Welcome to **AISec 2026!**



Sahar Abdelnabi, Giovanni Apruzzese, Matthew Jagielski

Program Outline

Morning sessions: AI for Security and Privacy

- 09:00–09:10 Opening and Welcome of the Chairs
- 09:10–10:00 "Autonomous and Secure AI Agents through Information-flow Control" (B. Köpf)
- 10:00–10:30 Coffee Break
- 10:30–11:15 Spotlight Talks (Four talks, 10 minutes each)
- 11:15–11:30 Reviewing Awards
- 11:30–12:30 Poster Session (15 papers)
- 12:30–13:30 Lunch

Afternoon sessions: Security and Privacy of AI

- 13:30–14:20 "The AI Trust Gap: When Expectations Meet Adversarial Reality" (A. Dmitrienko)
- 14:20–15:00 Spotlight Talks (Four talks, 10 minutes each)
- 15:00–15:30 Coffee Break
- 15:30–16:45 Poster Session (24 papers)
- 16:45–17:00 Paper Awards and Closure

Switches! (THANK YOU)

Due to recent disruptions, some posters / talks have been moved:

- **[Poster]** "*Exposure is All You Need: Using LLMs to Identify Critical Endpoints at Internet Scale*", presented by Stijn Pletinckx, will be presented in the Afternoon session (instead of in the Morning)
  - In contrast, "*Semi-Automated Detection of Gaps in LLM Security Knowledge*" presented by Jessica Staddon, will be presented in the Morning Session
- **[Poster and Spotlight talk]** "*Hacking Back the AI Hacker: Prompt Injection as a Defense Against LLM-driven Cyberattacks*", presented by Evgenios Kornaropoulos, will be presented in the Afternoon
  - In contrast, "*Trojan Hippo Bench: A Dynamic Benchmark for Persistent Memory Attacks and Defenses in LLM Agents Plugins*", presented by Debeshee Das, will be presented in the Morning
- **[Poster]** "*Ghost Approvals: Approval-Provenance Failure in Agentic AI Frameworks*" will be presented by Mina Mehrvaz (instead of Yuvraj Pradhan) in the Morning (instead of in the Afternoon)
  - Mina Mehrvaz will also present her poster on "*Compositional Dual-Trigger Backdoors for Independent Prediction and Explanation Manipulation*" (in the Afternoon)

Peer Review Process

Overview

- 263 total submissions
  - Up from 72 total submissions for both AISec ‘25 and AISec ‘24
- 221 eligible for peer review
  - 42 desk rejections due to: hallucinations (9), anonymity (6), formatting (17), quality (6), duplicate (4)
- 39 accepted papers

- 475 reviews have been submitted
  - (21 reviews were for papers that, unfortunately, have been administratively desk rejected)
- 257,002 words have been written across all reviews

Selection Process

Initially, we composed a Program Committee (PC) that would have enabled reviewing 120 submissions (with 3 reviews each).

The unforeseeable number of submissions forced us to change our plans

- We recruited ~25 additional PC members within ~1 week
- We extended the notification deadline by 10 days
- We changed the reviewing format:
  - Every paper would have 2 reviews,
  - PC members would be expected to engage in a discussion to reach a consensus on the decision
  - In cases where a decision could not be made, a third “tiebreaker” reviewer would be added to the pool
- The Workshop Chairs, who also interacted with PC members (if necessary), made the final call

Tiebreaker and Discussion

- The ensuing discussions were constructive and insightful
  - Some exceeded 2,500 words in length for a single paper.
  - 1,350 comments have been posted on HotCRP, spanning 88,574 words overall

- Of the 39 accepted papers, 19 have been accepted after the discussion led to identifying well-defined easy-to-implement changes (typically, rewordings or minor clarifications) that would have made the paper more accurate w.r.t. its contributions
  - The authors were strongly encouraged to integrate these changes while preparing the camera ready

- 13 papers received a third “tiebreaker” review:
  - Four of these papers were ultimately accepted.

Reviewing Form

Overall merit *

- A. Good paper, I will champion it
- B. OK paper, but I will not champion it
- C. Weak paper, though I will not fight strongly against it
- D. Reject

Usage of Generative AI *

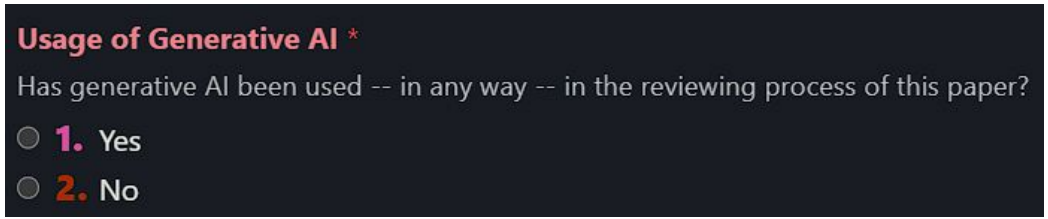
Has generative AI been used -- in any way -- in the reviewing process of this paper?

- 1. Yes
- 2. No

Usage of Generative AI

Out of 454 reviews submitted across the 221 papers:

- 156 had a “Yes”
- 298 had a “No”



Across 142 PC members that contributed to reviewing these 221 papers:

- 72 declared “Yes” at least once
  - Of these, 34 always picked “Yes”
- Whereas 70 never picked “Yes”.

Keynote



Title: *Autonomous and Secure AI Agents through Information-flow Control*

▼ **Boris Köpf, Microsoft Azure Research**

Short Bio: Boris Köpf is a scientist at Microsoft's Azure Research, where he works on security of AI agents (Project Roma) and on defenses against side-channel attacks (Project Venice).

His research brings together security, formal verification, programming languages, systems, and AI. He has published at venues including IEEE S&P, USENIX Security, ACM CCS, PLDI, CAV, ASPLOS, ICML, and ICLR. His work has received a Best Paper Award at IEEE S&P (2021) and Distinguished Paper Awards at USENIX Security (2017), ACM CCS (2024), and IEEE S&P (2026).

Spotlight Papers

Selection Process

- We divided all 39 accepted papers in two broad thematic groups:
  - AI for Security/Privacy (with 15 papers)
  - Security/Privacy of AI (with 24 papers)
- We shortlisted the top 60% of the papers in each group (using the submitted reviews), and asked the entire PC to express their preference on up to 3 papers for each group about the papers they deemed to be “particularly noteworthy”.
- 44 PC members voted: we picked the 4 papers of each group with the most votes
- Honorable mentions for “spotlight” papers (came 5th for each group):

Selection Process

- We divided all 39 accepted papers in two broad thematic groups:
  - AI for Security/Privacy (with 15 papers)
  - Security/Privacy of AI (with 24 papers)
- We shortlisted the top 60% of the papers in each group (using the submitted reviews), and asked the entire PC to express their preference on up to 3 papers for each group about the papers they deemed to be “particularly noteworthy”.
- 44 PC members voted: we picked the 4 papers of each group with the most votes
- Honorable mentions for “spotlight” papers (came 5th for each group):
  - **PkgScanArena: Benchmarking LLM-Based Package Scanners against Indirect Prompt Injection**, by Himadri Chowdhury, Ahmed Hatem, and Joseph Khoury (Louisiana State University)
  - **Position Matters: Feature Inversion Attacks in ViT Split Inference with Token Reduction and Shuffling**, by Stefano Leggio, Giulio Rossolini, Alessandro Biondi (Scuola Superiore Sant'Anna)

Spotlight Papers: *AI for Security/Privacy*

Research paper **Beyond the Syntax: Do Security Experts Trust LLMs for NIDS Rule Engineering?** ★ Spotlight

Lorenzo Di Filippo [1], Enkeleda Bardhi [2], Andrea Agiollo [2], Alessandro Palma [3], Silvia Bonomi [3], Fernando Kuipers [2]

[1] Sorbonne Université, LIP6, Paris, France [2] Delft University of Technology [3] Sapienza University of Rome

---

Benchmark paper **Trojan Hippo Bench: A Dynamic Benchmark for Persistent Memory Attacks and Defenses in LLM Agents Plugins**

★ Spotlight

Debeshee Das [1], Julien Piet [2], Darya Kaviani [2], Luca Beurer-Kellner [3], Florian Tramèr [1], David Wagner [2]

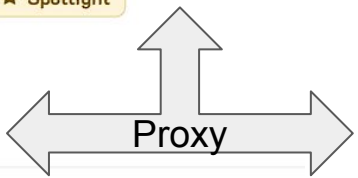
[1] ETH Zurich [2] University of California, Berkeley [3] Snyk



Benchmark paper **ScopeBench: Do Agents Preserve Engagement Boundaries Under Goal Pressure?** ★ Spotlight

Shane Caldwell, Max Harley, Ads Dawson, Michael Kouremetis, Vincent Abruzzo, Will Pearce

Dreadnode



Research paper **TypeSeg: Content-Type Segmentation for Mixed Text** ★ Spotlight

Martin Dallinger [1,2], Yanick Fratantonio [1], Luca Invernizzi [1]

[1] Google [2] Johannes Kepler University Linz

PC Members Recognitions

Motivation

- Few incentives to carry out “proper” service
- The few acknowledged incentives (i.e., “reviewing awards”) are opaque
  - What is really needed to achieve such awards?
- **We want to change this**

- There is much more that PC members can do than just “providing good reviews”
- **We want to promote positive behaviors from the entire community**
- We devised a set of “PC member recognitions” meant to highlight particularly positive behaviors that substantially contributed to the realization of AISec ‘26

Speedy 

- **Description:** awarded to the top 5% PC members who were the first to submit two reviews (with no late reviews)

- **Rationale:** submitting reviews early on is helpful for discussion—and to prevent “panicking” of PC chairs :)

- **Awardees (7):** Hannan Mahadik, Nik Kale, Md Sultanul Islam Ovi, Luca Pajola, Xiaokang Yin, Yuwen Pu, Maura Pintor

Commenter



- **Description:** awarded to the top 5% PC members with the highest word count across all comments or with the highest number of submitted comments.

- **Rationale:** interacting and engaging in discussion leads to better decisions

- **Awardees (9):** Kerri Prinos, Zeya Umayya, Advije Rizvani, Alberto Castagnaro, Jose Sanchez Vicarte, Sina Mavali, Bhavna Soman, Stefanos Koffas, Giuseppe Parrella

Elaborate



- **Description:** Awarded to the top 5% PC members with the highest cumulative word count across all reviews (with no late reviews)

- **Rationale:** lengthy reviews (typically) require more effort and (typically) provide more feedback

- **Awardees (7):** Stefanos Koffas, Hannan Mahadik, Jean-Charles Noirot Ferrand, Matteo Gioele Collu, Zeya Umayya, Shaoor Munir, William Corrias

Committed 

- **Description:** Awarded to those PC members who placed ≥10 bids, set topic preferences for all topics, were not late, and submitted ≥2 reviews

- **Rationale:** the above leads to better happiness (and better reviews), and prevents chairs from having to chase reviewers with emails

- **Awardees (16):** Adam Jones, Advije Rizvani, Andreas Happe, Kerri Prinos, Boyang Zhang, Marco Di Gennaro, Dimitri Galli, Bhavna Soman, Matteo Gioele Collu, Mabon Manoj Ninan, Konrad Rieck, Juhee Kim, Lea Schönherr, Kristina Nikolic, Kunyang Li, Lorenzo Cavallaro

Alacritous 

- **Description:** Awarded to those PC members who initiated the discussion for at least two distinct papers without being prompted to do so by the PC chair

- **Rationale:** proactivity is a sign of commitment, leads to better discussions, and saves time to organizers

- **Awardees (10):** Matteo Gioele Collu, Zeya Umayya, Balachandra Shanabhag, Ethan Wang, Gaurav Jain, Maura Pintor, Lea Schönherr, Sina Mavali, Juhee Kim, Xin Fan Guo

Firefighter 

- **Description:** Awarded to those PC members who submitted ≥2 reviews as well as an extra (emergency or tiebreaker) review

- **Rationale:** being able to submit reviews in a short time frame is extremely valuable (we simply could not have made it without these extra efforts)

- **Awardees (22):** Adam Jones, Andrew Paverd, Dimitri Galli, Gustavo Sánchez, Jacqueline Clare Mallett, Luca Pajola, Marco Di Gennaro, Maura Pintor, Patrick Dwyer, Ying Yuan, Matteo Gioele Collu, Zeyuan Chen, Daryna Oliynyk, Giulio Rossolini, Hari Venugopalan, Kerri Prinos, Marcello Maugeri, Carlo Mazzocca, Alessandro Brighente, William Corrias, Zeya Umayya, Kyle Domico

Appreciated 

- **Description:** Awarded to the top 5% PC members with the highest upvotes received on HotCRP.

- **Rationale:** upvotes (when they are released by fellow PC members) typically indicate high-quality reviews

- **Awardees (8):** Advije Rizvani, Jean-Charles Noirot Ferrand, Stefanos Koffas, Daryna Oliynyk, Hannan Mahadik, Ying Yuan, Luca Pajola, Gustavo Sanchez

Remark!

For determining “late” reviews: PC members who *proactively* informed us of anticipated delays and subsequently met their self-imposed deadlines **were not considered late**

Distinguished PC Member Awards



- **Description:** awarded to those PC members that met >=8 of the following 12 criteria:
  - submitting ≥ 10 bids;
  - setting topic preferences for paper-assignment across all topics;
  - submitting all reviews on time;
  - being ranked in the top 30% of PC members by total number of comments posted;
  - being ranked in the top 30% of PC members by aggregated length of comments;
  - having submitted ≥3 reviews;
  - being in the top 30% of PC members by average review length;
  - being in the top 30% by total upvotes received;
  - having initiated ≥1 discussion on an assigned paper;
  - having submitted one emergency or tiebreaker review;
  - having submitted all reviews at least 48 hours before the planned deadline (excluding extra reviews);
  - having voted in the spotlight paper selection.

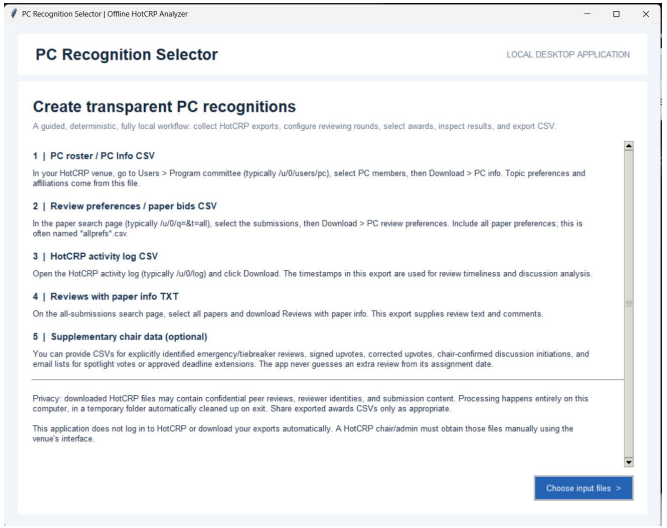
- **Awardees (24):** Advije Rizvani, Andrew Paverd, Bhavna Soman, Daryna Oliynyk, David Pape, Dimitri Galli, Giorgia Di Pietro, Giuseppe Parrella, Hannan Mahadik, Jean-Charles Noirot Ferrand, Jonathan Evertz, Jose Sanchez Vicarte, Kerri Prinos, Kristina Nikolić, Lea Schönherr, Lorenzo Cavallaro, Luca Pajola, Marco Di Gennaro, Matteo Gioele Collu, Maura Pintor, Stefanos Koffas, William Corrias, Zeya Umayya, Zied Ben Houidi

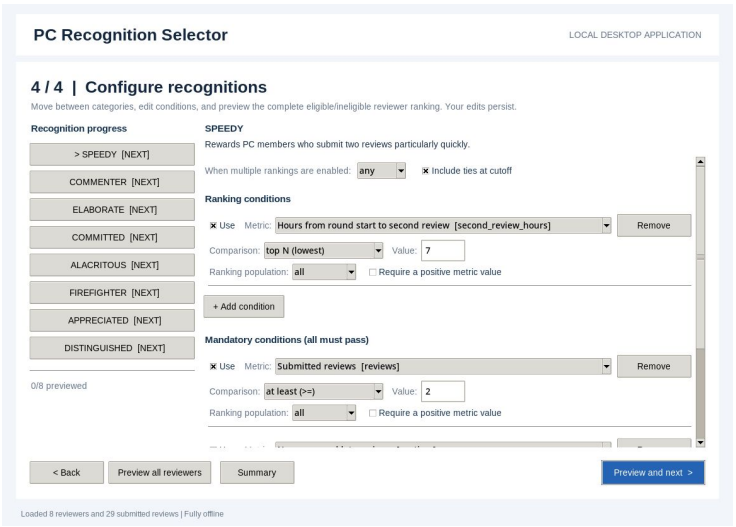
Outlook

- Overall, 53 PC members (out of 142) were recognized with at least one “distinction”, and a total of 103 “distinction badges” have been assigned.
- Our distinctions are: transparent, measurable, deterministic, and reusable

Outlook

- Overall, 53 PC members (out of 142) were recognized with at least one “distinction”, and a total of 103 “distinction badges” have been assigned.
- Our distinctions are: transparent, measurable, deterministic, and reusable
- To foster future adoption, we openly release the code, including a GUI-based APP, used to compute these recognitions: https://github.com/hihey54/distinguished_reviewer_selector





Poster Session
(AI for Security/Privacy)

Lunch begins at 12:30

Lunch

Submissions Statistics

Submitted Ratings (across 454 reviews)

Topic	Submitted	Share
A — Good paper, will champion it	11	2.4%
B — OK paper, will not champion it	94	20.7%
C — Weak paper, but will not strongly oppose it	182	40.1%
D — Reject	167	36.8%

Acceptance Rate by “Submission Topic”

Topic	Considered	Accepted	AR (%)
AI Security	145	24	16.6
Security Applications of AI	101	13	12.9
AI Privacy	35	6	17.1
Other (e.g., Federated Learning, Machine Unlearning)	29	2	6.9
Offensive AI	25	3	12.0
Human Factors in Security/Privacy aspects of AI	18	2	11.1

Acceptance Rate per Paper Category

Type	Considered	Accepted	AR (%)
Research Papers	161	31	19%
SoK	5	0	0%
Position	30	3	10%
Benchmark	25	5	20%

Acceptance Rate per Number of Paper Authors

Authors	Considered	Accepted	AR (%)
1	58	2	3.4
2	48	6	12.5
3	42	8	19.0
4	25	5	20.0
5	26	8	30.8
6	11	6	54.5
7	8	4	50.0
8	2	0	0.0
10	1	0	0.0

Acceptance Rate by Institutions' Country

Country / region	Considered	Accepted	AR (%)
USA	60	13	21.7%
Germany	24	10	41.7%
Canada	15	2	13.3%
India	13	1	7.7%
United Kingdom	11	2	18.2%
Italy	10	4	40.0%
South Korea	9	1	11.1%
Netherlands	7	2	28.6%
Japan	6	0	0.0%
Turkey	5	0	0.0%
Belgium	4	1	25.0%
China	4	0	0.0%
Greece	4	1	25.0%
Luxembourg	4	1	25.0%
Sweden	4	1	25.0%
Switzerland	4	2	50.0%
Taiwan	4	0	0.0%
Australia	3	1	33.3%
France	3	2	66.7%
Hong Kong	3	0	0.0%
Norway	3	1	33.3%
Poland	3	0	0.0%
Austria	2	1	50.0%
Denmark	2	1	50.0%

Country / region	Considered	Accepted	AR (%)
Estonia	2	1	50.0%
Finland	2	1	50.0%
Portugal	2	0	0.0%
Spain	2	0	0.0%
United Arab Emirates	2	0	0.0%
Argentina	1	0	0.0%
Bangladesh	1	0	0.0%
Brazil	1	0	0.0%
Bulgaria	1	0	0.0%
Croatia	1	1	100.0%
Cyprus	1	0	0.0%
Czechia	1	0	0.0%
Georgia	1	0	0.0%
Hungary	1	0	0.0%
Indonesia	1	0	0.0%
Ireland	1	1	100.0%
Israel	1	0	0.0%
Malta	1	0	0.0%
Romania	1	0	0.0%
Singapore	1	0	0.0%
Sri Lanka	1	0	0.0%
Unaffiliated	35	2	5.7%
Unknown affiliation	2	0	0.0%

Keynote



Title: *The AI Trust Gap: When Expectations Meet Adversarial Reality*

▼ Alexandra Dmitrienko, **University of Duisburg-Essen**

Short Bio: Alexandra Dmitrienko is a Full Professor of Computer Science at the University of Duisburg-Essen, Germany, where she holds the Chair of Trustworthy Systems. Previously, she held academic and research positions at the University of Würzburg, ETH Zurich, the Fraunhofer Institute for Secure Information Technology, and Ruhr University Bochum. She received her PhD in Information Technology from TU Darmstadt with distinction.

Her research interests span software and systems security, IoT security, and the intersection of artificial intelligence and cybersecurity. Her research regularly appears at the Big Four security conferences: ACM CCS, USENIX Security, NDSS, and the IEEE Symposium on Security and Privacy. She has delivered invited keynotes at AsiaCCS 2024 and ACNS 2025.

She is actively involved in the international security research community and has held several conference leadership positions, including Track Chair for ACM CCS 2025, TPC Vice Chair for USENIX Security 2024, and TPC Chair for ACM WiSec 2022. She regularly serves on the program committees of all Big Four security conferences and is serving as an Associate Chair for NDSS 2027 and Technical Program Co-Chair for IEEE CNS 2027.

Her research and academic contributions have been recognized with several awards, including the IEEE Symposium on Security and Privacy Best Student Paper Award (2013), the Intel Doctoral Student Honor Award (2013), the ERCIM STM Working Group Award (2016), the ACM AsiaCCS Distinguished Paper Award (2017), and the NDSS Distinguished Reviewer Award (2026).

Spotlight Papers

Spotlight Papers: *Security/Privacy of AI*

Research paper **Controlling Autonomous Vehicles using Continuous Adversarial Patches** ★ Spotlight

Pia Hanfeld [1,2], Erik Imgrund [1,2], Felix Weissberg [1,2], Thorsten Eisenhofer [3], Wolfgang Hönig [2], Konrad Rieck [1,2]

[1] BIFOLD [2] TU Berlin [3] CISPA Helmholtz Center for Information Security

Research paper **False Prophets: On the Security of World Models in Agentic Systems** ★ Spotlight

Erik Imgrund [1,2], Anna Wimbauer [1,2], Klim Kireev [1,2], Konrad Rieck [1,2]

[1] BIFOLD [2] TU Berlin

Research paper **PromptLeaks: Assessing Data Leakage in Guardrails and Automated Jailbreak Tools** ★ Spotlight

Kobe Sauwens, Simon Knuts, Tim Van Hamme, Lieven Desmet

DistriNet, KU Leuven

Research paper **Hacking Back the AI Hacker: Prompt Injection as a Defense Against LLM-driven Cyberattacks** ★ Spotlight

Dario Pasquini [1], Evgenios M. Kornaropoulos [2], Giuseppe Ateniese [2]

[1] Cracken [2] George Mason University



Distinguished Paper Awards

Process

- The 3 PC Chairs and the 4 members of the Steering Committee expressed, individually (and privately) their preference for the most noteworthy paper across the four spotlight papers of each group
  - In case of ties, the winner would have been the paper for which most PC members voted for

- Runners-up for “distinguished” papers (came second for each group):

Process

- The 3 PC Chairs and the 4 members of the Steering Committee expressed, individually (and privately) their preference for the most noteworthy paper across the four spotlight papers of each group
  - In case of ties, the winner would have been the paper for which most PC members voted for

- Runners-up for “distinguished” papers (came second for each group):
  - **ScopeBench: Do Agents Preserve Engagement Boundaries Under Goal Pressure?** by Shane Caldwell, Max Harley, Ads Dawson, Michael Kouremetis, Vincent Abruzzo, Will Pearce (Dreadnode)

Process

- The 3 PC Chairs and the 4 members of the Steering Committee expressed, individually (and privately) their preference for the most noteworthy paper across the four spotlight papers of each group
  - In case of ties, the winner would have been the paper for which most PC members voted for

- Runners-up for “distinguished” papers (came second for each group):
  - **ScopeBench: Do Agents Preserve Engagement Boundaries Under Goal Pressure?** by Shane Caldwell, Max Harley, Ads Dawson, Michael Kouremetis, Vincent Abruzzo, Will Pearce (Dreadnode)
  - **Trojan Hippo Bench: A Dynamic Benchmark for Persistent Memory Attacks and Defenses in LLM Agents** by Debeshee Das, Julien Piet, Darya Kaviani, Luca Beurer-Kellner, Florian Tramèr, David Wagner (ETH Zurich, UC Berkeley, Snyk)



The 19th ACM Workshop on Artificial Intelligence and Security (AISec '26)

presents to

Kobe Sauwens, Simon Knuts, Tim Van Hamme and Lieven Desmet

DistriNet, KU Leuven

the **Distinguished Paper Award** for the paper titled

"PromptLeaks: Assessing Data Leakage in Guardrails and Automated Jailbreak Tools"

15th November 2026, The Hague, Netherlands

Sahar Abdelnabi Workshop Chair	Giovanni Apruzzese Workshop Chair	Matthew Jagielski Workshop Chair

19th ACM Workshop on
Artificial Intelligence and Security (AISec '26)

presents to

Dario Pasquini[1], Evgenios M. Kornaropoulos[2] and Giuseppe Ateniese[2]

[1] Cracken [2] George Mason University

the **Distinguished Paper Award** for the paper titled

"Hacking Back the AI Hacker: Prompt Injection as a Defense Against LLM-driven Cyberattacks"

15th November 2026, The Hague, Netherlands

Sahar Abdelnabi
Workshop Chair

Giovanni Apruzzese
Workshop Chair

Matthew Jagielski
Workshop Chair

Thank you!

AISec 2026

See you next year!



Sahar Abdelnabi, Giovanni Apruzzese, Matthew Jagielski