Enhancing AI-Centered Social Cybersecurity Education through Learning Platform Design

CISSE 2024, Tampa, FL

Nishant Vishwamitra[*], Ebuka Okpala[‡], Mohammed Aldeen[‡], Pranav Silimkhan[‡], Song Liao [§], **Keyan Guo**[†], Sandeep Shah [¶], Yongkai Wu[‡], Hongxin Hu[†], Xiaohong Yuan[¶] and **Long Cheng**[‡]

[*] The University of Texas at San Antonio™

[†] University at Buffalo, The State University of New York

[‡] Clemson University

[§]

[¶]

AI has Worsened Existing Social Issues

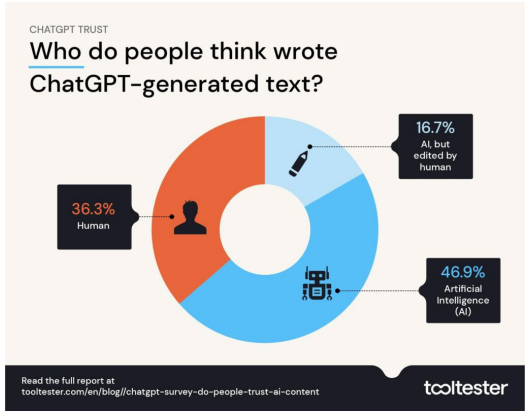
- Large language models (LLM) can automate hate speech, toxic language and disinformation
- StableDiffusion and DALL E can create realistic-looking non-consensual intimate images of women celebrities and social media users
- ChatGPT can compound social problems of fairness and ethics

Stable Diffusion 3

DALL·E

Efforts to Prevent GenAI-generated Harm

- Heightened anxiety and unease among governments, nations, and the research community



https://www.tooltester.com/en/blog/chatgpt-survey-can-people-tell-the-difference/

Italy curbs ChatGPT, starts probe over privacy concerns

By **Elvira Pollina** and **Supantha Mukherjee**

March 31, 2023 4:40 PM CDT · Updated 2 years ago

https://www.reuters.com/technology/italy-data-protection-agency-opens-chatgpt-probe-privacy-concerns-2023-03-31/

Schumer launches 'all hands on deck' push to regulate AI

The Senate leader urged lawmakers to advance 'comprehensive' legislation in coming months, amid pressure from critics for Congress to act

https://www.washingtonpost.com/technology/2023/06/21/ai-regulation-us-senate-chuck-schumer/

MAY 04, 2023

FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety

BRIEFING ROOM › STATEMENTS AND RELEASES

https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/

Motivation

- *Interdisciplinary Lens*: Need to engage students by integrating AI and social cybersecurity through an interdisciplinary approach
- *Experiential Learning*: Need Hands-on learning
- *Diversity*: Need to cater to a diverse audience

Lack of Education Materials

Our Approach

- Education Platform that integrates AI and social cybersecurity through an interdisciplinary approach
- Our Approach:
  - AI-based socially-relevant cybersecurity hands-on labs and education materials
  - In-depth understanding of social security problems and AI techniques through their own experimentation
- Hands-on labs cover different dimensions of AI-based cyberharassment detection systems and demonstrate the interplay between AI and cybersecurity:
  - AI for social cybersecurity
  - Vulnerabilities and social issues in AI algorithms

Research Questions

RQ1: How can AI-centered social cybersecurity education be enhanced for a diverse audience?

- Easy-to-use Learning Platform:
  - Labs based on PyTorch using the Jupyter Notebook on Google Colab platform

- Dual role of AI in social cybersecurity:
  - Labs 1-3 focus on how AI can be used to defend against social cybersecurity problems
  - Labs 4-6 focus on how AI could cause or exacerbate these problems

Research Questions

RQ2: What is the effectiveness of the novel learning platform in teaching AI-centered social cybersecurity concepts?

- Large-scale user study:
  - Students' understanding of AI-based social cybersecurity issues and their mitigation
  - Enthusiasm and interest in the topic before and after doing the labs
  - Use these results to measure the effectiveness of our learning platform in achieving our education goals

Lab Module Overview

- Labs consist of six lab modules that cover four dimensions of AI cybersecurity:
  - Positive use of AI for detecting cyberharassment and cyberbullying
  - Vulnerabilities of AI-classifiers for cyberharassment detection and negative use of AI (e.g., deepfakes) in engendering cyberharassment
  - Social issues in AI models for cyberharassment detection (e.g., bias, fairness, and trustworthiness)
  - Strengthening AI models for robust cyberharassment detection

Hands-on Labs

- Lab 1: AI for Text-based Cyberbullying Moderation
- Lab 2: AI for Multimodal Cyberbullying Moderation
- Lab 3: Interpretability of AI for Cyberbullying Detection
- Lab 4: Adversarial Attacks on Harmful Image Detection
- Lab 5: Disparity in AI-based Cyberharassment Models
- Lab 6: Debiasing AI-based Cyberharassment Models

Example Lab Module

The lab module consists of four steps:

- a brief introduction
- a broad presentation of general machine learning knowledge
- an introduction to cyberbullying and automated cyberbullying detection
- the steps to launch the AI-based cyberbullying detection on the Google Colab platform

Screenshot of Lab 1 on the Google Colab platform

Case Study of Text-based Cyberbullying Detection Lab

- We conducted pilot studies using Lab 1 and Lab 2 at N.C. A&T (the nation's largest HBCU) and Clemson University
- 201 students participated in our user study, including 123 male students and 78 female students, among which 58 participants were African American students
- we conducted both pre-survey and post-survey on the Qualtrics platform

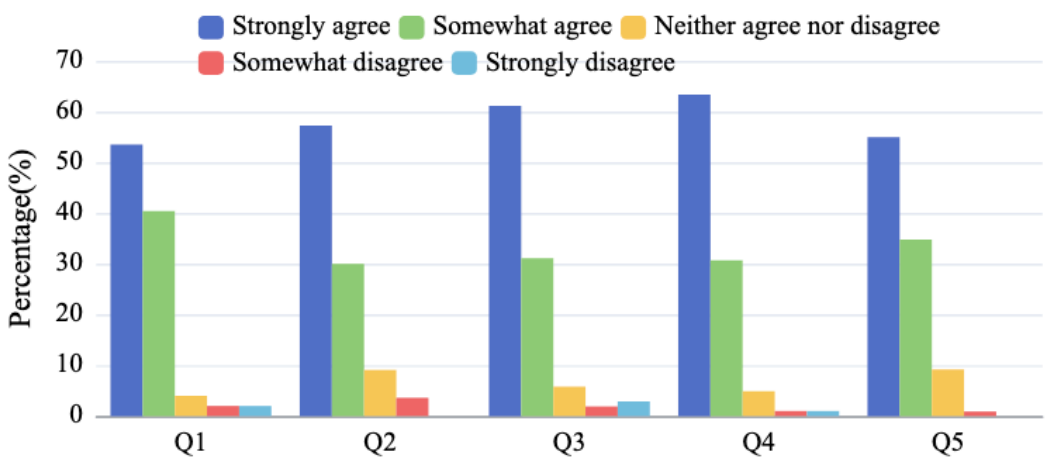
University	Course	Semester	Participants
N.C. A&T	COMP 365 AI and Machine Learning	Spring 2022	25
		Fall 2022	25
	SOCI 203 Social Statistics	Spring 2022	9
		Fall 2022	16
Clemson University	CPSC 4200/6200 Computer Security Principle	Summer 2022	11
		Fall 2022	39
		Spring 2023	76

Number of participants in our pilot studies

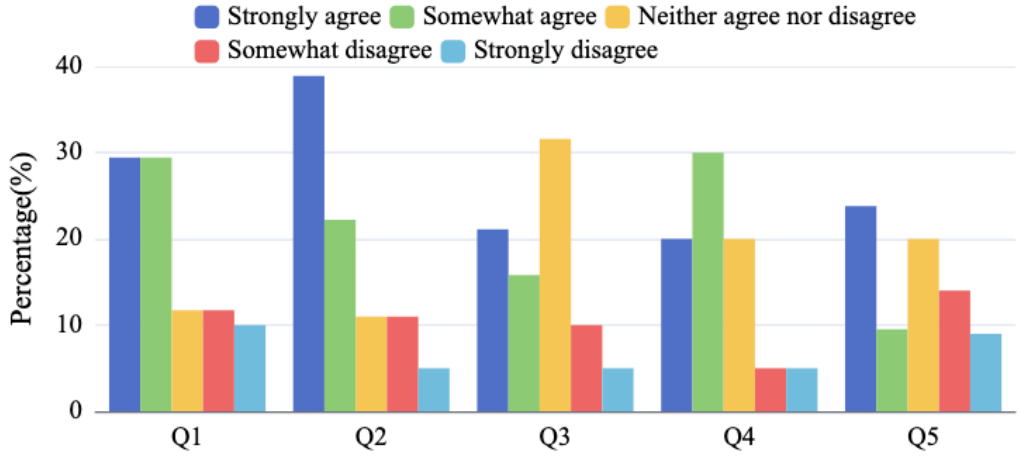
Demographic	CS	Non-CS
	Percentage	Percentage
Male	68.2%	12%
Female	31.8%	88%
White	13.1%	8%
African American	26.1%	88%
Asian	57.4%	4%
American Indian	3.4%	0

Demographics of students who participated

Post Survey



CS Students



Non-CS Students

Post survey results of both CS and Non-CS Students Show Improvement in Understanding AI Social Cybersecurity

Effectiveness Analysis for CS Students

Pre- /Post-Survey Question	Univ. A (CS)			Univ. B (CS)			Univ. A (Non-CS)		
	Pre.	Post.	Improve	Pre.	Post.	Improve	Pre.	Post.	Improve
1) Automated Cyber Harassment Detection	3.94	3.25	17.5% **	2.86	2.29	19.9% ***	4.38	3.31	24.4% *
2) State-of-The-Art Toxic Content Detectors	4.13	3.26	21.1% ***	2.98	2.45	17.8% **	4.71	3.44	27% **
3) How Machine Learning Works	3.13	2.83	9.6%	2.44	2.15	11.9% *	4.43	3.13	29.3% **
4) Cyberbullying Detection in images	4.15	2.78	33% ***	2.73	2.06	20.9% ***	-	-	-
5) AI-based classifier models selection	4.07	3.11	23.6% **	2.74	2.31	15.7% **	-	-	-
6) How to fine-tune an AI-based classifier	4.22	3	28.9% ***	2.83	2.35	17% **	-	-	-
7) What are true positive, true negative, false positive and false negative	3.7	2.56	30.8% **	2.45	2.09	14.7% *	-	-	-
8) Evaluation metrics (precision, recall, F1) of an AI classifier model	4.11	2.89	29.7% ***	2.6	2.12	18.5% **	-	-	-

Note. * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$

The designed labs are effective in enhancing the knowledge of Computer Science students in the areas of security

Effectiveness Analysis for Non-CS Students

Pre- /Post-Survey Question	Univ. A (CS)			Univ. B (CS)			Univ. A (Non-CS)		
	Pre.	Post.	Improve	Pre.	Post.	Improve	Pre.	Post.	Improve
1) Automated Cyber Harassment Detection	3.94	3.25	17.5% **	2.86	2.29	19.9% ***	4.38	3.31	24.4% *
2) State-of-The-Art Toxic Content Detectors	4.13	3.26	21.1% ***	2.98	2.45	17.8% **	4.71	3.44	27% **
3) How Machine Learning Works	3.13	2.83	9.6%	2.44	2.15	11.9% *	4.43	3.13	29.3% **
4) Cyberbullying Detection in images	4.15	2.78	33% ***	2.73	2.06	20.9% ***	-	-	-
5) AI-based classifier models selection	4.07	3.11	23.6% **	2.74	2.31	15.7% **	-	-	-
6) How to fine-tune an AI-based classifier	4.22	3	28.9% ***	2.83	2.35	17% **	-	-	-
7) What are true positive, true negative, false positive and false negative	3.7	2.56	30.8% **	2.45	2.09	14.7% *	-	-	-
8) Evaluation metrics (precision, recall, F1) of an AI classifier model	4.11	2.89	29.7% ***	2.6	2.12	18.5% **	-	-	-

Note. * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$

Marked improvement in the average knowledge scores across each question for non-CS students

Conclusion

- Open learning platform for AI-centered social cybersecurity
- AI to detect cyberharassment, as well as cybersecurity issues instigated by AI
- Improved student understanding

Thank you !

Keyan Guo,
PhD Candidate

keyanguo@buffalo.edu

Davis Hall,
Buffalo, NY 14260,
United States

Survey Questions

Index	Question	Stage
Q1	The lab engaged me in learning the topic of AI-Driven Socially-Relevant Cybersecurity.	Post Survey
Q2	I enjoyed the learning experience of this lab(s).	Post Survey
Q3	I think the learning experience with the lab(s) is effective.	Post Survey
Q4	I am satisfied with the level of effort the lab requires for learning this topic.	Post Survey
Q5	After using the lab(s), I have more confidence in describing the concepts learned.	Post Survey
Q6	In the following section, please rate your level of knowledge or skills: 1) Automated Cyberharassment Detection; 2) State-of-The-Art Toxic Content Detectors; 3) How Machine Learning Works; 4) Cyberbullying Detection in images; 5) AI-based classifier models selection; 6) How to fine-tune an AI-based classifier; 7) What are true positive, true negative, false positive, and false negative; and8) Evaluation metrics (precision, recall, F1) of an AI classifier model	Pre and Post Surveys.
Q7	What has been most helpful for your learning in using the lab(s) so far?	Post Survey
Q8	In terms of your learning, what has caused you the most difficulty in using the lab(s) so far?	Post Survey
Q9	What suggestion(s) can you make that would enhance your learning experience with the lab(s)?	Post Survey

TABLE III: List of survey questions.