.

Risks

The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during recollection of visiting a website.

Compensation & Costs

There is no compensation for participation in this study.
There will be no cost to you if you participate in this study other than your time.

Future Use of Information

In the future, once we have removed all identifiable information from your data (information or bio-specimens), we may use the data for our future research studies, or we may distribute the data to other investigators for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality

By participating in the study, you understand and agree that UCRE Team D2 may be required to disclose your consent form, data, and other personally identifiable information as required by law, regulation, subpoena, or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your research data will be stored in a secure location on the UCRE Team D2 property. By participating, you understand and agree that the data and information gathered during this study may be used by UCRE Team D2 and published and/or disclosed by UCRE Team D2 to others outside of UCRE Team D2. However, your name, address, contact information, and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by UCRE Team D2.

The researchers will take the following steps to protect participants' identities during this study: (1) Each participant will be assigned a number; (2) The researchers will record any data collected during the study by number, not by name; (3) Any original recordings or data files will be stored in a secured location accessed only by authorized researchers.

Optional Permission

The researchers may want to use a short portion of any video or audio recording for illustrative reasons in presentations of this work for scientific or educational purposes. I give my permission to do so provided that my name **[and face]** will not appear.

Please initial here: S O YES NO

Rights

Your participation is voluntary. You are free to stop your participation at any point. Refusal to participate or withdrawal of your consent or discontinuing participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. The researcher may at his/her discretion remove you from the study for any of several reasons. In such an event, you will not suffer any penalty or loss of benefits or rights to which you might otherwise be entitled.

Right to Ask Questions & Contact Information

If you have any questions about this study, you should feel free to ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact the researcher by mail, phone, or e-mail following the contact information listed on the first page of this consent.

Voluntary Consent

By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You are encouraged to ask questions about any aspect of this research study during the study and in the future. By signing this form, you agree to participate in this research study. A copy of the consent form will be given to you.

Sarah Ou _____
PRINT PARTICIPANT'S NAME

Sarah Ou _____ 4/23/2024 _____
PARTICIPANT SIGNATURE DATE

I certify that I have explained the nature and purpose of this research study to the above individual and I have discussed the potential benefits and possible risks of participation in the study. Any questions the individual has about this study have been answered and any future questions will be answered as they arise.

Ahana Banerjee _____ 4/23/2024 _____
SIGNATURE OF PERSON OBTAINING CONSENT DATE

SARAH x AHANA | INTERVIEW

- Engaging with developers should feel impactful, like they genuinely care about your experience.
 - Helps customize the platform to user preferences
 - Can accommodate based on situation and emotion of the user
- Language like "create thread" should be straightforward and easy to understand
 - However there is concern that it does not mention bias
 - It can be confusing to some people that it does not mention the fact that it has to do with bias
 - People might see the other posts and begin to understand a bit better
- How can you open multiple chats with ChatGPT given the side bar?
 - We might have to offer this solution as an icon or extension or figure out another way to correctly place it for ease for the user
- Offering various ways to sort threads allows users to tailor their experience to what matters most to them.
 - This helps users save threads that they care about and actively participate in those conversations
- To prevent abuse, having moderators in place ensures that the quality of posts remains high and the community stays healthy.
 - Makes it more fun to be on and a healthy enviornment
 - No trolls
- A community forum provides a space to share your ideas, engage with others, and explore different perspectives, fostering meaningful interactions.

Edgar's Notes

Consent Form for Participation in Research

Study Title: Lo-Fi Prototyping

Contact Information:

UCRE Team D2
5032 Forbes Avenue, Pittsburgh, PA, 15213
Email: ahanab@andrew.cmu.edu, irisliu@andrew.cmu.edu, eangeles@andrew.cmu.edu,
cliu@andrew.cmu.edu, alicewen@andrew.cmu.edu

Purpose of this Study

This study is part of a course in the Carnegie-Mellon University, Human-Computer Interaction program. Students need to learn about your experience as part of their course project. The purpose of the study is to test out a potential solution for identifying bias in generative AI systems.

Summary

You will be shown a rough prototype of our proposed solution and asked to complete some questions. Then you will be asked some questions about your experience and then the study will conclude.

Procedures

First, you will be sat down with the interviewer who will introduce you to the study and give you a consent form. Then you will be shown the prototype and then asked to complete some simple tasks. After the tasks, you will be asked some questions regarding the prototype. Then, after looking at any closing thoughts the study will end.

We would like to record the session for our record. However, this recording will have no identifiers for you and will be securely stored in a Google Drive. We will analyze the video and then delete the video after analysis. The only people that will have access to this video are the involved researchers as noted at the beginning of this form.

We expect this study to take no longer than 30 minutes. The study will be conducted remotely or in a quiet room located on the CMU campus.

Participant Requirements

To be eligible for this study the participant must be at least 18 years old. They must also have some experience with using generative AI. They must also be able to meet with a researcher in person or remotely.

Risks

Carnegie Mellon University

The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during recollection of visiting a website.

Compensation & Costs

There is no compensation for participation in this study.
There will be no cost to you if you participate in this study other than your time.

Future Use of Information

In the future, once we have removed all identifiable information from your data (information or bio-specimens), we may use the data for our future research studies, or we may distribute the data to other investigators for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality

By participating in the study, you understand and agree that UCRE Team D2 may be required to disclose your consent form, data, and other personally identifiable information as required by law, regulation, subpoena, or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your research data will be stored in a secure location on the UCRE Team D2 property. By participating, you understand and agree that the data and information gathered during this study may be used by UCRE Team D2 and published and/or disclosed by UCRE Team D2 to others outside of UCRE Team D2. However, your name, address, contact information, and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by UCRE Team D2.

The researchers will take the following steps to protect participants' identities during this study: (1) Each participant will be assigned a number; (2) The researchers will record any data collected during the study by number, not by name; (3) Any original recordings or data files will be stored in a secured location accessed only by authorized researchers.

Optional Permission

The researchers may want to use a short portion of any video or audio recording for illustrative reasons in presentations of this work for scientific or educational purposes. I give my permission to do so provided that my name [and face] will not appear.

Please initial here: _____ YES _____ NO

Rights

Your participation is voluntary. You are free to stop your participation at any point. Refusal to participate or withdrawal of your consent or discontinued participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. The researcher may at his/her discretion remove you from the study for any of a number of reasons. In such an event, you will not suffer any penalty or loss of benefits or rights to which you might otherwise be entitled.

Right to Ask Questions & Contact Information

If you have any questions about this study, please ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact the researcher by mail, phone, or e-mail by the contact information listed on the first page of this consent.

Voluntary Consent

By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You are encouraged ask questions about any aspect of this research study during the course of the study and in the future. By signing this form, you agree to participate in this research study. A copy of the consent form will be given to you.

Samuel Franco

PRINT PARTICIPANT'S NAME

Samuel Franco
PARTICIPANT SIGNATURE

DATE 04/21/24

I certify that I have explained the nature and purpose of this research study to the above individual and I have discussed the potential benefits and possible risks of participation in the study. Any questions the individual has about this study have been answered and any future questions will be answered as they arise.

Edgar Araya
SIGNATURE OF PERSON OBTAINING CONSENT

DATE 04/21/24

Thinks that they can talk to the developer, switch to the community tab, or talk to AI

They can see categories and see the Create thread

Talk to AI and it would prompt them that bias is there and move them to the threads

They think that the threads only matter if they have someone to see and raise awareness

They think it could be interesting to talk to someone with expertise

3 on the feature since you can't force people to do things and there might be motivation if there are popups

4 on impact since some people love to write their opinion

Prompt saying that says did you find bias and if you click the button it automatically leads you to the community tab

Iris' Notes

Consent form

2 messages

Iris Liu <irisliu@andrew.cmu.edu>
To: badajung8@gmail.com

Tue, Apr 23, 2024 at 12:40 AM

Consent Form for Participation in Research

Study Title: Lo-Fi Prototyping

Contact Information:

UCRE Team D2
5032 Forbes Avenue, Pittsburgh, PA, 15213
Email: ahanab@andrew.cmu.edu, irisliu@andrew.cmu.edu, eangeles@andrew.cmu.edu,
crliu@andrew.cmu.edu, alicewen@andrew.cmu.edu

Purpose of this Study

This study is part of a course in the Carnegie-Mellon University, Human-Computer Interaction program. Students need to learn about your experience as part of their course project. The purpose of the study is to test out a potential solution for identifying bias in generative AI systems.

Summary

You will be shown a rough prototype of our proposed solution and asked to complete some questions. Then you will be asked some questions about your experience and then the study will conclude.

Procedures

First, you will be sat down with the interviewer who will introduce you to the study and give you a consent form. Then you will be shown the prototype and then asked to complete some simple tasks. After the tasks, you will be asked some questions regarding the prototype. Then, after looking at any closing thoughts the study will end.

We would like to record the session for our record. However, this recording will have no identifiers for you and will be securely stored in a Google Drive. We will analyze the video and then delete the video after analysis. The only people that will have access to this video are the involved researchers as noted at the beginning of this form.

We expect this study to take no longer than 30 minutes. The study will be conducted remotely or in a quiet room located on the CMU campus.

Participant Requirements

To be eligible for this study the participant must be at least 18 years old. They must also have some experience with using generative AI. They must also be able to meet with a researcher in person or remotely.

Risks

The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during recollection of visiting a website.

Compensation & Costs

There is no compensation for participation in this study.
There will be no cost to you if you participate in this study other than your time.

Future Use of Information

In the future, once we have removed all identifiable information from your data (information or bio-specimens), we may use the data for our future research studies, or we may distribute the data to other investigators for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality

By participating in the study, you understand and agree that UCRE Team D2 may be required to disclose your consent form, data, and other personally identifiable information as required by law, regulation, subpoena, or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your research data will be stored in a secure location on the UCRE Team D2 property. By participating, you understand and agree that the data and information gathered during this study may be used by UCRE Team D2 and published and/or disclosed by UCRE Team D2 to others outside of UCRE Team D2. However, your name, address, contact information, and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by UCRE Team D2.

The researchers will take the following steps to protect participants' identities during this study: (1) Each participant will be assigned a number; (2) The researchers will record any data collected during the study by number, not by name; (3) Any original recordings or data files will be stored in a secured location accessed only by authorized researchers.

Optional Permission

The researchers may want to use a short portion of any video or audio recording for illustrative reasons in presentations of this work for scientific or educational purposes. I give my permission to do so provided that my name [and face] will not appear.

Please initial here: _____ YES _____ NO

Rights

Your participation is voluntary. You are free to stop your participation at any point. Refusal to participate or withdrawal of your consent or discontinued participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. The researcher may at his/her discretion remove you from the study for any of a number of reasons. In such an event, you will not suffer any penalty or loss of benefits or rights to which you might otherwise be entitled.

Voluntary Consent

By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You are encouraged ask questions about any aspect of this research study during the course of the study and in the future. By signing this form, you agree to participate in this research study. A copy of the consent form will be given to you.

PRINT PARTICIPANT'S NAME

PARTICIPANT SIGNATURE

DATE

I certify that I have explained the nature and purpose of this research study to the above individual and I have discussed the potential benefits and possible risks of participation in the study. Any questions the individual has about this study have been answered and any future questions will be answered as they arise.

SIGNATURE OF PERSON OBTAINING CONSENT

DATE

--
Iris Liu
irisliu@andrew.cmu.edu
Cognitive Science, Class of 2026
Dietrich College, Carnegie Mellon University

Bada Jung <badajung8@gmail.com>
To: Iris Liu <irisliu@andrew.cmu.edu>

Tue, Apr 23, 2024 at 12:42 AM

I, Bada Jung, consent to participate in this study (04.22.2024)
[Quoted text hidden]

- Not effective at all
 - Community complaining won't do much
 - It will help them feel better/social cause
- I would not use
 - Private person/don't wanna put public profile
 - Would want burner account
- Bad idea
 - AI is unbiased but humans are biased
 - No longer see
 - Alters the effectiveness of AI
 - Lowers quantity
 - Human input will be skewed
 - Assuming its offensive then its good
- Positive: more brave as a group to report (im not just gaslighting)
 - They'll just live in their own bubble/echo chamber
 - More extreme in their opinions
- Effective - customized
 - They have a lot of data on you
 - Targeted ads
 - Privacy concerns
 - Leaked to others
 - Echo chamber
- Good example
 - Community
 - informative , you want to know if you're doing something
- Good example
 - Direct contact with representative
 - Have someone to handle biases (internal dedication to fixing bias)
 - Someone is actually listening
 - MOST EFFECTIVE OUT OF ALL SOLUTIONS
- It's okay as long as it consents
 - For example kids content
 - Should only use info from consenting adult
- Direct forum
 - Community helps bc people have similar issues and devs can see
- This is effective
 - But not compared to the other options
 - Less

Alice's Notes

Consent Form for Participation in Research

Study Title: Lo-Fi Prototyping

Contact Information:

UCRE Team D2

5032 Forbes Avenue, Pittsburgh, PA, 15213

Email: ahanab@andrew.cmu.edu, irisliu@andrew.cmu.edu, eangeles@andrew.cmu.edu,
crliu@andrew.cmu.edu, alicewen@andrew.cmu.edu

Purpose of this Study

This study is part of a course in the Carnegie-Mellon University, Human-Computer Interaction program. Students need to learn about your experience as part of their course project. The purpose of the study is to test out a potential solution for identifying bias in generative AI systems.

Summary

You will be shown a rough prototype of our proposed solution and asked to complete some questions. Then you will be asked some questions about your experience and then the study will conclude.]

Procedures

First, you will be sat down with the interviewer who will introduce you to the study and give you a consent form. Then you will be shown the prototype and then asked to complete some simple tasks. After the tasks, you will be asked some questions regarding the prototype. Then, after looking at any closing thoughts the study will end.

We would like to record the session for our record. However, this recording will have no identifiers for you and will be securely stored in a Google Drive. We will analyze the video and then delete the video after analysis. The only people that will have access to this video are the involved researchers as noted at the beginning of this form.

We expect this study to take no longer than 30 minutes. The study will be conducted remotely or in a quiet room located on the CMU campus.

Participant Requirements

To be eligible for this study the participant must be at least 18 years old. They must also have some experience with using generative AI. They must also be able to meet with a researcher in person or remotely.

Risks

The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during recollection of visiting a website.

Compensation & Costs

There is no compensation for participation in this study.

There will be no cost to you if you participate in this study other than your time.

Future Use of Information

In the future, once we have removed all identifiable information from your data (information or bio-specimens), we may use the data for our future research studies, or we may distribute the data to other investigators for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality

By participating in the study, you understand and agree that UCRE Team D2 may be required to disclose your consent form, data, and other personally identifiable information as required by law, regulation, subpoena, or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your research data will be stored in a secure location on the UCRE Team D2 property. By participating, you understand and agree that the data and information gathered during this study may be used by UCRE Team D2 and published and/or disclosed by UCRE Team D2 to others outside of UCRE Team D2. However, your name, address, contact information, and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by UCRE Team D2.

The researchers will take the following steps to protect participants' identities during this study: (1) Each participant will be assigned a number; (2) The researchers will record any data collected during the study by number, not by name; (3) Any original recordings or data files will be stored in a secured location accessed only by authorized researchers.

Optional Permission

The researchers may want to use a short portion of any video or audio recording for illustrative reasons in presentations of this work for scientific or educational purposes. I give my permission to do so provided that my name [and face] will not appear.

Please initial here: _____ RL _____ YES _____ NO

Rights

Carnegie Mellon University

Your participation is voluntary. You are free to stop your participation at any point. Refusal to participate or withdrawal of your consent or discontinued participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. The researcher may at his/her discretion remove you from the study for any of a number of reasons. In such an event, you will not suffer any penalty or loss of benefits or rights to which you might otherwise be entitled.

Right to Ask Questions & Contact Information

If you have any questions about this study, please ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact the researcher by mail, phone, or e-mail by the contact information listed on the first page of this consent.

Voluntary Consent

By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You are encouraged ask questions about any aspect of this research study during the course of the study and in the future. By signing this form, you agree to participate in this research study. A copy of the consent form will be given to you.

RACHEL LUO

PRINT PARTICIPANT'S NAME



PARTICIPANT SIGNATURE

DATE 4/23/24

I certify that I have explained the nature and purpose of this research study to the above individual and I have discussed the potential benefits and possible risks of participation in the study. Any questions the individual has about this study have been answered and any future questions will be answered as they arise.



SIGNATURE OF PERSON OBTAINING CONSENT

DATE 4/23/24

Introduction

Hello! Thank you for your time and help with this project. My name is [name] and I am working on a project for a user research course this semester. In the class, we are trying to understand how to incentivize users of generative AI to identify and report biases they see in an effort to improve algorithms. Our group's specific goal is to transform the process of reporting harmful everyday AI behaviors into a collaborative, seamless, and community-based act. Today, we would like you to test our prototype that we think will motivate and encourage users to report instances of bias they encounter while using generative AI systems, such as chatGPT.

Consent

Before we get started here is a consent form that outlines what we will be doing in this session as well as some notes on filming and recording this session. Keep in mind that you do not have to consent to be recorded to be part of this study. The data collected will only be used for this project for the User Research and Evaluation course.

Please review this consent form, look over it carefully, and sign it if you have no oppositions. I can answer any questions or issues you may have with it.

Instructions

Today, we'll have you use a prototype we've made to simulate posting a thread that reports bias on a community forum. Note that this is not a fully working interface and is solely for the purpose of simulating our design implementation. Please visit the link I will send: prototype figma link

First, imagine that you've encountered an instance of bias. Please go about reporting this instance.

1. First, make your way to the community forum
 - Tell me what you see
 - Tell me what you think users will be able to do on this part of the site

Able to post on forum, able to vote if you think other users' posts display harmful AI activity or not

2. Next, create a thread
 - Looking at this page, what features do you see that you think would help users more easily report interface bias

Category tab helps users sort which bias their post is about
Community forum is on interface itself rather than a separate application, makes it easy for users to make a post, more accessible

- What features do you think are missing that would help users more easily report bias or further motivate them to do so?

adding a picture to post might help people visualize what the poster saw

3. Imagine you've completed all the fields and "submit" your report.

- If you were to actually have submitted a report using this method, do you think you would feel like you've made an impact on the issue? Why or why not?

I don't think I would feel like I made an impact, but rather feel I'm part of a community where we can share things. If that community grew to be big, that could make an impact if the devs started looking at it

- Look at the question "Would you also like to speak with an AI developer representative?". Would you answer yes or no, and why?

I would probably answer no since I would've put all my thoughts in my post and don't want to go through the hassle of further discussing with someone

- Rate on a scale of 1-5 how highly you think this would motivate users to report bias on an interface (compared to not having this feature at all). 1 being would not motivate users at all, and 5 being would greatly motivate users. Explain your rating.

on motivating users, would rate this 3.5. Since it is on the interface itself, it is easy to access the report page and make a post. People are also driven by communication/social interaction online, so being able to interact with others and share something with the hopes that other users will see your post would also motivate people. Only reasons rating is not higher is because people are lazy, so don't know if this feature would directly motivate them to share something

- Rate on a scale of 1-5 how highly this would make users feel they are making an impact. 1 being no impact at all, and 5 being a huge impact.

on making users feel they're making an impact, rate 3. wouldn't feel making immediate impact but overall impact of a strong community may draw attention. if devs were to also respond to posts and interact with users, would make me feel i'm making more of a direct impact since my thoughts are heard by people who have influence on the interface

- What can this design do better (any additions or anything to remove?) to motivate/encourage users to report any bias they encounter?

If devs also interact with and reply to users, would feel more that I'm making an impact since I would know people who have the ability to make a change in the interface are seeing my thoughts.

Clarise's Notes

Consent Form for Participation in Research

Study Title: Speed Dating Session on Generative AI

Contact Information:

UCRE Team D2

5032 Forbes Avenue, Pittsburgh, PA, 15213

Email: ahanab@andrew.cmu.edu, irisliu@andrew.cmu.edu, eangeles@andrew.cmu.edu,
criiu@andrew.cmu.edu, alicewen@andrew.cmu.edu

Purpose of this Study

This study is part of a course in the Carnegie-Mellon University, Human-Computer Interaction program. Students need to learn about your experience as part of their course project. The purpose of the study is to propose some possible solutions for creating a sense of community in generative AI.

Summary

You will be shown a couple of storyboards and asked what your thoughts were on them as well as some other questions.

Procedures

First, you will be sat down with the interviewer who will introduce you to the study and give you a consent form. Then you will be shown some scenarios described through a storyboard. After the storyboards, you will be asked some questions regarding the storyboards. Then, after looking at all the storyboards the study will end.

We would like to record the session for our record. However, this recording will have no identifiers for you and will be securely stored in a Google Drive. We will analyze the video and then delete the video after analysis. The only people that will have access to this video are the involved researchers as noted at the beginning of this form.

We expect this study to take no longer than 30 to 40 minutes. The study will be conducted remotely or in a quiet room located on the CMU campus.

Participant Requirements

To be eligible for this study the participant must be at least 18 years old. They must also have some experience with using generative AI. They must also be able to meet with a researcher in person or remotely.

Risks

The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during recollection of visiting a website.

Compensation & Costs

There is no compensation for participation in this study.

There will be no cost to you if you participate in this study other than your time.

Future Use of Information

In the future, once we have removed all identifiable information from your data (information or bio-specimens), we may use the data for our future research studies, or we may distribute the data to other investigators for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Confidentiality

By participating in the study, you understand and agree that UCRE Team D2 may be required to disclose your consent form, data, and other personally identifiable information as required by law, regulation, subpoena, or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your research data will be stored in a secure location on the UCRE Team D2 property. By participating, you understand and agree that the data and information gathered during this study may be used by UCRE Team D2 and published and/or disclosed by UCRE Team D2 to others outside of UCRE Team D2. However, your name, address, contact information, and other direct personal identifiers will not be mentioned in any such publication or dissemination of the research data and/or results by UCRE Team D2.

The researchers will take the following steps to protect participants' identities during this study:

(1) Each participant will be assigned a number; (2) The researchers will record any data collected during the study by number, not by name; (3) Any original recordings or data files will be stored in a secured location accessed only by authorized researchers.

Optional Permission

The researchers may want to use a short portion of any video or audio recording for illustrative reasons in presentations of this work for scientific or educational purposes. I give my permission to do so provided that my name [and face] will not appear.

Please initial here: MY _____ YES _____ NO

Rights

Your participation is voluntary. You are free to stop your participation at any point. Refusal to participate or withdrawal of your consent or discontinued participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. The researcher may at his/her discretion remove you from the study for any of a number of reasons. In such an event, you will not suffer any penalty or loss of benefits or rights to which you might otherwise be entitled.

Right to Ask Questions & Contact Information

If you have any questions about this study, please ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact the researcher by mail, phone, or e-mail by the contact information listed on the first page of this consent.

Voluntary Consent

By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You are encouraged ask questions about any aspect of this research study during the course of the study and in the future. By signing this form, you agree to participate in this research study. A copy of the consent form will be given to you.

Maristella Yim

PRINT PARTICIPANT'S NAME

Maristella Yim

PARTICIPANT SIGNATURE

04/23/2024

DATE

I certify that I have explained the nature and purpose of this research study to the above individual and I have discussed the potential benefits and possible risks of participation in the study. Any questions the individual has about this study have been answered and any future questions will be answered as they arise.

SIGNATURE OF PERSON OBTAINING CONSENT

Clarise Liu

DATE

04/23/2024

Introduction

Hello! Thank you for your time and help with this project. My name is Clarise and I am working on a project for a user research course this semester. In the class, we are trying to understand how to incentivize users of generative AI to identify and report biases they see in an effort to improve algorithms. Our group's specific goal is to transform the process of reporting harmful everyday AI behaviors into a collaborative, seamless, and community-based act. Today, we would like you to test our prototype that we think will motivate and encourage users to report instances of bias they encounter while using generative AI systems, such as chatGPT.

Consent

Before we get started here is a consent form that outlines what we will be doing in this session as well as some notes on filming and recording this session. Keep in mind that you do not have to consent to be recorded to be part of this study. The data collected will only be used for this project for the User Research and Evaluation course. Please review this consent form, look over it carefully, and sign it if you have no oppositions. I can answer any questions or issues you may have with it.

Instructions

Today, we'll have you use a prototype we've made to simulate posting a thread that reports bias on a community forum. Note that this is not a fully working interface and is solely for the purpose of simulating our design implementation. Please visit the link I will send: prototype figma link

First, imagine that you've encountered an instance of bias. Please go about reporting this instance.

- Lost at first, starts clicking randomly for the first clickable thing
 - Only community tab works
 - Unable to chat with a technician, eyne though that was the first instinct
- The lo-fi prototype is so low fidelity that there isn't much opportunity to naturally engage with the application
- Navigated to the community forum, made a post, and gave up because nothing else worked

First, make your way to the community forum

Tell me what you see

- Quickly acknowledges all of the key features of the application
- Looks like a lot of posts with the option to mark as harmful, like, and comment
- Users have a profile

Tell me what you think users will be able to do on this part of the site

- Rant
- Complain about bias
- Treat it like a social media platform where you can like and dislike things
- Chat with a developer if needed

- Create their own thread
- Doesn't really care about the category filtering
 - Thinks the UI might not be the most effective, especially as more categories are formed
 - Would prefer to have a search feature

Next, create a thread

Looking at this page, what features do you see that you think would help users more easily report interface bias

- Creating a title
- Selecting a category
- Free-form text to elaborate on perspective

What features do you think are missing that would help users more easily report bias or further motivate them to do so?

- Unsure how to guarantee all categories are correctly organized
- Not sure what to do if the title is not descriptive/what counts as a good title
- There should be an option to insert a screenshot
- Looks pretty straightforward; not sure what else can be added

Imagine you've completed all the fields and "submit" your report.

If you were to actually have submitted a report using this method, do you think you would feel like you've made an impact on the issue? Why or why not?

- No
- Not sure who cares or pays attention to this issue
 - Probably one of the thousands of posts
- Based on the harmful vs unharmed, seems like posts only get noticed if many people mark it as harmful
- Would be nice to have fun animations after submitting something as a general incentive/sense of completion after writing a repost

Look at the question "Would you also like to speak with an AI developer representative?". Would you answer yes or no, and why?

- No
- Generally antisocial and not willing to talk to people
- Would be more willing to say yes if it was a text, since texting is less social labor than in-person conversation

Rate on a scale of 1-5 how highly you think this would motivate users to report bias on an interface (compared to not having this feature at all). 1 being would not motivate users at all, and 5 being would greatly motivate users. Explain your rating.

- 1; probably not effective
 - You can't report unless you go into the community tab
 - Why would I ever want to go into the community tab while using ChatGPT? Would be better to use "bias" or "report" in the tab name instead.

Rate on a scale of 1-5 how highly this would make users feel they are making an impact. 1 being no impact at all, and 5 being a huge impact.

- Depends on the user
- 1.5 - feels like user would make an impact, but only if many people validate the post by marking it as "harmful"
 - This feels difficult to do, especially on popular platforms with lots of users—and thus a lot of posts

What can this design do better (any additions or anything to remove?) to motivate/encourage users to report any bias they encounter?

- Instead of chat with developer, might be good to change it to "report bias" or something. That button should go on the community tab, rather than having the extra action of having to click on the community tab
 - Forces people to interact with others
- Add a search function within the posts and filter things based on recentness and harmfulness
- Doesn't really understand the star
 - Add a place where you can view all stars to give it value
- Should be able to see your history so that the user can see their past interactions on the website

Group Synthesis

I would probably answer no since I would've put all my thoughts in my post and don't want to go through the hassle of further discussing with someone

Alice Wen

the low fidelity prototype is so lo-fi that the participant could easily figure out what we wanted them to do, and thus struggled to organically engage with it

Clarise Liu

- Would not speak to dev
- Feel like they wouldn't respond, too many people who want to talk

Itz

chat with developer addresses the impact portion -- makes you feel like they care, makes the experience customizable

Ahana Banerjee

would prefer to reword to "text with developer", since texting is less social energy than talking

Clarise Liu

Thinks that they can talk to the developer, switch to the community tab, or talk to AI

Edgar Angeles

They think it could be interesting to talk to someone with expertise

Edgar Angeles

"chat with a developer" implies in-person conversation, either by phone or video call

Clarise Liu

simple

Ahana Banerjee

the structure of making a post is easy to understand

Clarise Liu

They can see categories and see the Create thread

Edgar Angeles

"create thread" language makes sense

Ahana Banerjee

unsure of how to guarantee that all titles are well-formatted

Clarise Liu

having the "chat with developer" button go right to the community tab ensures everyone engages with the tab, as opposed to having to explicitly tap into it

Clarise Liu

how can you open multiple chats with chatGPT, then the side menu would need to be moved.

Ahana Banerjee

helpful to sort by type of bias to find patterns

Itz

wants to a search function within the posts and filter things based on recency and harmfulness

Clarise Liu

wants to add a search feature

Clarise Liu

could have different ways to sort the threads and users can choose what they want to see

Ahana Banerjee

quality of posts could be abused, so could have moderators to clean it up

Ahana Banerjee

unsure of how to guarantee that all categories are correctly organized

Clarise Liu

It could be not taken seriously
Trolls could post whatever they want

Itz

Prompt saying that says did you find bias and if you click the button it automatically leads you to the community tab

Edgar Angeles

Talk to AI and it would prompt them that bias is there and move them to the threads

Edgar Angeles

wants to be able to add a screenshot

Clarise Liu

Would be nice to have fun animations after submitting something as a general incentive/sense of completion after writing a post

Clarise Liu

adding a picture to post might help people visualize what the poster saw

Alice Wen

The buttons (harmful/unharmful) - not like currency but rather to improve the AI = feels more impactful than like/dislike

Itz

wants to be able to see user history so that the user can see their past interactions on the website

Clarise Liu

wants to add a function where users can view all starred posts

Clarise Liu

people will use it to express themselves/validate their feelings

Iris

Motivates others because there's an option to share your emotions
Similar to social media

Iris

4 on impact since some people love to write their opinion

Edgar Angeles

the labeling of the community forum feels like it could become a social media platform

Clarise Liu

I don't think I would feel like I made an impact, but rather feel I'm part of a community where we can share things. If that community grew to be big, that could make an impact if the devs started looking at it

Alice Wen

Comments might end up buried
Some way to track the diff biases? Find the most issues/patterns

Iris

the categories could become too saturated and difficult to navigate, especially as more categories are added

Clarise Liu

too many users >> too many posts >> post can easily be lost if they aren't heavily interacted with

Clarise Liu

community forum is good to be able to share your ideas and think about other people's ideas and have a way to interact with other ppl

Ahana Banerjee

on making users feel they're making an impact, rate 3. wouldn't feel making immediate impact but overall impact of a strong community may draw attention. If devs were to also respond to posts and interact with users, would make me feel I'm making more of a direct impact since my thoughts are heard by people who have influence on the interface

Alice Wen

Prompt saying that says did you find bias and if you click the button it automatically leads you to the community tab

Edgar Angeles

Talk to AI and it would prompt them that bias is there and move them to the threads

Edgar Angeles

wants to be able to add a screenshot

Clarise Liu

Would be nice to have fun animations after submitting something as a general incentive/sense of completion after writing a post

Clarise Liu

adding a picture to post might help people visualize what the poster saw

Alice Wen

The buttons (harmful/unharmful) - not like currency but rather to improve the AI - feels more impactful than like/*dislike*

Iris

Why would i ever want to go into the community tab while using ChatGPT?

Clarise Liu

Would be better to use "bias" or "report" in the community tab name to better show its purpose

Clarise Liu

They think that the threads only matter if they have someone to see and raise awareness

Edgar Angeles

If you put it out first it's easier for others to chime in

Iris

on motivating users, would rate this 3.5. Since it is on the interface itself, it is easy to access the report page and make a post. People are also driven by communication/social interaction online, so being able to interact with others and share something with the hopes that other users will see your post would also motivate people. Only reasons rating is not higher is because people are lazy, so don't know if this feature would directly motivate them to share something

Alice Wen

3 on the feature since you can't force people to do things and there might be motivation if there are popups

Edgar Angeles