CYBERSECURITY AND PRIVACY INSTITUTE
UNIVERSITY OF WATERLOO

Digital Overground

Cybersecurity and Privacy Institute Student Newsletter

If we say that Spring has sprung, does that mean Fall has fallen? Also, as Halloween is coming up, make sure to trick-or-treat yourself to some candy and costumed silliness! A little sugar rush whilst dancing around in a giant banana costume has been proven to lower stress, enhance learning, and increase your awesomeness on a cellular level; science for the win!

With the leaves changing and the weather cooling off, we would like to remind you to always bring a sweater, update your passwords, and enjoy your October!

If you are interested in contributing to this newsletter, please email us at CPI Students <cpi.students@uwaterloo.ca> we welcome the help!

Upcoming Events

Treaty Girl exhibit at Longhouse Labs

Open Access Week: Reproducibility and Replicability in Research

Open Access Week: Increasing Research Impact and Academic Prestige through Open Access Publishing

Bridge 2024: Honouring the Lives of Missing and Murdered Indigenous Women, Girls, and Two Spirit People

Smartizen Halloween party

To Hell with the 90's

Cyclist Workshop

Bloody Berlin Walking Tour

Student Support and Resources

Campus Wellness and Counselling Services

CPI for Students

Current Students Pathways

CPI Undergraduate Award

CPI Excellence Graduate Scholarship

The Vector Digital Talent Hub

Research

Out of the Ordinary: Spectrally Adapting Regression for Covariate Shift
Benjamin Eyre, CPI Member Elliot Creager,
David Madras, Vardan Papyan, &Richard Zemel

The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks
Kent F. Hubert, Kim N. Awa & Darya L. Zabelina

Managing Heterogeneous Datacenters with Tokens

CPI Member Seyed Majid Zahedi, Songchun Fan, & Benjamin C. Lee

FITS: Inferring Intermediate Taint Sources for Effective Vulnerability Analysis of IoT Device Firmware
Puzhuo Liu, Yaowen Zheng, CPI Member Chengnian Sun, Chuan Qin, Dongliang Fang, Mingdong Liu, & Limin Sun

Reinforcement Learning and Collusion
CPI Member Clemens Possnig

Remembering to Be Fair:
Non-Markovian Fairness in Sequential Decision Making
Parand A. Alamdari, Toryn Q. Klassen, CPI Member Elliot Creager, & Sheila A. McIlraith

Distributed Strategies for Computational Sprints
Songchun Fan, CPI Member Seyed Majid Zahedi, & Benjamin C. Lee

An Empirical Study of Data Disruption by Ransomware Attacks
Yiwei Hou, Lihua Guo, Chijin Zhou, Yiwen Xu, Zijing Yin, Shanshan Li, CPI Member Chengnian Sun, & Yu Jiang

Learning to Best Reply:
On the Consistency of Multi-Agent Reinforcement Learning
CPI Member Clemens Possnig

Open Calls

The Vector Digital Talent Hub encourages students to create profiles on their website to apply for a variety of employment opportunities. | Vector Institute

ICITST 2024 : International Conference for Internet Technology and Secured Transactions

New York Annual Conference on Cyber Security 2024
December 14-15, 2024,
New York City

International Journal on Cybernetics & Informatics (IJCI)

WatITis 2024 Conference

In the Media

- **Podcast of the Month: Cybersecurity Today: Wayback Machine Read-Only, AI-Driven Phishing, and Quantum Computing Breakthroughs - In this episode of Cybersecurity Today, host Jim Love discusses the recent cyber incident with the Internet Archive's Wayback Machine, which is now back online in read-only mode. He outlines sophisticated AI-**

driven Gmail phishing schemes that are fooling even tech experts and reports on Chinese researchers' breakthrough using a Canadian quantum computer to potentially crack military-grade encryption. Jim also shares practical advice on staying vigilant against such cyber threats.

- AI begins its ominous split away from human thinking
- Warning! This is how cars are hacked. Just like in Mr Robot.
- How to Create a Beautiful Python Visualization Dashboard with Panel/Hvplot
- Create An AI Song and Music Video That's Actually Good
- AI Models in Cybersecurity: From Misuse to Abuse
- Gryphon Healthcare, Tri-City Medical Center Disclose Significant Data Breaches
- Mastercard to Acquire Threat Intelligence Firm Recorded Future for $2.6 Billion
- Organizations Faster at Detecting OT Incidents, but Response Still Lacking: Report
- How Lessons Learned From the 2016 Campaign Led US Officials to Be More Open About Iran Hack
- Homebrew Security Audit Finds 25 Vulnerabilities

Seen anything that you think should be on this list for our next edition? Let us know!

CPI Students <cpi.students@uwaterloo.ca>

Student Spotlights

Optimizing Adaptive Attacks Against Content Watermarks for Language Models

Abdulrahman Diaa, Toluwani Aremu, Florian Kerschbaum, Nils Lukas

UNIVERSITY OF WATERLOO | DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Our October Student Spotlight features the CPI Poster competition winner, **Abdulrahman Diaa**, Supervisor: Florian Kerschbaum CS, with their work entitled: **Optimizing Adaptive Attacks Against Content Watermarks for Language Models**

Large Language Models (LLMs) can be \emph{misused} to spread online spam and misinformation. Content watermarking deters misuse by hiding a message in model-generated outputs, enabling their detection using a secret watermarking key. Robustness is a core security property, stating that evading detection requires (significant) degradation of the content's quality. Many LLM watermarking methods have been proposed, but robustness is tested only against \emph{non-adaptive} attackers who lack knowledge of the watermarking method and can find only suboptimal attacks. They formulate the robustness of LLM watermarking as an objective function and propose preference-based optimization to tune \emph{adaptive} attacks against the specific watermarking method. Their evaluation shows that (i) adaptive attacks substantially outperform non-adaptive baselines. (ii) Even in a non-adaptive setting, adaptive attacks optimized against a few known watermarks remain highly effective when tested against other unseen watermarks, and (iii) optimization-based attacks are practical and require less than seven GPU hours. Their findings underscore the need to test robustness against adaptive attackers.

Find Out More about Digital Overground

Our mailing address is:

200 University Ave W. DC 3147 Waterloo, Ontario N2L 3G1