Sahar Abdelnabi

AI/ML Security and Safety Researcher

sahar.s.abdelnabi@gmail.com
https://s-abdelnabi.github.io/

Research Focus

I am interested in the broad intersection of machine learning with security, safety, and sociopolitical aspects. This includes the following areas: 1) Understanding, probing, and mitigating the failure modes of machine learning models, their biases, and their misuse scenarios. 2) How machine learning models could amplify or help counter existing societal and safety problems (e.g., misinformation, biases, and stereotypes). 3) Emergent safety challenges posed by new foundation and large language models.

Current Position

AI Security Researcher *2024 - Ongoing*

Microsoft Security Response Center (MSRC), Microsoft Research Cambridge, UK
Role:
- Conducting AI security and safety research
- Assessing AI security vulnerabilities reported through Microsoft's AI Bug Bounty program

Education

Ph.D. in Computer Science *2019 - 2024*

CISPA Helmholtz Center for Information Security, Germany
Advisor: Prof. Dr. Mario Fritz

M.Sc. in Computer Science *2017 - 2019*

Saarland University, Germany
GPA: 1.2/1.0 (Thesis GPA: 1.0)
Advisor: Prof. Dr. Mario Fritz

M.Sc. in Computer and System Engineering *2013 - 2017*

Ain Shams University, Egypt
GPA: 3.7/4.0

B.Sc. in Computer and System Engineering *2008 - 2013*

Ain Shams University, Egypt

Miscellaneous

CISPA Summer School on Trustworthy AI, 2022
ELLIS Doctoral Symposium on "AI for Good", 2022

Previous Research and Industry Experience

PhD Candidate *2019 - 2024*

CISPA Helmholtz Center for Information Security, Germany

Research Assistant *2017 - 2019*

Max Planck Institute for Informatics, Germany
Advised by Prof. Dr. Andreas Bulling

Quality Assurance Engineer *2013 - 2017*

Mentor Graphics (Currently, Siemens EDA), Egypt

Teaching Experience

Teaching Assistant *2020- 2023*

Saarland University, Germany
Classes: "Opportunities and Risks of Large Language Models and Foundation Models" seminar, "Machine Learning in Cybersecurity", "High-level Computer Vision"

Tutor *2017- 2019*

Saarland University, Germany
Classes: "Image Processing and Computer Vision", "Neural Networks Implementation and Applications", "Interactive Systems"

Publications

[1] **Sahar Abdelnabi**, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. "Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation". In: *NeurIPS Datasets and Benchmarks*. 2024.

[2] Edoardo Debenedetti, Javier Rando, Daniel Paleka, Silaghi Fineas Florin, Dragos Albastroiu, Niv Cohen, Yuval Lemberg, Reshmi Ghosh, Rui Wen, Ahmed Salem, Giovanni Cherubin, Santiago Zanella-Beguelin, Robin Schmid, Victor Klemm, Takahiro Miki, Chenhao Li, Stefan Kraft, Mario Fritz, Florian Tramèr, **Sahar Abdelnabi**, and Lea Schönherr. "Dataset and Lessons Learned from the 2024 SaTML LLM Capture-the-Flag Competition". In: *NeurIPS Datasets and Benchmarks*. **Spotlight**. 2024.

[3] **Sahar Abdelnabi**, Aideen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. "Are you still on track!? Catching LLM Task Drift with Activations". In: *arXiv* (2024).

[4] Egor Zverev, **Sahar Abdelnabi**, Mario Fritz, and Christoph H Lampert. "Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?" In: *arXiv* (2024).

[5] Ivaxi Sheth, **Sahar Abdelnabi**, and Mario Fritz. "Hypothesizing Missing Causal Variables with LLMs". In: *arXiv* (2024).

[6] Sarath Sivaprasad, Pramod Kaushik, **Sahar Abdelnabi**, and Mario Fritz. "Exploring Value Biases: How LLMs Deviate Towards the Ideal". In: *arXiv* (2024).

[7] **Sahar Abdelnabi***, Kai Greshake*, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection". In: *AISec Workshop, in conjunction with CCS*. *: Equal contribution. **Oral Presentation, Best Paper Award**. 2023.

[8] **Sahar Abdelnabi** and Mario Fritz. "Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks against Fact-Verification Systems". In: *USENIX Security Symposium (USENIX Security)*. 2023.

[9] Giada Stivala, **Sahar Abdelnabi**, Andrea Mengascini, Mariano Graziano, Mario Fritz, and Giancarlo Pellegrino. "From Attachments to SEO: Click Here to Learn More about Clickbait PDFs!" In: *Proceedings of the 39th Annual Computer Security Applications Conference*. 2023.

[10] Protik Bose Pranto, Waqar Hassan Khan, **Sahar Abdelnabi**, Rebecca Weil, Mario Fritz, and Rakibul Hasan. "Poster: Understanding the Effect of Private Data in Disinformation Propagation". In: *the 19th Symposium on Usable Privacy and Security (SOUPS) Poster Session*. 2023.

[11] **Sahar Abdelnabi**, Rakibul Hasan, and Mario Fritz. "Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[12] **Sahar Abdelnabi** and Mario Fritz. "Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding". In: *IEEE Symposium on Security and Privacy (S&P)*. 2021.

[13] Ning Yu, Vladislav Skripniuk, **Sahar Abdelnabi**, and Mario Fritz. "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data". In: *International Conference on Computer Vision (ICCV)*. **Oral presentation**. 2021.

[14] **Sahar Abdelnabi** and Mario Fritz. "What's in the box: Deflecting Adversarial Attacks by Randomly Deploying Adversarially-Disjoint Models". In: *the 8th ACM Workshop on Moving Target Defense, in conjunction with CCS*. 2021.

[15] **Sahar Abdelnabi**, Katharina Krombholz, and Mario Fritz. "Visualphishnet: Zero-day phishing website detection by visual similarity". In: *ACM Conference on Computer and Communications Security (CCS)*. 2020.

[16] **Sahar Abdelnabi**, Michael Xuelin Huang, and Andreas Bulling. "Towards High-Frequency SSVEP-Based Target Discrimination with an Extended Alphanumeric Keyboard". In: *IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019.

[17] **Sahar Abdelnabi**, Seif Eldawlatly, and Mahmoud I Khalil. "Epileptic seizure prediction using zero-crossings analysis of EEG wavelet detail coefficients". In: *IEEE conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2016.

Academic Activities

Competition Organizations

One of the main organizers of IEEE SaTML'25 challenge "LLMail-Inject: Adaptive Prompt Injection Attacks"

Helped organize the IEEE SaTML'24 challenge "LLM CtF"

Reviewing

PC member of IEEE SaTML'24, CCS'23 and '24 AISec Workshop, USENIX Security'25, AAAI'25, CCS'25

Reviewer for ICLR'25, ICML'24, ICLR'24, NeurIPS'23, ICML'23 Neural Conversational AI Workshop, ICCV'23, CVPR'23, ECCV'22, CVPR'22, IEEE TPAMI (2024, 2022, 2021), ICLR'21 Workshop on "Synthetic Data Generation – Quality, Privacy, Bias"

Grant reviewing for Cooperative AI

Talks

On New Security and Safety Challenges Posed by LLMs and How to Evaluate Them
Keynote at HIDA PhD Meet-up and MLSec seminars, 2024

On Evaluating Language Models and Their Security and Safety Implications
Vector Instuite and ETH Zürich, 2023

LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games
SIGSEC talk, 2023

Compromising LLMs: The Advent of AI Malware
Black Hat USA, 2023

Panel: Security of Generative AI and Generative AI in Security
Invited panelists at DIMVA, 2023

Not what you've signed up for: Investigating the Security of LLM-Integrated Applications
Privacy and Security in ML Seminars, 2023

How to Improve Automated Fact-Checking
Max Planck Institute for Software Systems, 2022

Multi-modal Fact-checking: Out-of-Context Images and How to Catch Them
UCL Information Security seminars, 2022

Awards

Best paper award and oral presentation at AISec, 2023

Academic Research Grant for Google Cloud research credits, 2023

Academic Research Grant for Google Cloud research credits, 2021

Oral presentation at ICCV, 2021

Saarland University scholarship for international students (DAAD STIBET III scholarship grant), 2019

IEEE Computational Intelligence Society (CIS) Outstanding Student-Paper Travel Grant for CIBCB, 2016

Media and Press Coverage

Podcasts and documentaries

ChatGPT: What happens when the AI takes over?
Y-Kollektiv Documentary, 2023

A dark side to LLMs
CyberWire Podcast, 2023

Deepfakes and Fingerprinting
CISPA tl;dr Podcast, 2022

Articles

Our work on "indirect prompt injection" has been featured in Vice, Wired, Zeit, MIT Technology Review, and others.

Languages

English (Bilingual)
German (Intermediate)
Colloquial Egyptian (Mother Tongue)

Referees

Mario Fritz
Faculty, CISPA Helmholtz Center for Information Security. Professor, Saarland University
fritz@cispa.de