Sahar Abdelnabi

sahar.abdelnabi@cispa.de @AbdelnabiSahar
sahar-abdelnabi Google Scholar Personal Page

Research Interests

I am interested in the broad intersection of machine learning with security, safety, and sociopolitical aspects. This includes the following areas: 1) Understanding 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 challenges posed by new foundation and large language models.

Education

2019 – Now **Ph.D. Computer Science**
CISPA Helmholtz Center for Information Security and Saarland University, Germany.
Advisor: Prof. Dr. Mario Fritz.

2017 – 2019 **M.Sc. Computer Science**
Saarland University, Germany.
GPA: 1.2/1.0 (thesis GPA: 1.0).
Thesis Advisor: Prof. Dr. Mario Fritz.

2013 – 2017 **M.Sc. Computer and Systems Engineering**
Ain Shams University, Egypt.
GPA: 3.7/4.0.

2008 – 2013 **B.Sc. Computer and Systems Engineering**
Ain Shams University, Egypt.

Research Experience

2019 – Now **PhD Candidate**
CISPA Helmholtz Center for Information Security, Germany.

2017 – 2019 **Student Research Assistant**
Max Planck Institute for Informatics, Germany.
I worked in the "Perceptual User Interfaces" group with Prof. Dr. Andreas Bulling in the domain of Brain-Computer Interfaces.

Industry Experience

2013 – 2017 **Quality Assurance Engineer**
Mentor Graphics (currently, Siemens EDA), Egypt.
I worked in the quality assurance team of ModelSim (multi-language HDL simulation environment). I was working on developing test plans for each product release, build the test environment, and analyze customers' problems.

Teaching Experience

2023 **Teaching Assistant**
Seminar: "Opportunities and Risks of Large Language Models and Foundation Models".
Saarland University.

2021 – 2022 **Teaching Assistant**
Course: "Machine Learning in Cybersecurity".
Saarland University.

2020 **Teaching Assistant**
Course: "High-level Computer Vision".
Saarland University.

2017 – 2019 **Tutor**
Tutored several courses during my M.Sc. studies ("Image Processing and Computer Vision", "Neural Networks Implementation and Applications", and "Interactive Systems').
Saarland University.

Publications

1. K. Greshake*, **S. Abdelnabi***, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection," in *arXiv*, *: Equal contribution, 2023.
2. **S. Abdelnabi** and M. Fritz, "Fact-saboteurs: A taxonomy of evidence manipulation attacks against fact-verification systems," in *USENIX Security Symposium (USENIX Security)*, 2023.
3. **S. Abdelnabi**, R. Hasan, and M. 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.
4. **S. Abdelnabi** and M. Fritz, "Adversarial watermarking transformer: Towards tracing text provenance with data hiding," in *IEEE Symposium on Security and Privacy (S&P)*, 2021.
5. **S. Abdelnabi** and M. 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.
6. N. Yu, V. Skripniuk, **S. Abdelnabi**, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," in *International Conference on Computer Vision (ICCV)*, 2021.
7. **S. Abdelnabi**, K. Krombholz, and M. Fritz, "Visualphishnet: Zero-day phishing website detection by visual similarity," in *ACM Conference on Computer and Communications Security (CCS)*, 2020.
8. **S. Abdelnabi**, M. X. Huang, and A. Bulling, "Towards high-frequency ssvep-based target discrimination with an extended alphanumeric keyboard," in *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019.
9. **S. Abdelnabi**, S. Eldawlatly, and M. 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.

Awards

2023 Academic Research Grant for Google Cloud research credits.

2021 Academic Research Grant for Google Cloud research credits.

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

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

Academic Activities

Reviewing NeurIPS'23, ICCV'23, CVPR'23, ECCV'22, CVPR'22, IEEE TPAMI (2022, 2021), ICLR'21 Workshop on "Synthetic Data Generation – Quality, Privacy, Bias".

Miscellaneous CISPA Summer School on Trustworthy AI, ELLIS Doctoral Symposium'22 on "AI for Good".

Talks and Media Coverage

Invited Talks

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

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

Compromising LLMs: The Advent of AI Malware
at Black Hat USA, to happen in August 2023.

Conference Presentations

CVPR'22: Multi-modal Fact-checking - [Video Recording].

S&P'21: Adversarial Watermarking Transformer - [Video Recording].

Podcasts and Documentaries

ChatGPT: What happens when the AI takes over? - Documentary: [Link]

A dark side to LLMs - Podcast: [Link]

Deepfakes and Fingerprinting - Podcast: [Link]

Articles

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

Additional Skills – Languages

English	Fluent
German	Beginner/Intermediate
Colloquial Egyptian	Mother Tongue

Referees

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