Hidden Layer: Intellectual Privacy and Generative AI

Workshop Lesson Plan

- Workshop guide: https://guides.libraries.psu.edu/berks/AI/
- Workshop slides: https://docs.google.com/presentation/d/1sV7re95Y43ffx3ToQK_egWrpmtLea9eFd5ShibGF34M/edit?usp=sharing
  This workshop is designed for a 60-minute session as indicated by activity time stamps, but can be extended to fill the time available.

Learning Objectives

Facilitator learning objectives

During this workshop, participants will

- Apply prompt engineering techniques to elicit information from text-to-text generative AI (genAI) platforms
- Appreciate a range of intellectual privacy implications posed by genAI, including:
  - personal data;
  - intellectual property (copyright, patent, proprietary and sensitive data);
  - AI alignment (social bias, content moderation, AI guardrails, censorship, prompt injection);
  - synthetic media;
  - AI hallucination and mis/dis/malinformation; and
  - data sovereignty and data colonialism.
- Engage in a simulation to develop a conceptual understanding of how the hidden layer in the neural networks underpinning large language models works
- Synthesize their knowledge of genAI intellectual privacy considerations to analyze an ethical case study using the Agent-Impact Matrix for Artificial Intelligence (AIM4AI).

Participant learning outcomes

During this workshop, participants will

- Interact with genAI to explore its possibilities and limitations
- Discuss the intellectual privacy implications of genAI, including intellectual property considerations
- Evaluate the ethics of genAI for its impact on human agency

Learning Activities

Welcome (3 mins.)

Participant learning outcomes and agenda

[Activity] Think-Pair-Ask AI-Share (10 mins.)

This is a prompt engineering exercise in which participants will use text-to-text genAI platforms to elicit information about the privacy implications of AI.

Introduce the research question: What are the privacy implications of AI?

Direct participants to

1. **Think**: Brainstorm prompts on your own
2. **Pair**: Discuss prompts with a partner, and select 1-2 to prompt engineer and use.
    a. Provide participants with prompt engineering tips, such as
        i. ChatGPT Cheat Sheets from AIfire.co
        ii. CLEAR Framework for prompt engineering (Lo, 2023)
3. **Ask AI**: Explore the research question by prompting genAI using any of the following (or a platform of their choosing):
    a. AI Playground
    b. Perplexity.AI
    c. ChatGPT (login)
    d. Google Bard (login)
    e. Microsoft Copilot (login + Edge or Chrome)
4. **Share**: What did you learn from AI? *What did AI learn from you?*

Provide an online posting board (ex. Padlet) to preserve participant anonymity.

Ask participants to share their prompt and some GenAI output
Reflect on:

- What worked well?
- What did you need to tweak?
- Did you get the information you were looking for?
- Does it seem accurate? How do you know?
- What did you learn from AI?
- What did AI learn from you?

Facilitate a large-group discussion based on participant responses.

Inspired by "Think-Pair-Share with ChatGPT" proposed by Sarah Dillard.

Transition: What is generative AI and how does it work?

[Lecture] Introduction to generative AI (10 mins.)

Note: For details, refer to content and speaker notes in the workshop slides!

Explain ChatGPT, its relationship to large language models, and how they are trained.

Define neural networks and deep learning.

Introduce the Agent-Impact Matrix for Artificial Intelligence (AIM4AI) and the intellectual privacy implications of genAI:

> One axis of AIM4AI looks at agency on a spectrum from machine autonomy to human autonomy. Compared to other forms of model training, like supervised learning with trained data sets and target output values, the deep learning of neural networks is characteristic of machine autonomy.
>
> The other axis of AIM4AI considers impact on the spectrum from input to output. Input broadly refers to ways in which these models are trained or prompted, including deep learning, while output refers to the ways these models are used. As you learn about and interact with AI, think about ways that it can be used to enhance human agency by augmenting our intellectual activities, rather than progressing solely as an independent form of machine intelligence.

Review the Six Private I's Privacy Conceptual Framework (Hartman-Caverly & Chisholm, 2019) with a particular focus on the Intellect frame (Richards, 2015):

> Privacy is a critical element of human agency. The Six Private I's framework demonstrates six ways that privacy benefits us in everyday life, including by protecting our sense of identity, safeguarding our intellect and the activities of our mind, maintaining the contextual integrity of our personal information flows and our bodily integrity through spatial privacy and medical autonomy, securing the intimacy of our closest personal relationships, and preserving our freedom of association or interaction, as well as our ability to voluntarily withdraw into seclusion or isolation.
>
> This workshop will focus specifically on intellectual privacy, what Neil Richards calls "a zone of protection that guards our ability to make up our minds freely" (2015, p. 95). Intellectual privacy also protects your rights to your intellectual property, including any creative works that are eligible for copyright protection, or useful inventions that are eligible for patent protection.

Take a deeper dive into ChatGPT's history with a focus on training data.

Connect this back to intellectual privacy, including use of personal data and creative expressions in model training, the identity implications of AI output and hallucination, and intellectual property considerations.

Revisit AIM4AI by analyzing some case studies related to the impact of AI input and output on intellectual privacy.

Note: It is useful to demonstrate how the same case can be placed in a different quadrant on the matrix depending on whether it is considered from the perspective of input vs. output (impact) or machine autonomy vs. human autonomy (agent). In the example slide, the same article from Futurism about how

leaky prompts from Amazon employees probably divulged sensitive company information that appeared in ChatGPT output is used. It is analyzed as an example of both human autonomy in the input domain resulting in privacy harms, and of human autonomy in the output frame by applying ChatGPT to coding tasks to augment intellect.

It is recommended to conclude the AIM4AI analysis with an example of AI hallucination to facilitate the transition to the Hidden Layer Simulation.

Transition: If genAI applications like ChatGPT have access to so much information, why do they hallucinate?

[Activity] Hidden Layer Simulation (10 mins.)

Direct participants to access the Hidden Layer Simulation:

> For this simulation, our neural network has one input node, three parallel hidden layer nodes, and one output node.
>
> You will answer three questions to perform the analysis of the three hidden layer nodes, and answer a fourth and final question to predict the next token in the sequence as the output node.

Note: The Hidden Layer Simulation uses a source text written in Central Atlas Tamazight using the tifinagh script, an indigenous language of Morocco. This is an intentional choice to surface the issues of data sovereignty, data colonialism, and the language gap of large language models along with their implications for intellectual privacy. Tamazight is of personal significance to Hidden Layer Simulation creator Sarah Hartman-Caverly. An alternate language, real or fictitious, can be used, as long as it is likely to be unfamiliar to participants and the source text contains the same features that are used in the simulation (verse with repeated words).

Transition:

> This exercise simulates the activity of the hidden layer in a neural network to give you a conceptual understanding of how machine learning works in generative AI like ChatGPT.
>
> You were able to predict the next token in a text sequence, despite not being able to comprehend or interpret the input text. (In this case, the input text is an AI-generated translation of a nursery rhyme in Central Atlas Tamazight, an indigenous language of Morocco!)
>
> Similarly, **AI does not interpret, understand, or create meaning** - it is only performing sophisticated mathematical functions to predict the most likely desired output. It isn't magical - it's **math**emagical.

[Lecture] Math and Meaning (6 mins.)

Revisit earlier concepts about deep learning in large language models by emphasizing that they are manifestations of statistical computations over large bodies of text. These include algorithms for unsupervised machine learning, like data clustering.

Note: In the workshop slides, a data flow diagram for the transformer – the key component of GPT (generative pre-trained transformer) AI models – depicts some of the mathematical formulae that are programmed into this neural network architecture. The point is not to understand the math so much as to recognize that the math is there!

Emphasize that math does not know meaning.

Note: The workshop slides explore the process of generating the Central Atlas Tamazight translation of the nursery rhyme, Twinkle Twinkle Little Star, using Perplexity.ai. The purpose is to surface issues of data sovereignty, data colonialism, and the language gap of large language models along with their implications for intellectual privacy.

Introduce the concepts of data colonialism, data sovereignty, and the language gap.

Revisit AIM4AI by analyzing case studies related to data sovereignty, data colonialism, and the language gap as they relate to intellectual privacy.

Note: See workshop slides for examples.

Transition: What are the implications of language and other model training gaps for AI bias?

[Activity] AI Bias (5 mins.)

Direct participants to explore the image galleries in "How AI Reduces the World to Stereotypes" by Rest Of World.

Facilitate a brief discussion about their observations of AI bias based on the galleries.

Transition: Bias isn't only an artifact of the hidden layer of AI - it is also present in the 'human layer.' The decisions we make, from what data sets to use for model training, to how the data is labeled, to the selection and tuning of model parameters, to the evaluation of AI output for reinforcement learning, to the implementation of AI guardrails and other alignment, safety, and content moderation strategies can all introduce bias to generative AI.

[Activity] AIM4AI Case Study Analysis (12 mins.)

Reintroduce AIM4AI in the context of Neil Richard's definition for intellectual privacy:

> "a zone of protection that guards our ability to make up our minds freely" and "protection from surveillance or unwanted interference by others when we are engaged in the processes of generating ideas and forming beliefs" (2015, p. 5, 95).

Provide a curated collection of case studies in categories related to intellectual privacy like Alignment, Hallucination, Data Sovereignty, Intellectual Property, and Synthetic Media.

Direct participants to select and skim a case study and consider the following questions:

> Impact Dimension

1. Does the case address **input** or **output** from an AI system?
2. At what point does human-machine interaction occur in your case (ex. training during machine learning, fine-tuning, or in response to output)?

Agency Dimension

1. Who is doing the input - humans or machines?
2. Who is impacted by the output? How is the output evaluated for fairness, accountability, and transparency?
3. How transparent is the interaction? Does it enhance or undermine human agency?

Map your case onto the Agency-Interaction Matrix for AI (AIM4AI) (via Markup.io)

Provide an online posting board (ex. Padlet, Markup.io) to preserve participant anonymity.

Facilitate a large-group discussion based about the impact of AI on intellectual privacy and human agency based on participant responses.

Workshop review and closing (3 min.)

Assessment

1. This workshop taught me something new about **generative AI**, including its possibilities and limitations. [Likert scale 1 = strongly disagree 5 = strongly agree]
2. This workshop gave me a new way to think about **intellectual privacy**, including how it is impacted by generative AI. [Likert scale 1 = strongly disagree 5 = strongly agree]
3. This workshop gave me a new way to think about the **ethics** of generative AI, including how it can impact human agency and augment human intellect. [Likert scale 1 = strongly disagree 5 = strongly agree]
4. My **top takeaway** *or* **suggestion** for improvement is: [free-text response]