A CONSENT AND DATA COLLECTION PROCESSES

To obtain informed consent and anonymize students' data, the requisite processes started by the course instructor emailing the following:

> Dear all,
> As I have mentioned in the course syllabus and in prior lectures, my team and I plan to conduct a research study assessing the effectiveness of several educational tools and resources I will utilize in this course during the semester. We are writing to invite you to participate in this study.
> You will not be asked to do anything beyond the normal activities and assignments that are part of the course curriculum. You have the choice of opting out of data collection for research purposes. Participants will not receive any compensation, and those opting out will not not have their course outcome adversely impacted in any way.
> Please see the opt-out form attached. If you wish to opt out of the study, please sign the form indicating that and send it to [student advised by instructor but not part of teaching staff] preferably before the class on [date of first research activity]. **Please do NOT cc the course staff. This way the teaching staff (including me) won't know whether you opted out until after all grades have been posted and submitted. This is to ensure that you won't feel coerced to participate in our study.**
> If you have any questions or concerns at any point, please don't hesitate to get in touch with me.
> Best regards,
> [Instructor]

The opt-out consent form contained the following information regarding data collection and usage:

> Data collection: (1) In the classroom, you (the study participant) will complete a brief pre-study questionnaire aiming to assess your initial understanding of the concepts. (2) You will then learn about the concepts and the corresponding tool through the instructor. (3) You will be asked to work in teams with the tool and submit a short report in collaboration with your teammates. If all students in your team consent, we will ask you to record the audio of your deliberation via [university's] Zoom rooms and share the recording or its transcript with us. (4) You will complete a survey and participate in a class-wide evaluation of the activity. (5) Finally, one week after the class activity, you will complete a post-activity questionnaire/quiz designed to assess your final understanding of the concepts.
> Most questions in our questionnaires will require open-ended responses. Please do not reveal any private or personally-identifiable information about yourselves or others in your answers to the open-ended questions. We will also ask you to provide us with basic information about your educational background and demographics. The purpose of these questions is for us to detect any significant variations in responses across the corresponding dimensions. Answering demographic questions is entirely optional.

In accordance with this outlined protocol, a Python script was run to anonymize all data by converting all email addresses and other identifiable information to random IDs in the form of integers starting from 1. No members of the research team that were also teaching staff observed identifiable information prior to running this script.

B PRE-CLASS QUESTIONNAIRE (VERBATIM)

This questionnaire aims to understand your background, knowledge, and goals as they relate to Machine Learning (ML), Artificial Intelligence (AI), and the topics covered in [...] (our class). (Please respond to the best of your knowledge and memory and don't consult any external resources. There is no right or wrong answer to any of these questions). It also asks an optional set of questions about your demographics. Please feel free to leave that last part un-answered.

Your answers to this questionnaire will:

(1) inform the themes, topics, and tools we will focus on during the semester.
(2) If you consent, a de-identified version of your collective responses will be utilized in a research study by [...] (the instructor), in which she aims to assess the effectiveness of the educational tools she utilizes in this course. The opt-out consent form for the study can be found here. Please follow the instructions there if you wish to opt out.

We expect the questionnaire will take 10-15 minutes to complete (on average).

B.1 Educational Background

(1) Which one best describes your educational background?
   A. Science, Technology, Engineering, Mathematics (STEM)
   B. Humanities, Social Sciences and the Arts (HSA)
   C. Other…(Short Answer)
(2) What is your current major (i.e., the name of the degree program you are enrolled in)? (Short Answer)
(3) What is the highest degree or level of school you have completed?
   A. High school or equivalent
   B. Bachelor degree or equivalent
   C. Master degree or equivalent
   D. Doctoral degree or equivalent

Table 1: Students were offered a brief description of values taken verbatim from prior work of Jakesch et al. [32]

RAI value	Description
Transparency	A transparent AI system produces decisions that people can understand. Developers of transparent AI systems ensure, as far as possible, that users can get insight into why and how a system made a decision or inference.
Fairness	A fair AI system treats all people equally. Developers of fair AI systems ensure, as far as possible, that the system does not reinforce biases or stereotypes. A fair system works equally well for everyone independent of their race, gender, sexual orientation, and ability
Safety	A safe AI system performs reliably and safely. Developers of safe AI systems implement strong safety measures. They anticipate and mitigate, as far as possible, physical, emotional, and psychological harms that the system might cause.
Accountability	An accountable AI system has clear attributions of responsibilities and liability. Developers and operators of accountable AI systems are, as far as possible, held responsible for their impacts. An accountable system also implements mechanisms for appeal and recourse.
Privacy	An AI system that respects people's privacy implements strong privacy safeguards. Developers of privacy-preserving AI systems minimize, as far as possible, the collection of sensitive data and ensure that the AI system provides notice and asks for consent.
Autonomy	An AI system that respects people's autonomy avoids reducing their agency. Developers of autonomy-preserving AI systems ensure, as far as possible, that the system provides choices to people and preserves or increases their control over their lives.
Performance	A high-performing AI system consistently produces good predictions, inferences or answers. Developers of high-performing AI systems ensure, as far as possible, that the system's results are useful, accurate and produced with minimal delay.

   E. Other…(Short Answer)
(4) Have you completed any courses with a significant ML component in the past? If yes, please list those courses. (Short Answer)
(5) Which one best describes your level of familiarity with AI/ML?
   A. None
   B. Elementary
   C. Intermediate
   D. Advanced
   E. Other…(Short Answer)
(6) Have you received any form of training on ethical and societal considerations around the use of AI or ML?
   A. Yes
   B. No
(7) If your response to the previous question was "yes", please briefly describe the nature of the training. (Short Answer)
(8) Please briefly describe what motivated you to take this course. (Long Answer)
(9) What do you hope to learn from/gain out of this course? (Long Answer)
(10) Are there any specific *topics, tools, or applications* you would like to learn about as a part of this course?

B.2 ML in Society

(1) Please name a use case/application of ML in society that you are excited about. (Short Answer)
(2) How did you learn about this use case? (Short Answer)
(3) Please name a use case/application of ML in society that you are concerned about. (Short Answer)
(4) How did you learn about this use case? (Short Answer)
(5) Which one of the following broad categories of values do you believe is most urgent to address to promote the responsible use of ML in socially high-stakes domains? (A high-level description of each value–taken from prior work–is provided below. Options are ordered randomly).
   A. Fairness
   B. Safety
   C. Transparency
   D. Privacy
   E. Accountability & governance
   F. Human autonomy & agency
   G. Performance & efficiency
   H. Other…(Short Answer)
(6) How do you believe *Machine Learning experts and engineers* can contribute to promoting the above value? Please elaborate. (Long Answer)

B.3 AI News

(1) How often do you normally hear about AI-related news (e.g., new use cases, significant advances, etc.)?
   A. Every day
   B. Every week
   C. Every month
   D. Every year

E. Rarely

(2) Where (or from what source) do you usually hear about AI-related news? (Options below are ordered randomly.)
   A. Educational resources (e.g., lectures, course material, university mailing lists, ...)
   B. Social media (e.g., Twitter, Facebook, LinkedIn ...)
   C. Traditional media (e.g., TV, Radio, printed or online newspapers, ...)
   D. Word of mouth (e.g., through classmates, friends, and acquaintances)
   E. Other...(Short Answer)
(3) Please briefly describe an example of recent AI news you have heard about. (Long Answer)

B.4 Demographic Information (Optional)

The purpose of this section is to understand whether there are significant variations in your responses to the previous questions along demographic dimensions. Answering this part is entirely optional, so please feel free to leave a question unanswered if you prefer not to disclose the corresponding information about yourself. All categorical alternatives in this section have been ordered alphabetically.

(1) Which one better describes your political views – please pick the closest one if neither is an exact fit?
   A. Conservative
   B. Liberal
   C. Libertarian
   D. Other...(Short Answer)
(2) Which one best describes your gender?
   A. Female
   B. Male
   C. Non-binary
   D. Other...(Short Answer)
(3) What is your age group?
   A. 18-26
   B. 27-40
   C. 41 or older
   D. Other...(Short Answer)
(4) Which one best describes your race?
   A. Asian
   B. Black or African American
   C. Native Hawaiian or Other Pacific Islander
   D. White
   E. American Indian or Alaskan Native
   F. Other...(Short Answer)
(5) Do you believe you belong to a marginalized/disadvantaged group or community?
   A. Yes
   B. No
   C. Other...(Short Answer)

B.5 Conclusion

Thank you so much for your participation!

(1) If you have any feedback for the teaching instructor or the research team about the questionnaire, please leave your comments here. (Long Answer)

C IN-CLASS ACTIVITY

C.1 Individual activity

(1) Go to "Table View" tab, the search page corresponding to your number.
(2) Read the summary of the 10 incidents in your page.
(3) Among the 10, pick the incident that grabs your attention (e.g., because it's news to you, it's surprising, the magnitude of impact could be significant, ...)
(4) Read the incident report carefully.
(5) Then provide 1–2 word responses to the next questions.
   - What was the source of the story (e.g., which website/author published it)?
   - In what application domain did the incident occur?

- What was the nature of the incident (i.e., the concern that was raised)?
- Who was (partially) responsible for the incident (e.g., because they developed, deployed, or used the AI system)?
- Who was (potentially) harmed?
- How was the incident ultimately addressed (Was there a penalty? Was the tool discontinued?)

C.2 Team activity

(1) Form a team with the 4 classmates sitting closest to you.
(2) Search for a recent AI/ML incident that has not been submitted to the database.
   - If you don't know where to start, go back to the AIID and look at alternative views of the data (e.g., spatial, entities, ...) or the review queue for inspiration.
   - It is okay if you can't find such a story after looking for 15 minutes.
(3) Produce a team report as follows and upload it to the course website individually.
   - Submit a report on your story to AIID. Take a snapshot of your report.
   - If you didn't find a story, briefly describe your search process (e.g., queries, websites).

D POST-ACTIVITY QUESTIONNAIRE (VERBATIM)

This questionnaire aims to assess the efficacy of the first class activity in which we explored the AI Incident Database (AIID) by eliciting your feedback about the tool. As before, there is no right or wrong answer to any of these questions.

If you consent, a de-identified version of your collective responses will be utilized in a research study by [...] (the instructor), in which she aims to assess the effectiveness of the educational tools she utilizes in this course. The opt-out consent form for the study can be found here. Please follow the instructions there if you wish to opt out.

We expect the questionnaire will take around 7 minutes to complete (on average).

D.1 Assessing the impact of AIID

In the previous questionnaire, we asked you about

- your motivation for taking this course and the topics you'd like to learn about,
- the applications of ML in society that you are excited/concerned about,
- the values you believe should be prioritized,
- and the role of ML experts in promoting those values.

The goal of the following questions is to understand whether working with AIID has impacted your response to any of the above questions.

(1) Having explored AIID, has your motivation for taking this course changed at all?
   A. It hasn't changed.
   B. It has increased.
   C. It has decreased.
   D. Other...(Short Answer)
(2) Please briefly describe why you selected the answer above. (Long Answer)
(3) Having explored the AIID, are there any new/additional uses/applications of ML in society that you are now excited about? (Long Answer)
(4) Having explored the AIID, are there any new/additional uses/applications of ML in society that you are now concerned about? (Long Answer)
(5) Having explored the AIID, which one of the following broad categories of values do you now believe is most urgent to address to promote the responsible use of ML in socially high-stakes domains? (A brief description of each value–taken from prior work–is provided below. Options are ordered randomly).
   A. Fairness
   B. Safety
   C. Transparency
   D. Privacy
   E. Accountability & governance
   F. Human autonomy & agency
   G. Performance & efficiency
   H. Other...(Short Answer)
(6) Has exploring the AIID changed your belief about how *Machine Learning experts and engineers* can contribute to promoting the above value? Please elaborate. (Long Answer)
(7) Having explored the AIID, are there any additional tools or topics you would like to learn about in this course? (Long Answer)

D.2 Feedback on AIID

(1) How likely are you to use AIID in the future?
   A. I will refer to it frequently (i.e., at least every 1-2 weeks).
   B. I will refer to it occasionally (i.e., every few months).
   C. I will refer to it only if the need arises.
   D. I will likely never use it again.
   E. Other…(Short Answer)
(2) Do you have any additional thoughts or comments on the limitations of AIID? (Long Answer)
(3) How easy/difficult did you find it to work with the user interface of AIID?
   (a) Very Easy
   (b) Easy
   (c) Neutral
   (d) Hard
   (e) Very Hard
(4) Do you have any suggestions or ideas regarding how AIID can be improved? (Long Answer)

D.3 Conclusion

Thank you so much for your completing this assignment!

(1) If you have any feedback for the teaching instructor or the research team about this questionnaire in particular or the first session in general, please leave your comments here. (Long Answer)

E IN-CLASS FINDINGS

Q1: What was the source of the story?

1. The Guardian
2. Harvard paper
3. standard.co.uk
4. globalcitizen.org
5. Seattle Times
6. Detroit Free Press
7. BBC
8. USA Today
9. NPR
11. WashingtonPost
12. Bulletin of the Atomic Scientists
13. WashingtonPost
14. Forbes.com
16. Keen Security Lab
17.thedrive.com
18. wired.com
19. politico.eu
20. The Atlantic
22. Vice.com
23. Statnews
24. prnewswire
29. Nabla.com
30. Avaaz
32. snopes.com
33. BBC
34. Slate
35. arstechnica.com
36. MIT Tech Review
37. CNBC
39. Institute for Strategic Dialogue
40. splinter news

Q2: In which domain did the incident occur?

1. Information
2. Advertising
3. mobile devices
4. Ecommerce
5. Information
6. Self-driving
7. Social Media
8. Entertainment
9. Fact checking
11. Public safety
12. War/defense
13. Healthcare
14. Advertising
16. Self-driving
17. Transportation
18. AVs
19. Public Safety
20. Politics/War
21. Transport
22. Security
23. Healthcare
24. Public Safety
29. Healthcare
30. Social media
32. Social media
34. Education
35. Politics/Election
36. Computer Vision
37. Security
39. Social media
40. Healthcare

Q3: What was the nature of the incident?

1. Conflict between bots
2. Racial discrimination
3. Algo hacking
4. sexual bias
5. Gender bias
6. Harm to physical health/safety
7. Political comments
8. Racial bias
9. Inaccuracy
11. facial recognition abuse
12. autonomous killer robot
13. Discrimination
14. Inaccuracy
16. Adversarial attack
17. Inaccuracy
18. imperfect human
19. Misclassification
20. Deepfake - Misinformation
21. Car accident
22. Biased Facial Recognition
23. Poor performance
24. Waze blocking fire escape routes
29. Inaccuracy
30. Misinformation
32. Fake Users
33. Discriminatory algorithm
34. Cheating
35. misinformation
36.Bias/Stereotypes
37. Incorrect unlocks
39. Inaccuracy/Bias
40. Privacy

Q4: Who was (partially) responsible for the incident

1. Wikipedia
2. Google
3. Apple
4. Amazon
5. LinkedIn
6. Tesla
7. Microsoft
8. Niantic Labs
9. Facebook
11. Pimeyes
12. GNA-AF (Libyan military)
13. Optum
14. Brand safety tech
16. Tesla
17. Bath government
18. auto company
19. Spanish Ministry of Interior
20. Hackers, Ukraine 24 TV channel
21. Tesla
22. SN Technologies
23. IBM
24. Waze, Google
29. OpenAI
30. Youtube
32. The BL
33. UK Home Office
34. Students, OpenAI
35. Youtube
36. OpenAI
37. Google
39. Facebook
40. Meta

Q5: Who was or could have potentially been harmed?

1. Wikipedia users
2. Racial minority groups
3. iPhone users
4. Female applicants
5. Women
6. Driver
7. CCP
8. Pokemon Go players in minority neighborhoods
9. Facebook users interested in US election and covid news
11. vulnerable individuals with an online footprint
12. people living in Libya during war
13. seriously ill patients from minority groups
14. Marketing agencies
16. drivers
17. drivers
18. drivers
19. Women at risk of gender violence
20. Ukrainians, Soldiers
21. Pedestrians
22. Students
23. Cancer patients
24. Town residents
29. Patients
30. Youtube users
32. Facebook users
33. UK Visa applicants
34. Students + Teachers
35. users & citizens
36.Minority Groups
37. Phone owners
39. Facebook users
40. Patients

Q6: How was the incident ultimately addressed?

1. Unclear
2. Unclear
3. Unclear
4. Fired the AI algorithm
5. Removed search for men specifically, tweaked search engine
6. unclear
7. Reprogramming
8. Attributed to legacy app
9. reduced post distribution
11. unclear
12. UN comissioned a report
13. none
14. no
16. Tesla fixed it
17. Ticket thrown out
18. no
19. Call for auditing
20. Video removed, Zelenskyy clarified
21. Tesla taxi discontinued
22. Not
23. TBD/ algo update
24. Congress has set up a special committee
29. Warning issued
30. Misinformation reports
32. Fake users were deleted
33. UK home office agreed to stop using algorithm
34. No
35. No
36.Unclear
37. No
39. TBD
40. Unclear

Figure 5: Results of students' explorations of AIID.

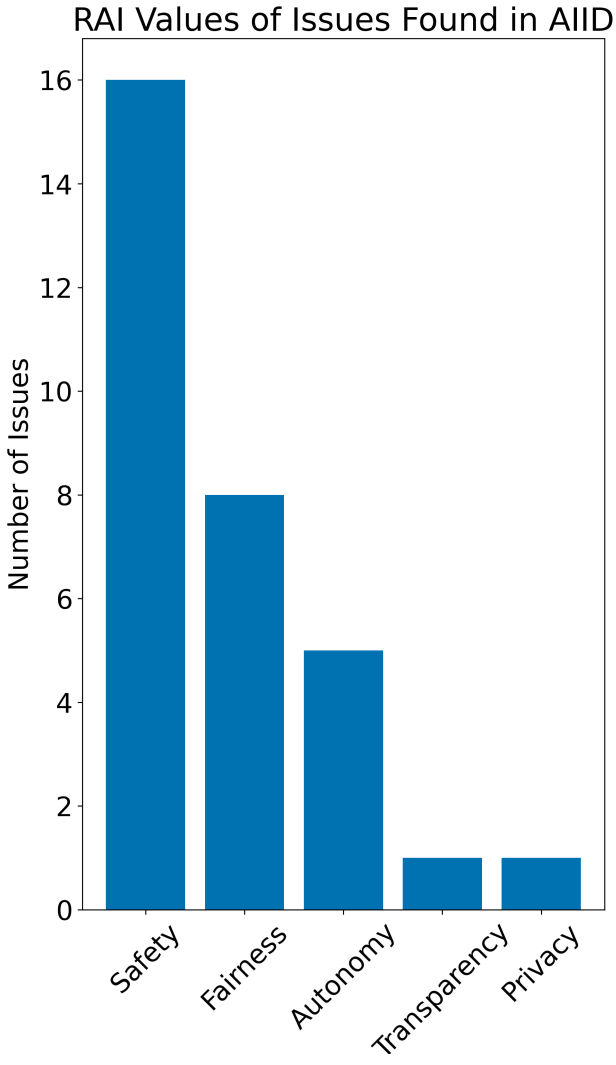

Figure 6: Value distribution of issues found interesting by students. Of the 31 issues explored, we identified 16 as pertaining to the RAI value of Safety as per Jakesch et al. [32], 8 as pertaining to Fairness, 5 as pertaining to Autonomy, 1 as pertaining to Privacy, and 1 as pertaining to Transparency. We additionally note that Accountability & Governance cannot be a reason for an incident of harm as calls for oversight stem from the results of harm.

F PARTICIPANT DEMOGRAPHICS

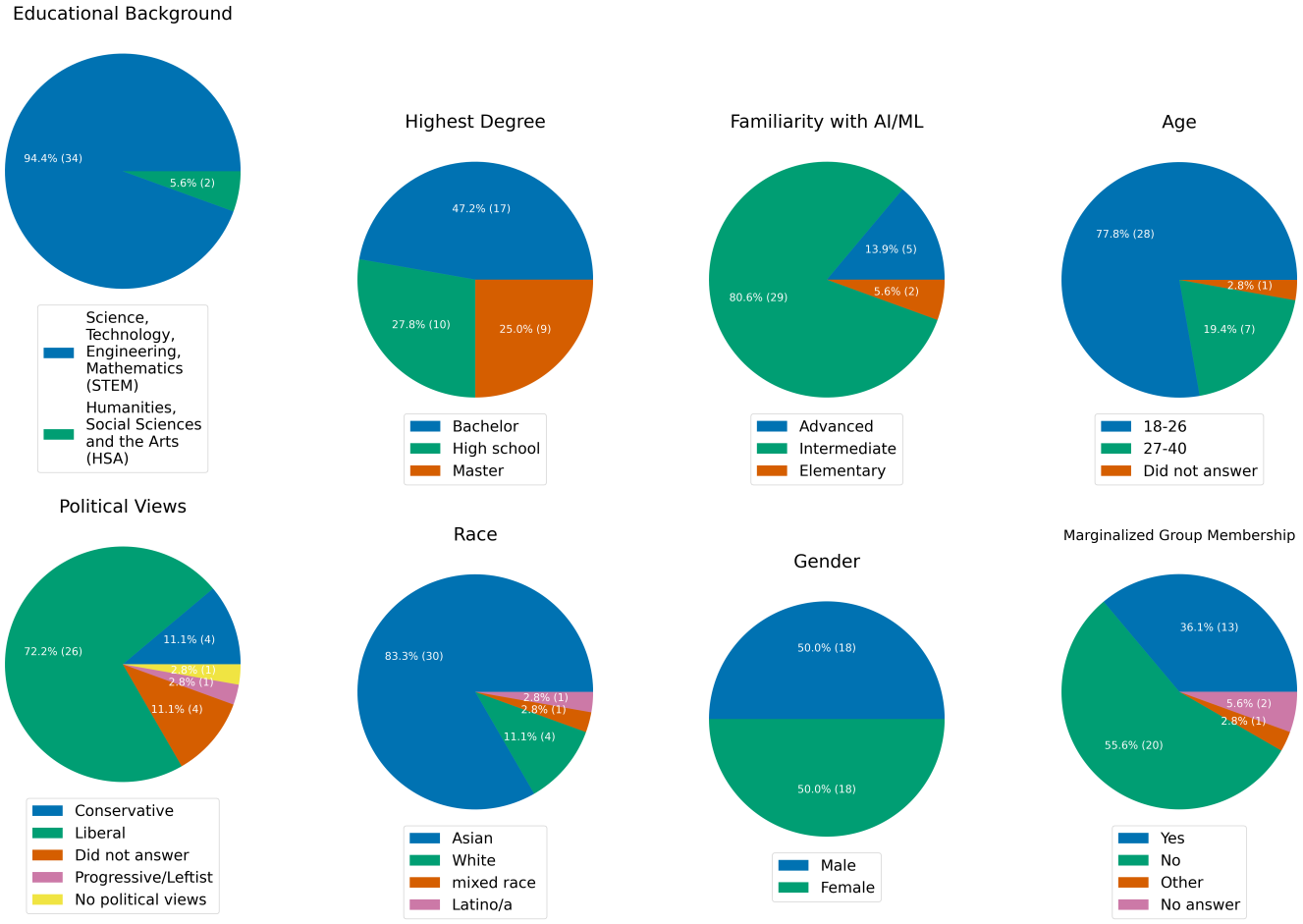

Figure 7: Participants' demographics and backgrounds. The sample appear relatively diverse (e.g. across gender, political leanings, and marginalized group membership) while it is highly homogeneous in other respects (e.g. most students have a STEM background, and belong to the same age and racial groups).

G STATISTICAL TESTS

Using the pre-activity and post-activity questionnaire responses pertaining to different RAI values as per [32], we generated a two-way contingency table as shown in Table 2. Note that this is a matrix form of the changes represented in Figure 1. Given this categorical data, we then performed two statistical tests of marginal homogeneity: a Bhapkar test [9] and a Stuart-Maxwell test [47, 69]. More specifically, we conducted these tests to investigate whether we could reject the null hypothesis: our classroom activity did not induce a change in value prioritization and we observed these differences by chance. Even though as per Keefe [35], the Bhapkar test is generally preferred as it is more powerful, we report results with both tests here.

Based on implementations in the DescTools package in R, the Bhapkar test yielded a p-value of 0.2217 and the Stuart-Maxwell test yielded a p-value of 0.34. At significance level $\alpha = 0.05$, neither of these results are statistically significant, which in turn means we fail to reject the null. However, we refer to existing literature such as Vuolo et al. [72] that highlights how tests like Stuart-Maxwell require hundreds of paired samples in order to test for conventional levels of significance. Given that our results with only 30 samples still yielded p-values somewhat close 0.05, we argue that this evidences the need for further research with more data for generalization and significance purposes.

	0	1	2	3	4	5
0	2	1	0	0	0	0
1	2	4	0	0	3	0
2	0	0	0	0	0	0
3	0	1	0	2	0	0
4	0	0	0	0	8	1
5	2	0	1	0	1	2

Table 2: Contingency table based on pre-activity and post-activity data from our classroom activity. Rows correspond to pre-activity responses and columns correspond to post-activity responses. Each row or column refers to a different RAI value as outlined in Jakesch et al. [32]. Category 0 refers to 'Accountability & governance', category 1 refers to 'Fairness', category 2 refers to 'Performance & efficiency', category 3 refers to 'Privacy', category 4 refers to 'Safety', and category 5 refers to 'Transparency'. This information is also reflected in Figure 1.

H REPORTED AIID USABILITY RATINGS

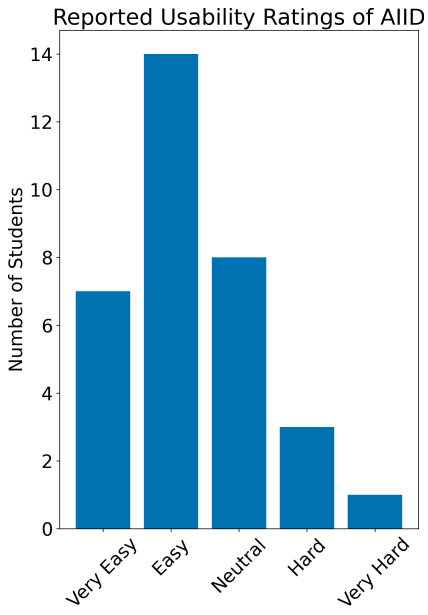

Figure 8: Student feedback on the usability of the AI Incident Database. While the majority of students reported that the database was "Easy" or "Very Easy" to use, a nontrivial number of students were "Neutral" or stated that it was "Hard" or "Very Hard" to use. This suggests that while it is already useful, enhancements to the tool would increase its utility.

I AIID REVIEW QUEUE

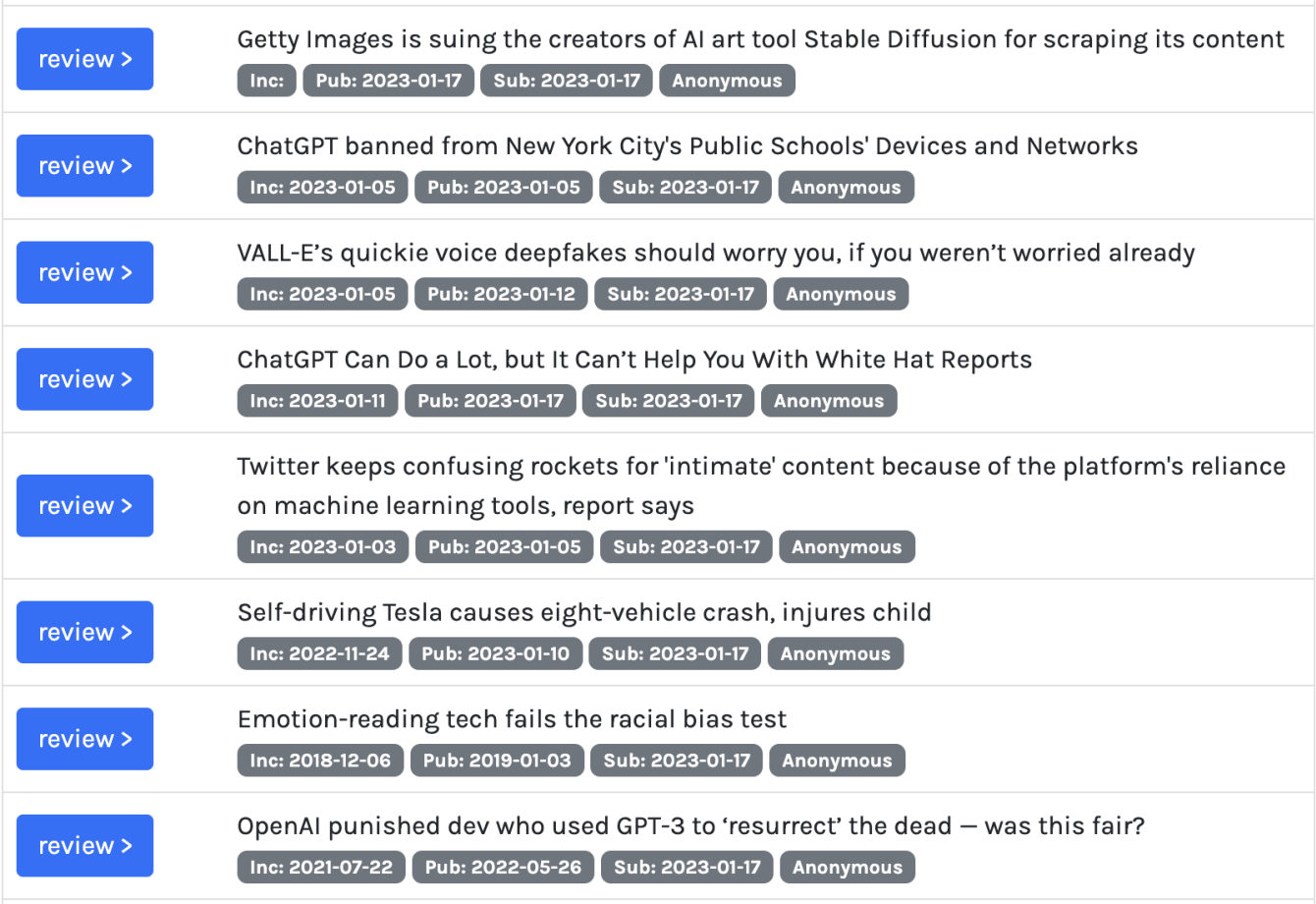

Figure 9: Snapshot of review queue containing reports submitted as part of our activity.