Enhancing AI-Centered Social Cybersecurity Education through Learning Platform Design

Nishant Vishwamitra
Information Systems and Cyber Security
The University of Texas at San Antonio
San Antonio, USA
0000-0002-3728-1921

Ebuka Okpala
School of Computing
Clemson University
Clemson, USA
0000-0002-5816-8194

Song Liao
Department of Computer Science
Texas Tech University
Lubbock, USA
0000-0002-5264-7573

Keyan Guo
Department of Computer Science and Engineering
University at Buffalo
Buffalo, USA
0000-0001-9961-2442

Sandeep Shah
Department of Computer Science
North Carolina A&T State University
Greensboro, USA
0000-0001-9076-6162

Hongxin Hu
Department of Computer Science and Engineering
University at Buffalo
Buffalo, USA
0000-0001-8710-247X

Xiaohong Yuan
Department of Computer Science
North Carolina A&T State University
Greensboro, USA
0000-0002-1295-9812

Long Cheng
School of Computing
Clemson University
Clemson, USA
0000-0003-1736-0873

Abstract—Artificial Intelligence (AI) technologies have become increasingly pervasive in our daily lives. Recent breakthroughs such as large language models (LLMs) are being increasingly used globally to enhance their work methods and boost productivity. However, the advent of these technologies has also brought forth new challenges in the critical area of social cybersecurity. While AI has broadened new frontiers in addressing social issues, such as cyberharassment and cyberbullying, it has also worsened existing social issues such as the generation of hateful content, bias, and demographic prejudices. Although the interplay between AI and social cybersecurity has gained much attention from the research community, very few educational materials have been designed to engage students by integrating AI and socially relevant cybersecurity through an interdisciplinary approach. In this paper, we present our newly designed open-learning platform, which can be used to meet the ever-increasing demand for advanced training in the intersection of AI and social cybersecurity. The designed platform, which consists of hands-on labs and education materials, incorporates the latest research results in AI-based social cybersecurity, such as cyberharassment detection, AI bias and prejudice, and adversarial attacks on AI-powered systems, are implemented using Jupyter Notebook, an open-source interactive computing platform for effective hands-on learning. Through a user study of 201 students from two universities, we demonstrate that students have a better understanding of AI-based social cybersecurity issues and mitigation after doing the labs, and they are enthusiastic about learning to use AI algorithms in addressing social cybersecurity challenges for social good.

Keywords*—Artificial Intelligence (AI), Social Cybersecurity, AI Bias and Prejudice, Cyberharassment Detection, Hands-on Learning Platform*

I. INTRODUCTION

Artificial Intelligence (AI) technologies have been steadily permeating our everyday lives. From AI-powered chatbots such as ChatGPT [3] and Bard [6] to image generation AI such as StableDiffusion [12] and DALL·E 2 [5], AI-powered technologies are being globally used by people to optimize their work strategies and enhance productivity [35]. For instance, OpenAI's ChatGPT achieved 100 million monthly active users in January 2023, making it the fastest-growing application in history [30]. While the rise of these technologies has expanded new frontiers in addressing social cybersecurity [9] challenges such as the detection of cyberharassment (e.g., cyberbullying and cyberhate) [41], they have also exacerbated other challenges. For instance, Large language models (LLM), such as ChatGPT [3], can be potentially used to automate the creation and dissemination of large amounts of hate speech and toxic language and generate disinformation at an unprecedented scale [48]. Generative AIs such as StableDiffusion [12] and DALL•E 2 [5] have been used to create realistic-looking nonconsensual intimate images of women celebrities and social media users [13]. Concerns have also been raised about the capability of ChatGPT [3] in compounding social problems of fairness and ethics through the use of biased training data [32], [34]. The threat of AI on the social security and safety of our cyberspaces has heightened anxiety and unease among governments, nations, and the research community. The Italian government has recently blocked ChatGPT citing privacy concerns [4], proposals calling for the regulation of generative AI development have been recently introduced in the US senate owing to national security concerns [11], and researchers have called for a 6-month moratorium on training systems that are "more powerful than GPT-4" [1]. Moreover, a recently published White House fact sheet has called for urgent actions to promote *socially secure* generative AI use for the public good [14].

We argue that AI has immense potential to solve critical social cybersecurity problems, but also has the capability to exacerbate these problems further. Automatic detection methods of both text-based and image-based cyberbullying using AI techniques have emerged [46]. Internet companies such as Facebook and Google have also deployed AI algorithms to detect toxic content on social media [2], [7]. Meanwhile, adversaries may exploit vulnerabilities of AI-based classifiers to evade existing cyberharassment detectors [25], [31], [45]. In addition, there exist *social problems*, such as fairness and ethics, in AI models for cyberharassment detection. For example, some particular demographic groups are unfairly treated by AI-based detectors [39]. Concerns have been raised that the vulnerabilities of AI models as well as the robustness against attacks are biased towards underrepresented groups [34]. As such, an unfair AI-based cyberharassment detection system may perpetuate and aggravate existing prejudices and inequalities in society.

Despite progress in the research community, very few educational materials have been designed to engage students by integrating AI and social cybersecurity through an interdisciplinary approach. This paper presents our initial progress in developing AI-based socially-relevant cybersecurity hands-on labs and education materials, which provide students with an in-depth understanding of social security problems and AI techniques through their own experimentation. Our labs cover different dimensions of AI-based cyberharassment detection systems and demonstrate the interplay between AI and cybersecurity: i) AI for social cybersecurity, and ii) vulnerabilities and social issues in AI algorithms. These labs have been developed using IPython-based Jupyter Notebook (a web-based interactive software development environment) and can be made available using several cloud-based platforms, such as Google's Colaboratory (Colab) [19] and CloudLab [23]. These labs function as standalone modules, allowing flexibility in their integration into different curricula. Instructors can choose to incorporate them as independent learning units, supplementing relevant course topics or using them as optional enrichment activities.

We have conducted pilot studies using the text-based cyber-harassment detection lab at two universities and 201 students participated in surveys before and after taking the lab. Our survey results demonstrate that students have a better understanding of AI-based cyberharassment detection after participating in our pilot lab, and they also acquired research interests in using AI algorithms to address cyberharassment issues. We maintain a project website for our hands-on labs and the detailed instructions for our labs and datasets are available at https://cuadvancelab.github.io/. It is the first version and we will keep updating them as necessary.

II. DESIGN OF AI-CENTERED SOCIAL CYBERSECURITY LEARNING PLATFORM

We drive our work with the following research questions:

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

An important objective of our design is to build an easy-to-use learning platform to foster AI-centered social cybersecurity education and research. To this end, we develop our labs based on PyTorch [37] programming framework using the Jupyter Notebook [8] on Google Colab platform [19]. The PyTorch framework features high extensibility and customization, thereby allowing quick additions of new labs and seamless inclusion of several necessary algorithms used in our labs. Google Colab is a cloud-based Jupyter Notebook that enables users to run and execute Python code remotely with minimal setup. Its cell-based design allows for easy visualization of code outputs, which is useful for debugging machine learning models. It can be easily integrated with machine learning frameworks, such as TensorFlow, and PyTorch, and data exploration tools, such as Pandas. Importantly, users have free access to GPUs required for training machine learning models. Furthermore, we develop two versions of each lab, suitable for a student audience with a technical background and without a technical background, respectively.

A secondary objective of our work is to teach and demonstrate a dual role of AI in social cybersecurity - one where AI is used to defend against social cybersecurity threats, such as detecting cyberbullying and online hate, and the other where AI is misused to perpetrate or exacerbate social cybersecurity issues, such as bias and disparity. To this end, labs 1-3 focus on how AI can be used to defend against social cybersecurity problems, and labs 4-6 focus on how AI could cause or exacerbate these problems.

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

To address the second research question, we conducted a large-scale user study consisting of students from diverse backgrounds from two public universities in the United States. In the study, we measure students' understanding of AI-based social cybersecurity issues and their mitigation, as well as their enthusiasm and interest in the topic before and after doing the labs, and use these results to measure the effectiveness of our learning platform in achieving our education goals.

A. Lab Module Overview

The AI-based cyberharassment detection labs consist of six lab modules that cover four dimensions of AI-cybersecurity: 1) positive use of AI for detecting cyberharassment and cyberbullying; 2) vulnerabilities of AI-classifiers for cyberharassment detection and negative use of AI (e.g., deepfakes [42]) in engendering cyberharassment; 3) social issues in AI models for cyberharassment detection (i.e., bias, fairness, and trustworthiness); and 4) strengthening AI models for robust cyberharassment detection. The objectives are to teach students how to utilize AI techniques to detect cyberharassment, the vulnerabilities of these algorithms to adversarial attacks, and potential social problems in AI models.

Each lab module offers a detailed problem definition, learning goal, quizzes, and brainstorming questions (challenging tasks). Quizzes allow us to evaluate how well students understand each lab and the concepts involved. The brainstorming questions are designed to train students to think critically about future research needed to address the limitations of the current solutions and encourage them to participate in research on AI-centered cybersecurity. In what follows we briefly describe the six lab modules. More specifically, we tailored each lab module to cater to two distinct student audiences. This means that each lab comes in two different versions. One version is designed for students with little to no background in computer science. In this version, we aim for students to gain foundational knowledge in machine learning and AI-related cybersecurity, thereby appreciating the pivotal role AI can play in our digital ecosystem, understanding the risks that exist, and exploring new ideas in AI-centered social cybersecurity in light of their own unique backgrounds. The other version is tailored for students majoring in computer science or those with substantial programming skills. For these students, the lab introduces more advanced machine-learning concepts. Furthermore, students are expected to engage in programming tasks to fully realize the lab's objectives.

1) *Lab 1: AI for Text-based Cyberbullying Moderation*

Problem Definition: Cyberbullying involves perpetrators bullying victims by sending or sharing negative, harmful, false, or mean content, where the text (e.g., text message, email, or tweet) is the most common form used by perpetrators. As more teens increase their online presence, especially social media usage, the number of cyberbullying victims is expected to increase. Automatically detecting cyberharassment makes it easier and faster to protect the most vulnerable and enables platforms to give their users the ability to limit exposure to such toxic content.

Learning Objectives: Students will understand how to utilize an AI model to detect cyberharassment and understand the iterative nature of developing an AI model by experimenting with different hyperparameters. Students will gain firsthand experience in using AI models to distinguish cyberbullying from non-cyberbullying content, including model structure selection, feature selection, training, and prediction. Students will also learn the main metrics used in evaluating AI-based classifiers, such as false positives, false negatives, precision, recall, and F1 score.

2) *Lab 2: AI for Multimodal Cyberbullying Moderation*

Problem Definition: Recent technological advancements have led to a new cyberbullying paradigm, where perpetrators use visual media to bully their victims by sending and distributing images with cyberbullying content. The detection of multimodal cyberbullying is challenging given its highly personalized and contextual nature.

Learning Objective: For the visual factor identification and extraction, students will learn to use state-of-the-art tools to extract visual factors from images, e.g., using OpenPose [21] to estimate the body-pose and using OpenFace [17] to extract the emotions of a person in images. Students will learn how to combine low-level image features with high-level visual factors using feature fusion techniques. Students will also learn to employ deep neural networks (e.g., ResNet [27]) for cyberbullying detection in images.

3) *Lab 3: Interpretability of AI for Cyberbullying Detection*

Problem Definition: AI-based models have been deployed in practical applications to detect cyberharassment. Yet, much of their inner workings are still a "black box", i.e., it is not clear to a human, how an AI model makes predictions. The importance of interpreting AI predictions is more crucial in security applications such as cyberharassment detection, because of the far-reaching consequences of such predictions.

Learning Objective: Students will learn AI interpretability in both text and image-based models. Students will learn how to interpret text-based models with techniques such as Local Interpretable Modelagnostic Explanations (LIME) [38]. Students will also learn how to interpret the predictions made by image-based models using techniques such as Class Activation Maps [47] (CAM) and Gradient-based CAMs [40] (Grad-CAMs).

4) *Lab 4: Adversarial Attacks on Harmful Image Detection*

Problem Definition: While AI can be a useful tool in the fight against cyberharassment, it is vulnerable to adversarial attacks. For example, attackers may poison the dataset to trick the model or install backdoors so that the AI model operates normally until a trigger is presented to cause misclassification. Several attacks based on images and text, e.g., Fast Gradient Sign attack [24] (FGSM), DeepFool [33], and TextFooler [28] have demonstrated the vulnerability of AI systems to adversarial attacks.

Learning Objective: Students will obtain an understanding of potential vulnerabilities in AI systems. Students will learn how to craft adversarial attacks in both image and text-based cyberharassment detection systems. Students will also learn representative defense techniques against adversarial attacks in AI models.

5) *Lab 5: Disparity in AI-based Cyberharassment Models*

Problem Definition: AI is capable of enhancing cyber-harassment detection whereas it may inherit demographic prejudice from training data or aggravate social inequalities during the design and training phases unintentionally. Consequently, the cyberharassment from/against some groups may be misrecognized by such AI detection models, which leads to disparity and inequalities against minorities.

Learning Objective: Students will build/apply cyber-harassment detection models from/to real-world data and understand the demographic disparity of detection models against certain underrepresented groups. Students will investigate how demographic bias [22] and gender bias [20] inadvertently impact a cyberharassment detection model.

6) Lab 6: Debiasing AI-based Cyberharassment Models

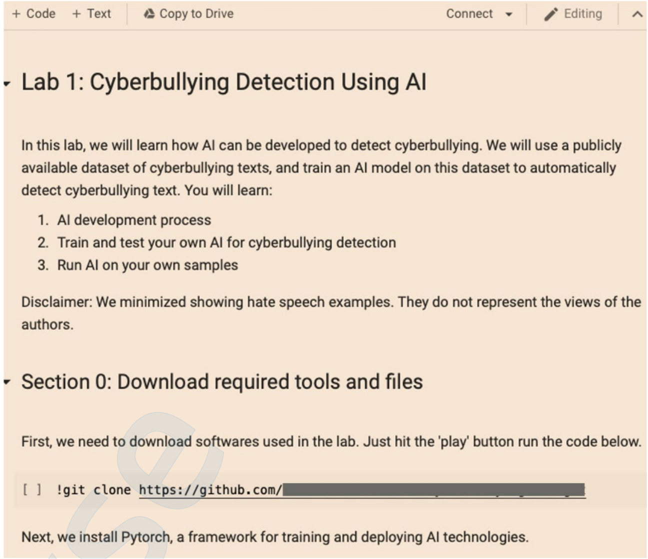
Problem Definition: Fairness-enhanced AI models have been extensively studied in traditional tasks, e.g., classification, regression, and dimensional reduction. It is imperative to mitigate bias from the unfair cyberharassment detection models. Fairness in cyberharassment detection can be enhanced by eliminating the bias from the training data and representation learning, incorporating debiasing constraints in AI models, and remedying unfair predictions made by biased AI models.

Learning Objective: In this lab, students will learn representative techniques for debiasing AI models, such as debiased word embeddings [20], adversarial debiasing [44], [18], and dynamic upsampling [15]. These techniques will engage students in understanding debiasing methods from several different perspectives, such as data representations, AI algorithms, and prediction correction.

B. Example Lab Module

In this section, we use Lab 1 (i.e., text-based cyberbullying detection) as an example to articulate the details of our lab design, and demonstrate how we use the lab to teach students how to use AI to moderate cyberbullying content. The lab module consists of four steps: 1) a brief introduction about the IPython-based Jupyter Notebook and Google Colab; 2) a broad presentation of general machine learning knowledge; 3) an introduction about cyberbullying and automated cyberbullying detection; and 4) the steps to launch the AI-based cyberbullying detection on the Google Colab platform. In Section 3, we present the results of our preliminary user studies using Lab 1 in two universities.

Our machine learning background tutorial covers the basics of AI, such as the concepts of classification, clustering, supervised learning, unsupervised learning, and semi-supervised their functionalities (including dataset preparation, embedding, model training, deployment, hyperparameter tuning, etc.). For example, the first section shown in Figure 1 includes the Python code to download required tools and files, such as installing PyTorch and importing software dependencies. With the cell-based design, students can run code cells step by step to learn how the AI-based cyberbullying detector is developed. Lab 1 allows students to tune hyper-parameters, such as learning rate, training epochs, number of neurons in the hidden layer, etc. Therefore, students can improve the performance of the AI model by themselves. In addition, the lab module is highly visualizable, e.g., students can observe the training process of the AI model, including the training time required for each epoch and the corresponding loss and accuracy.



Lab 1: Cyberbullying Detection Using AI

In this lab, we will learn how AI can be developed to detect cyberbullying. We will use a publicly available dataset of cyberbullying texts, and train an AI model on this dataset to automatically detect cyberbullying text. You will learn:

1. AI development process
2. Train and test your own AI for cyberbullying detection
3. Run AI on your own samples

Disclaimer: We minimized showing hate speech examples. They do not represent the views of the authors.

Section 0: Download required tools and files

First, we need to download softwares used in the lab. Just hit the 'play' button run the code below.

```
[ ]  !git clone https://github.com/[redacted]
```

Next, we install Pytorch, a framework for training and deploying AI technologies.

Fig. 1. Screenshot of Lab 1 on the Google Colab platform.

III. CASE STUDY OF TEXT-BASED CYBERBULLYING DETECTION LAB

We conducted pilot studies using Lab 1 and Lab 2 described in Section II-B at North Carolina A&T State University (which is one of the nation's top producers of African American STEM undergraduates) and Clemson University. In total, 201 students participated in our user study, including 123 male students and 78 female students, among which 58 participants were African American students. Table I lists the participants and courses they took in the user study. The demographics of our user study have been presented in Table II. At North Carolina A&T State University, we conducted four user studies involving 50 Computer Science (CS) students enrolled in COMP365 — AI and Machine Learning, and 25 no-CS students, who were from Social Science (SS), enrolled in SOCI203 — Social Statistics. Meanwhile, at Clemson University, we carried out three user studies with 126 students from the CS department who were part of the CPSC 4200/6200 — Computer Security Principle course. Given the time limitations in Spring 2022, we excluded lectures on general machine learning knowledge and automated cyberharassment detection during our initial user studies. This strategy provided us an opportunity to gauge the significance of the background introduction and refine both our labs and the associated survey questions, especially for non-CS students.

TABLE I. Number of participants in our pilot studies.

University	Course	Semester	Participants
North Carolina A&T State University	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	Spring 2022	11
		Fall 2022	39
		Spring 2023	76

TABLE II. Demographics of students who participated in Lab 1 and Lab 2.

Demographic	CS Percentage	Non-CS 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

TABLE III. List of 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; and 8) 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 IV. Sample t-test results for pre- and post-surveys.

	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 $p < .001$

A. *User Study Design*

To evaluate student interest in AI-driven socially-relevant cybersecurity, we conducted both pre-survey and post-survey based on the Qualtrics platform [10]. Table III lists the survey questions in our pilot study. The surveys consist of multiple-choice and open-ended questions. For Q1∼Q5, we used a 5-point scale to measure the student's interest and knowledge of cyberbully detection ranging from 1 (Strongly agree), 2 (Somewhat agree), 3 (Neither agree nor disagree), 4 (Somewhat disagree) to 5 (Strongly disagree). In Q6, the levels of knowledge/skills include 1 (Proficient), 2 (Good), 3 (Moderate), 4 (A little), and 5 (None). Q7∼Q9 are open-ended questions that can help us further improve our labs based on student comments. The IRB offices of both universities have approved the survey questions and the user study protocols.

B. *Effectiveness Analysis for CS Students*

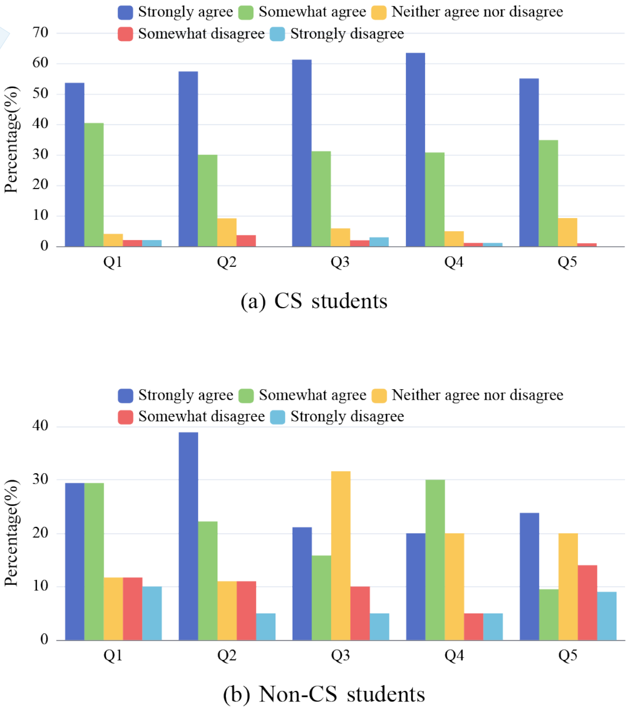
To analyze our designed labs' effectiveness, we compared the average knowledge score between pre-survey and post-survey. More specifically, Table IV depicts the results for the students majoring in Computer Science since the Fall 2022 semester. We skipped Spring 2022 because our lab was still in the development phase at that time, and the associated survey questions had not yet been completed. The results for the first three questions are from Lab 1 and the remaining data are from Lab 2. For both North Carolina A&T State University and Clemson University, there is a noticeable improvement in knowledge scores after students complete the labs. These results strongly support the effectiveness of the designed labs in enhancing the knowledge of Computer Science students in the areas of security. The majority of the subjects showed statistically significant improvements in both universities. Although the lack of statistically significant improvement for the topic *"How Machine Learning Works"* at North Carolina A&T State University, the overarching trend across topics reinforces the positive impact of the labs on the bulk of the students. It is pertinent to note that a considerable number of CS students had previously undertaken introductory machine-learning courses prior to our labs. Given that the pre-survey average knowledge score for this topic was the lowest—indicating a higher proficiency—it stands to reason that this particular subject witnessed a more muted enhancement compared to others.

In addition, we look into the result for each post-survey question to understand the feedback from the students. Figure 2a shows the user study results for CS students at Universities A and B. Most students were seniors or graduate students. More than 70% of the students gave positive feedback on all the questions. For Q1∼Q5 questions, over 90% of student feedback is positive. 53.5% of the students strongly agreed, and 40% of students somewhat agreed that they "enjoyed the learning experience of the lab". 87% of students agreed that the learning experience with these labs is effective, and 91.5% of students agreed that after using the lab, they have more confidence in describing the concepts learned. Concerning the knowledge/skills in Q6, there was a unanimous improvement. Many students transitioned from having no knowledge/skills to possessing some. These results reinforce the importance and relevance of utilizing such labs to bolster students' understanding of AI and social cybersecurity knowledge.

C. *Effectiveness Analysis for Non-CS Students*

Students from the Social Statistics class at North Carolina A&T State University were presented only with the questions corresponding to Lab 1 (questions 1–3 of Q6) from Table III due to the incompletion of Lab 2. As indicated by the pre-survey average knowledge scores, the majority of non-CS students possessed a minimal grasp of the relevant subjects. However, post-engagement with our bespoke lab, there was a marked improvement in the average knowledge scores across each question. Moreover, the accompanying p-values confirm the statistical significance of these improvements.

The feedback derived from our post-survey, as illustrated in Figure 2b, is predominantly positive. For Q1 and Q2, over 60% of the students concurred that our labs are engaging, expressing enjoyment in the learning journey. Further, 37% of the non-CS students deemed our labs effective for assimilating AI-centric social cybersecurity knowledge. Half of these students felt that the effort required by the lab was appropriate for grasping the associated subjects. Ultimately, 33.3% of non-CS students expressed an elevated confidence in articulating the concepts post-participation in our labs. Such results highlight the efficacy of our labs, emphasizing their potential to enhance learning outcomes, even for students with minimal exposure to computer science.



(a) CS students

(b) Non-CS students

Fig. 2. Post-survey results in our user study.

D. *Open-Ended Questions Analysis*

To continually refine and evolve our AI and social cybersecurity labs, we have dedicated three questions to gauge student feedback and identify areas of improvement. Table V showcases representative answers to Q7 and Q8 and suggestions for enhancing Lab 1. A majority of the students emphasized the utility of step-by-step instructions in facilitating their learning. However, there was feedback suggesting room for refinement in the instructions. Particularly, non-CS students encountered challenges with the specialized terminologies presented in the lab. Additional recommendations encompassed updating the code regularly to preempt future errors and infusing greater interactivity into the learning process. In general, students were enthusiastic about learning AI and social cybersecurity knowledge. While non-CS students enjoyed that the code was provided for them, allowing them to focus on the main concepts, CS students wanted the opportunity to complete some of the code themselves.

TABLE V. Representative responses to open-ended questions.

Question	**What has been most helpful for your learning in using the labs?**
Responses	Easier to understand, not complicated, clear instructions. (COMP 365)
	The step-by-step instructions as well as the visual materials. (COMP 365)
	Following everything step by step and running previous codes to make sure the rest work. (COMP 365)
	Being able to edit and rerun the code quickly to see how different settings affect the accuracy was very cool. (CPSC 4200/6200)

Question	**In terms of your learning, what has caused you the most difficulty in using the labs?**
Responses	The terminology. (SOCI 203)
	Understanding given code. (CPSC 4200/6200)
	Understanding all of the data preprocessing steps. (COMP 365)
	I didn't understand what was happening at first. (COMP 365)

Question	**What suggestion(s) can you make that would enhance your learning experience with the lab(s)?**
Responses	Make it more interactive. (SOCI 203)
	Spend more time with the basics of the lab. (CPSC 4200/6200)
	Please check that the code is in-date. We had to fix errors if some libs updated. (COMP 365)
	For me adding more explanation to the lab webpage, as I don't understand what is happening at first glance. (COMP 365)

E. *Limitation*

Our work, while promising, has its limitations. Our study is based on a limited sample size from two universities, both of which are public institutions in the United States. This confines the breadth of our findings and the generalizability. The modest number of participants, particularly from non-CS disciplines, further narrows the scope. We have primarily focused on sociology students, omitting potential interest from students in other areas like psychology and public health. Additionally, while our data indicates that our lecture introductions are beneficial, there might be external factors influencing the results that we have not yet identified. Moreover, our current work does not encompass some of the very latest AI technologies, such as ChatGPT.

IV. RELATED WORK

Cybersecurity education is now more important than ever, which is essential to protecting the national infrastructure, government, industry, and personal security and privacy. Švábenský *et al*. conducted a literature review of 71 cybersecurity education papers from SIGCSE and ITiCSE [43]. The authors found that the primary cybersecurity topics in existing works are secure programming, network security, offensive security (e.g., cyber-attacks and exploitation), human aspects (e.g., privacy and social engineering), cryptography, and authentication/authorization. Despite an increase in the literature on cybersecurity education [29], [16], to the best of our knowledge, there is no previous work on teaching and learning of AI-based socially relevant cybersecurity.

The integration of AI with social cybersecurity is a nascent area in both Computer Science and Social Science. Few studies have pursued this interdisciplinary approach. Our literature review revealed no prior studies specifically designing AI labs for cyberharassment detection. Despite the lack of literature in AI and social cybersecurity teaching modules, prior work proposed an AI-assisted cybersecurity course for malware analysis in real time [26], and hands-on labs to engage students in learning about Software-Defined

Networking (SDN) [36]. In [26], the authors developed a six-module course consisting of lectures and labs designed to teach students how to leverage AI in malware analysis. The authors in [36] developed five hands-on labs to enable students to learn SDN security issues. In a user study using two out of the five labs, survey analysis of the 35 students who participated in the study shows that more than 90% agree that the hands-on labs helped them understand the SDN security issues and 92% showed interest in SDN security research after completing the labs.

V. CONCLUSION

In this paper, we presented our newly designed open learning platform to meet the ever-increasing demand for advanced training in AI-centered social cybersecurity. Through the hands-on labs in our platform, students learn how to use AI to detect cyberharassment, as well as study the cybersecurity issues instigated by AI. Also, students learn how to tune AI models to improve prediction and the vulnerabilities in models trained to detect cyberharassment. Our user study results showed that students enjoyed the learning experience of the lab and were interested in learning about AI-centered socially-relevant cybersecurity.

REFERENCES

[1] 1,100+ notable signatories just signed an open letter asking ‘all AI labs to immediately pause for at least 6 months’. https://techcrunch.com/2023/03/28/1100-notable-signatories-just-signed-an-open-letter-asking-all-ai-labs-to-immediately-pause-for-at-least-6-months/.

[2] AI advances to better detect hate speech. https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech/.

[3] ChatGPT. https://openai.com/blog/chatgpt.

[4] ChatGPT banned in Italy over privacy concerns. https://www.bbc.com/news/technology-65139406.

[5] DALL•E 2. https://openai.com/product/dall-e-2.

[6] Google Bard. https://bard.google.com.

[7] Google’s Hate Speech Detection A.I. Has a Racial Bias Problem. https://fortune.com/2019/08/16/google-jigsaw-perspective-racial-bias/.

[8] Jupyter Notebook. https://jupyter.org.

[9] Microsoft Security Copilot. https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot.

[10] Qualtrics XM Platform. https://www.qualtrics.com/.

[11] Schumer Launches Major Effort To Get Ahead Of Artificial Intelligence. https://www.democrats.senate.gov/newsroom/press-releases/schumer-launches-major-effort-to-get-ahead-of-artificial-intelligence.

[12] Stable Diffusion. https://stability.ai/blog/stable-diffusion-public-release.

[13] Stable diffusion made copying artists and generating porn harder and users are mad. https://www.theverge.com/2022/11/24/23476622/ai-image-generator-stable-diffusion-version-2-nsfw-artists-data-changes. Accessed: 2022-11-24.

[14] White House Fact Sheet, 2023. https://bidenwhitehouse.archives.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/.

[15] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.

[16] Muhammad Rizwan Asghar and Andrew Luxton-Reilly. A case study of a cybersecurity programme: Curriculum design, resource management, and reflections. *In Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE)*, 2020.

[17] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Open-face: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[18] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint* arXiv:1707.00075, 2017.

[19] Ekaba Bisong. *Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners*. Apress, 2019.

[20] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[21] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint* arXiv:1812.08008, 2018.

[22] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint* arXiv:1905.12516, 2019.

[23] Dmitry Duplyakin, Robert Ricci, Aleksander Maricq, Gary Wong, Jonathon Duerig, Eric Eide, Leigh Stoller, Mike Hibler, David Johnson, Kirk Webb, Aditya Akella, Kuangching Wang, Glenn Ricart, Larry Landweber, Chip Elliott, Michael Zink, Emmanuel Cecchet, Snigdhaswin Kar, and Prabodh Mishra. The design and operation of cloudlab. *In USENIX USENIX Conference on Usenix Annual Technical Conference (ATC)*, page 1–14, 2019.

[24] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint* arXiv:1412.6572, 2014.

[25] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[26] Maanak Gupta, Sudip Mittal, and Mahmoud Abdelsalam. Ai assisted malware analysis: A course for next generation cybersecurity workforce. *arXiv preprint* arXiv:2009.11101, 2020.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[28] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025, 2020.

[29] Ge Jin, Manghui Tu, Tae-Hoon Kim, Justin Heffron, and Jonathan White. Game based cybersecurity training for high school students. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE)*, 2018.

[30] Krystal Hu. ChatGPT sets record for fastest-growing user base. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Accessed: 2023-02-02.

[31] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA*, February 24-27, 2019. The Internet Society, 2019.

[32] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[33] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[34] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. *CoRR*, abs/2006.12621, 2020.

[35] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Available at SSRN 4375283*, 2023.

[36] Younghee Park, Hongxin Hu, Xiaohong Yuan, and Hongda Li. Enhancing security education through designing sdn security labs in cloudlab. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 2018.

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, editors, Advances in *Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[38] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why Should I Trust You?”: Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 13-17, 2016, pages 1135–1144. ACM, 2016.

[39] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics.

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[41] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. *In IEEE Symposium on Security and Privacy (SP)*, pages 473–493, 2021.

[42] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131 – *arXiv preprint arXiv:2001.00179*, 2020.

[43] Valdemar Švábenský, Jan Vykopal, and Pavel. What are cybersecurity education papers about? a systematic literature review of sigcse and iticse conferences. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE)*, 2020.

[44] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[45] Wei Emma Zhang, Quan Z. Sheng, Ahoud Abdulrahmn F. Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3):1–41, 2020.

[46] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI’16, page 3952–3958, 2016.

[47] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In Proceedings of the *IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[48] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In Proceedings of the 2023 *CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.