AI Audit Project

Teaching Critical AI Literacies Starter Kit

Dr. Daniel Estrada, NJIT
estrada@njit.edu

Project Overview

The AI Audit Project is a five week group activity that provides students the opportunity to critically engage with popular AI services for text and image generation through an extended probing exercise. Students are asked to perform an external audit of several generative AI services (chatGPT, CoPilot, DALL-E, Gemini, etc), modeled on audit frameworks discussed in the scholarly literature. Through this activity, students will learn to operate different AI services, to engineer effective prompts, and to compare the quality of their results. Students will also experience first hand the limitations and failure modes of these services, in order to reflect on the ethical and social impact of these technologies. Finally, students will learn about product safety protocols commonly used in AI and other industrial contexts, such as FMEA charts and social impact statements.

This project was developed as the final group project for a senior humanities seminar with 18 students. Students are expected to read, write, and conduct research at a college level, but no other technical background was required for the course. The length and scope of the project can be adjusted to fit other learning contexts and groups.

Primer on AI Audits

An AI Audit or Algorithmic Audit is a process through which an AI service is reviewed and evaluated to identify potential risks and failure modes in its operation. According to Raji et al. (2020), a thorough AI audit can serve as “accountability measures for deployed systems”, and can operate as “a mechanism to check that the engineering processes involved in AI system creation and deployment meet declared ethical expectations and standards, such as organizational AI principles.” AI Audits can be conducted internally, by the AI developer’s organization itself, or externally by third party reviewers. AI Audits became popular in the AI industry following research by Boulamwini and Gebru (2018), which conducted an external audit of popular facial recognition software. This work demonstrates clear and systematic failures in the software when attempting to recognize the faces of women with darker skin tones, reflecting inadequate training data. See Appendix A for background readings in the AI Audit literature.

Raji et al (2020) lays out a complete framework for conducting internal algorithmic audits. The framework develops through several stages in which the scope of the audit is made explicit, key documents and perspectives are collected, and systematic testing is conducted to thoroughly assess the potential risks and social impact of some AI service. In each stage there are key issues to consider and documents to complete, and Raji et al. (2020) provide sample

documents and a worked example to demonstrate the framework in operation. Although the worked example is fairly simple, these resources provide excellent materials for turning an AI audit into an extended classroom activity.

The activity plan below has adapted the audit framework from Raji et al. to fit the learning goals of an undergraduate AI Ethics course. Specifically, the framework has been simplified into two stages: the **Scoping stage**, responsible for laying out the scope of the audit, the ethical principles at stake, and the social impact assessment; and the **Testing stage**, responsible for conducting adversarial testing of the services to document potential risks and failure modes. These stages are discussed in more detail below.

Activity Setup

My class of 18 students is sorted into two groups of nine, each running parallel audits. One group was tasked with auditing text generation services (chatGPT, Copilot, Gemini), the other group was tasked with auditing image generation services (Dall-E, Copilot, Firefly, Midjourney). Within these groups students were sorted into a **Scoping Team**, responsible for the Scoping stage and scoping documents, and a **Testing Team**, responsible for the Testing stage and testing documents, with each team composed of 4-5 students. For a five week activity, teams are asked to deliver a weekly progress report as in-class presentations.

Warning: Since students will be running adversarial probes that may get flagged or banned by service moderators, it is strongly recommended that students open “burner” accounts for these services, rather than using their primary accounts.

Week 1: Pre-audit preparation
In the preparation week, Scoping and Testing teams should become familiar with the available generative AI services, and should commit to specific systems to target for the audit. If students have not used these services before, instructor-led demonstrations in class can be useful for highlighting some of the strengths and weaknesses of the services. It is recommended that groups run the same testing regime against several different services to compare their results and give some context for overall AI performance on those issues.

No formal presentations are expected in Week 1. Scoping and Testing teams should work together to agree on the specific issues or potential failure modes they want to probe with their audits, similar to formulating a research question. For instance, groups might decide to explore one of the following issues:

- **Quality and accuracy of results:** Does gen AI produce high quality essays? Can gen AI produce high quality research? Are its answers to questions in specific domains accurate?
- **Bias:** Does gen AI responses reflect biases in training data? Does it reflect or show sensitivity to cultural biases or stereotypes? What is the potential impact of these biases?

- **Safety:** Can gen AI be used to perform dangerous or illegal acts? What guardrails exist to prevent gen AI from assisting in dangerous activity? To what extent can these guardrails be circumvented?
- **Intellectual property:** Will gen AI reproduce copyrighted information, such as extended text passages from existing material, logos, mascots and animated characters, etc? What guardrails exist to protect this information, and to what extent can they be circumvented?

After agreeing on a set of issues, groups should write up the specific concerns they intend to probe with the audit, and should document their initial expectations of the AI performance.

Week 2: Preliminary Scoping Report
The Scoping Team is responsible for establishing the scope of the audit, and providing an analysis of the ethical stakes and the potential social impact of the AI service. The Scoping Team should identify the specific services targeted by the audit, and should lay out the issues, concerns, and failure modes that will be the focus of the audit. The Scoping Team should also provide some background analysis on the service and developers for the audit.

In Week 2, the Scoping team presents their preliminary work to the class. Some tasks and questions for the Scoping Team:

Put the service in context:
- Give some background on the service and its developers. How does the service work? How was the model trained? What data was used to train it? Is this information publicly available? What financial, political, environmental, etc circumstances are relevant for understanding the service?
- What use instructions are provided with the service? What ethical principles or safety concerns do the developers explicitly mention (on their website, for instance) as informing the development of their service?
- Beyond what the devs state explicitly, what ethical principles, safety policies, laws and regulations, or other concerns should inform an assessment of this service?

Preliminary Social Impact Assessment:
- How is the app typically used? Who are its typical users?
- What is the potential social impact (positive or negative) if the service operates according to expectations regarding the topic of your audit?
- What is the potential social impact (positive or negative) if the service fails to operate according to expectations?
- What are the worst case scenarios that might arise from these failures? How likely will the worst case scenarios happen?

Document your group's expectations:
- Which tasks do you expect that gen AI will perform well?
- Which tasks do you expect gen AI to perform poorly?
- How often do you expect to see explicit failures?

- How will students evaluate the performance of gen AI? What counts as performing well vs performing poorly?

Week 3: Preliminary Testing Report
The Testing Team is responsible for testing and assessing the performance of AI services regarding the issues and concerns developed by the Scoping Team. The Testing Team should develop a systematic testing regime that can be applied consistently and repeatedly across several different AI services. Individual tests should be run multiple times on the same service with small variations to ensure consistent, rigorous results. Be aware that for some gen AI services, prompt history may impact future results; this should be accounted for in running tests. The Testing Team is also responsible for developing an assessment framework for evaluating and comparing the performance of AI services. The prompts and results of tests should be carefully documented for the final report.

In Week 3, the Testing team presents their proposed testing regime and shares preliminary results. Some tasks and questions for the Testing Team:

Develop a rigorous testing regime and a clear framework for evaluation

- What prompts are you using? How many times do you repeat the prompts? How are you varying prompts across different tests? How many variations are you using? Does the service return consistent responses, or is it sensitive to small variations?
- How does the prompt/response history impact AI performance on the current prompt?
- How are you evaluating the results of your probe? How are you ensuring consistent, unbiased evaluation of the results? (eg, what counts as a "good deep fake"?)
- How are you comparing results across services? What challenges complicate these comparisons?

Document and assess AI performance and failure modes

- In what circumstances does the service perform to expectations? In what circumstances does the service exceed expectations? What is the benchmark for expectations?
- In what circumstances does the service demonstrate some error, bias, or other unambiguous mistake? In what circumstances does the service fail? How serious are these failures? Can these failures be reliably replicated?
- What guardrails did you encounter during testing? (eg, warning statements, prompt restrictions, banning, etc) Were these guardrails used effectively? How easy is it to get around these guardrails?

Week 4: Final Testing Report

By Week 4, the Testing Teams are expected to have completed their testing and assessed the final results, and should be prepared to present their findings to class. The Testing Team should complete their key documents, the FMEA chart and the Heatmap, and should compile their findings (prompts and responses) for a presentation and final report. The Testing team should provide an evaluation of their findings in light of the audit goals laid out by the scoping team.

Week 5: Final Scoping Report

In Week 5, the Scoping Team completes the audit by giving a final analysis of the testing results in light of the audit scope. The Scoping team should complete their documents, the Ethical Use Case Analysis and Social Impact Assessment, and should compile their analysis for a final presentation and the final report. The Scoping team should discuss their group's evaluation of the AI service performance, highlighting safety and ethical concerns. Finally, the Scoping Team should present the group's final Go/No-Go analysis:

- Is the service safe for continued public use, or should the service be taken off-line? How serious are the risks from misuse or other failure modes?
- What immediate changes to the service should be made to protect the safety and ethical interests of the public?
- How can the guardrails, warnings, and other precautions be improved to protect public safety and welfare?
- What other suggestions does the group have for developers or users of these services?

Final Report

Each group should compile their work into a single document that includes the research, testing results, analysis, and key documents from this project. All group members should contribute to the report, and the names of all contributing group members should appear on the front. **Please organize the report to be as readable as possible!** Finished reports will be made available to students in future semesters. Include research, commentary, and analysis at the front of the document. Format any in-text charts to make the document readable and understandable. Save any datasets (images, etc) and make them available either in an appendix or a supplementary document.

Post-Audit Podcast

After completing final presentations and submitting the final report, I ask students to record a short "podcast" conversation about the audit project. Groups are asked to discuss their responses to the following questions:

- Group performance:
    - What went well? What didn't go well?
    - What challenges did this project present? How did you address these challenges?

    - If you could start the project over, what would you do differently?
- What you learned:
    - What, if anything, surprised you about your findings on this project?
    - What insights have you gained about the success or failure modes of generative AI?
    - How has this project changed your attitudes towards AI?
    - Having completed this project, what is the most important thing you think the public should understand about generative AI?
    - Share any other thoughts you have about AI Ethics and the work you've done this semester.

The full conversation should take between 30-45 minutes.

Key Documents

The documents below are samples from the Smile Detection worked example from Raji et al (2020), demonstrating completed versions of these forms. I have included a blank sample of the documents for use in this activity.

Scoping:

- Ethical use case analysis (Blank template)
- Social Impact Assessment (Blank template)

Testing:

- Algorithmic failure modes and effects analysis (FMEA) chart (Blank template)
- Ethical risk chart heatmap (Blank template)

Appendix A: Reference material

- Raji et al. (2020). Closing the AI accountability gap
    - This paper lays out a complete framework for conducting internal audits. The paper links to a sparsely worked sample audit which includes templates for documentation, including the primary documents for this activity.
- Buolamwini and Gebru (2018) Gender Shades
    - This paper documents an audit of facial recognition software, demonstrating a systematic bias against accurate recognition of dark skinned female faces. The authors develop a novel database for evaluating facial recognition software in order to demonstrate the biases of existing off-the-shelf software and the failures of existing training databases.
- Raji and Buolamwini, J. (2019). Actionable auditing
    - This followup paper documents the impact of Gender Shades on the tech industry. Within a few months of publication, standard facial recognition software from major developed had been retrained on the new, more representative

database, demonstrating the potentially rapid response a successful AI audit might generate. These results helped to launch AI audits as an important tool in evaluating and regulating AI services. .

- Coded Bias (2020, documentary)
    - This Netflix documentary looks at Joy Buolamwini's experience as a grad student at MIT leading to the Gender Shades study. A great introduction to AI Ethics generally, and the social challenges it presents.
- The papers below represent some recent critical assessment of the effectiveness of AI audits, and especially of their use in "ethics washing", or creating the false appearance of ethical behavior. AI audits are not a complete solution to challenges in AI Ethics, and students should be aware of the limitations of this approach.
    - Costanza-Chock, Raji, & Buolamwini (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem
    - Keyes and Austin (2022). Feeling fixes: Mess and emotion in algorithmic audits
    - Birhane et al. (2024) AI Auditing: The Broken Bus on the Road to AI Accountability