Andrea Agiollo

Postdoctoral Researcher in Trustworthy AI Systems

Delft, The Netherlands andagio.github.io A.Agiollo-1@tudelft.nl
scholar.google.com 0000-0003-0531-1978 github.com

Profile

Research. I am a postdoctoral researcher at TU Delft working on trustworthy AI. My research revolves around understanding and improving the privacy, robustness, and reliability of machine learning and Neuro-Symbolic (NeSy) AI systems. I study what models memorize and leak, from loss geometry-based memorization estimates for dataset pruning to the membership inference risks of sharpness-aware training, and stress test NeSy models against adversarial threats such as backdoor attacks. I bring these questions to AI systems that learn from distributed and federated data, such as energy-aware serverless federated learning. Cybersecurity serves as the main application domain and testbed, for example, grounding intrusion detection in structured threat knowledge such as attack graphs. I regularly publish at top-tier venues including ICLR, KDD, IEEE EuroS&P, and ACM Computing Surveys, with a spotlight paper at ACM AISec 2026 (top 8 of 263 submissions).

Education. Supervised six MSc theses (two ongoing), of which two graduated cum laude and two led to first-author student papers currently under submission, and two groups of BSc research projects. Following University Teaching Qualification (UTQ) courses at TU Delft.

Leadership & service. Committee member of the TU Delft PostDoc Network, giving postdocs a voice in university decision making. Web Chair of ECAI 2025 as well as organizer of four international workshops. PC member for top-tier venues such as NeurIPS, ICML, ICLR, KDD, and AAAI, completing 40+ reviews per year and recognized with four reviewer awards.

Areas of expertise: Trustworthy AI - Neuro-Symbolic AI - Distributed Learning - Resource-Efficient AI - Cybersecurity

Academic Positions

PostDoc at Multimedia Computing Group at EEMCS Faculty *TU Delft* **Delft, NL** *10/2026 - Ongoing*
Research on trustworthy and privacy-preserving graph learning, building on my work on memorization and loss sharpness of machine learning models. Currently, working under the supervision of Dr. Megha Khosla.

PostDoc at Cybersecurity Group at EEMCS Faculty *TU Delft* **Delft, NL** *10/2024 - 09/2026*
I developed a research line on trustworthy learning systems for cybersecurity applications and data, where I evaluated the privacy and robustness of neuro-symbolic and deep models against membership inference and backdoor attacks, and designed neuro-symbolic intrusion detection pipelines that ground learned detectors in structured threat knowledge such as attack graphs. I supervised three MSc theses in this line, two of which led to first author student papers currently under submission.

Visiting Research Scholar at Purdue University **West Lafayette, Indiana, USA** *08/2023 - 11/2023*
My research efforts focused on the intersection of neuro-symbolic integration techniques and the properties of neural network learning process, focusing on memorization, minimum flatness and many others. Throughout this experience, I had the pleasure of collaborating with an international research group, strengthening my knowledge of the inner properties of sub-symbolic models.

Visiting Research Scholar at Delft University of Technology **Delft, NL** *10/2022 - 02/2023*
My research efforts focused on eXplainable Artificial Intelligence (XAI) techniques to extract knowledge from a trained Large Language model used for text classification and morality alignment. Throughout this experience, I had the pleasure to collaborate with an international research group, strengthening my knowledge of XAI and Natural Language Processing.

Education

Ph.D. in Computer Science and Engineering *University of Bologna* **Bologna, IT** *11/2020 - 06/2024*
Thesis entitled "*Embedding AI into Constrained Devices: A Multi-Faceted Approach*"

MSc in ICT for Internet and Multimedia *University of Padua* **Padua, IT** *09/2018 - 07/2020*
Graduation Grade: 110/110 cum laude

MSc in Communication Engineering *National Taiwan University* **Taipei, TWN** *09/2018 - 07/2020*
Graduation Grade: 4.23/4.3

BSc in Information Engineering *University of Padua* **Padua, IT** *10/2015 - 07/2018*
Graduation Grade: 110/110 cum laude

Erasmus Exchange *University of Aberdeen* **Aberdeen, Scotland, UK** *09/2017 - 01/2018*

Awards and Honors

- **Spotlight Paper** (top 8 of 263 submissions) *19th ACM Workshop on Artificial Intelligence and Security (AISec-26)*
- **Gold Reviewer** *43rd International Conference on Machine Learning (ICML-26)*
- **Top Reviewer** *42nd International Conference on Machine Learning (ICML-25)*
- **Outstanding Reviewer** *31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-25)*
- **Top Reviewer** *39th Conference on Uncertainty in Artificial Intelligence (UAI-23)*
- **Student Award** *Double Degree Honours, Master's Degree Programme at the University of Padua (2020)*

Publications

Bibliometrics: *520 citations, h-index: 12, i10-index: 14 (Google Scholar, November 2026).*

Journal Articles:

[1] Giovanni Ciatto, **Andrea Agiollo**, Matteo Magnini, and Andrea Omicini. "Large language models as oracles for instantiating ontologies with domain-specific knowledge". In: *Knowledge-Based Systems* 310 (2025), p. 112940

[2] **Andrea Agiollo**, Luciano Cavalcante Siebert, Pradeep Murukannaiah, and Andrea Omicini. "From Large Language Models to Small Logic Programs: Building Global Explanations from Disagreeing Local Post-hoc Explainers". In: *Autonomous Agents and Multi-Agent Systems* 38.2, 32 (2024). DOI: 10.1007/s10458-024-09663-8

[3] Giovanni Ciatto, Federico Sabbatini, **Andrea Agiollo**, Matteo Magnini, and Andrea Omicini. "Symbolic Knowledge Extraction and Injection with Sub-symbolic Predictors: A Systematic Literature Review". In: *ACM Computing Surveys* 56.6 (June 2024), 161:1–161:35. DOI: 10.1145/3645103

[4] **Andrea Agiollo**, Paolo Bellavista, Matteo Mendula, and Andrea Omicini. "EneA-FL: Energy-aware Orchestration for Serverless Federated Learning". In: *Future Generation Computer Systems* 154 (May 2024), pp. 219–234. DOI: 10.1016/j.future.2024.01.007

[5] **Andrea Agiollo**, Enkeleda Bardhi, Mauro Conti, Nicolò Dal Fabbro, and Riccardo Lazzeretti. "Anonymous Federated Learning via Named-Data Networking". In: *Future Generation Computer Systems* 152 (2024), pp. 288–303. DOI: 10.1016/j.future.2023.11.009

[6] **Andrea Agiollo**, Andrea Rafanelli, Matteo Magnini, Giovanni Ciatto, and Andrea Omicini. "Symbolic Knowledge Injection meets Intelligent Agents: QoS metrics and experiments". In: *Autonomous Agents and Multi-Agent Systems* 37.2 (June 2023), 27:1–27:30. DOI: 10.1007/s10458-023-09609-6

[7] **Andrea Agiollo**, Mauro Conti, Pallavi Kaliyar, TsungNan Lin, and Luca Pajola. "DETONAR: Detection of Routing Attacks in RPL-based IoT". in: *IEEE Transactions on Network and Service Management* 18.2 (2021), pp. 1178–1190. DOI: 10.1109/TNSM.2021.3075496

[8] **Andrea Agiollo** and Andrea Omicini. "Load Classification: A Case Study for Applying Neural Networks in Hyper-Constrained Embedded Devices". In: *Applied Sciences* 11.24 (Dec. 2021). Special Issue "Artificial Intelligence and Data Engineering in Engineering Applications". DOI: 10.3390/app112411957

Conference Papers:

[9] Young In Kim, **Andrea Agiollo**, Pratiksha Agrawal, Johannes O. Royset, and Rajiv Khanna. "Membership Privacy Risks of Sharpness Aware Minimization". In: *International Conference on Learning Representations (ICLR)*. 2026. arXiv: 2310.00488. URL: https://proceedings.iclr.cc/paper_files/paper/2026/hash/82f77023563b21fbd423b2d9b165c281-Abstract-Conference.html

[10] **Andrea Agiollo**, Young In Kim, and Rajiv Khanna. "Approximating Memorization Using Loss Surface Geometry for Dataset Pruning and Summarization". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2024, pp. 17–28. DOI: 10.1145/3637528.3671985

[11] **Andrea Agiollo** and Andrea Omicini. "Measuring Trustworthiness in Neuro-Symbolic Integration". In: *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*. Vol. 35. Annals of Computer Sciences and Information Systems. Sept. 2023, pp. 1–10. DOI: 10.15439/2023F6019

[12] **Andrea Agiollo**, Enkeleda Bardhi, Mauro Conti, Riccardo Lazzeretti, Eleonora Losiouk, and Andrea Omicini. "GNN4IFA: Interest Flooding Attack Detection With Graph Neural Networks". In: *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. Delft, Netherlands, July 2023, pp. 615–630. DOI: 10.1109/EuroSP57164.2023.00043

[13] **Andrea Agiollo** and Andrea Omicini. "GNN2GNN: Graph Neural Networks to Generate Neural Networks". In: *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence, UAI 2022*. Aug. 2022, pp. 32–42. URL: https://proceedings.mlr.press/v180/agiollo22a.html

[14] Umberto Michieli, Maria Camporese, **Andrea Agiollo**, Giampaolo Pagnutti, and Pietro Zanuttigh. "Region Merging Driven by Deep Learning for RGB-D Segmentation and Labeling". In: *13th International Conference on Distributed Smart Cameras (ICDSC2019)*. 2019, 9:1–9:6. URL: https://doi.org/10.1145/3349801.3349810

Workshop Papers:

[15] Lorenzo Di Filippo, Enkeleda Bardhi, **Andrea Agiollo**, Alessandro Palma, Silvia Bonomi, and Fernando Kuipers. "Beyond the Syntax: Do Security Experts Trust LLMs for NIDS Rule Engineering?" In: *Proceedings of the 19th ACM Workshop on Artificial Intelligence and Security (AISec), co-located with ACM CCS*. 2026. DOI: 10.1145/3847352.3848108. arXiv: 2607.05916. **Spotlight paper** (top 8 out of 263 submissions)

[16] Mattia Passeri, **Andrea Agiollo**, and Andrea Omicini. "Peer-Reviewed Federated Learning". In: *WOA 2023 – 24th Workshop "From Objects to Agents"*. Ed. by Rino Falcone, Cristiano Castelfranchi, Alessandro Sapienza, and Filippo Cantucci. Vol. 3579. CEUR Workshop Proceedings. Sun SITE Central Europe, RWTH Aachen University, Nov. 2023, pp. 49–65

[17] **Andrea Agiollo**, Luciano C. Siebert, Pradeep K. Murukannaiah, and Andrea Omicini. "The Quarrel of Local Post-hoc Explainers for Moral Values Classification in Natural Language Processing". In: *Explainable and Transparent AI and Multi-Agent Systems. Fifth International Workshop, EXTRAAMAS 2023, London, UK, May 29, 2023, Revised Selected Papers*. Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2023, pp. 97–115. DOI: 10.1007/978-3-031-40878-6_6

[18] **Andrea Agiollo**, Andrea Rafanelli, and Andrea Omicini. "Towards Quality-of-Service Metrics for Symbolic Knowledge Injection". In: *WOA 2022 – 23rd Workshop "From Objects to Agents"*. Vol. 3261. CEUR Workshop Proceedings. 2022, pp. 30–47. URL: https://ceur-ws.org/Vol-3261/paper3.pdf

[19] **Andrea Agiollo**, Giovanni Ciatto, and Andrea Omicini. "Graph Neural Networks as the Copula Mundi between Logic and Machine Learning: A Roadmap". In: *WOA 2021 – 22nd Workshop "From Objects to Agents"*. Ed. by Roberta Calegari, Giovanni Ciatto, Enrico Denti, Andrea Omicini, and Giovanni Sartor. 2021, pp. 98–115. URL: http://ceur-ws.org/Vol-2963/paper18.pdf

[20] **Andrea Agiollo**, Giovanni Ciatto, and Andrea Omicini. "Shallow2Deep: Restraining Neural Networks Opacity through Neural Architecture Search". In: *Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers*. Cham, Switzerland, 2021, pp. 63–82. URL: http://link.springer.com/10.1007/978-3-030-82017-6_5

Book Chapters:

[21] **Andrea Agiollo**, Roberta Calegari, Giovanni Ciatto, Matteo Magnini, Andrea Omicini, and Federico Sabbatini. "Intelligent agents from symbolic to neurosymbolic systems: the quest for integration". In: *The Agents Journey: Twenty-Five Years of Multi-agent Systems*. Springer, 2026, pp. 320–339

Patents and Inventions:

Patent: "Dishwasher for Treating Washware" **Pordenone, IT**

- *Inventors*: Andrea Comina, **Andrea Agiollo**, and Andrea Omicini.
- *Status*: Granted in the United States as US 12,569,109 B2 (03/2026). European application EP 4427656 A1 (filed 03/2023, published 09/2024) and Chinese application CN 118614825 A pending.
- *Full access*: US12569109B2, EP4427656A1, CN118614825A.
- Done as part of my PhD program in collaboration with Electrolux Professional

Submitted: († MSc student supervised by me)

- Raoul Saad†, Enkeleda Bardhi, **Andrea Agiollo**, and Fernando Kuipers. "DeepWeave: Weaving Neural and Symbolic Evidence for ML-based Network Intrusion Detection System". Under submission
- Sofia V. Burkow†, **Andrea Agiollo**, Enkeleda Bardhi, Fernando Kuipers, Kaitai Liang, and Georgios Smaragdakis. "DeepChronos: Temporal Neuro-Symbolic Reasoning for Multi-Step Attack Detection". Under submission
- Marco Antonio Corallo, **Andrea Agiollo**, Mauro Conti, and Alberto Giaretta. "Does Reasoning Mitigate Backdoor Attacks? A Neuro-Symbolic Perspective". Under submission. arXiv: 2609.00464
- **Andrea Agiollo**, Enkeleda Bardhi, Alessandro Palma, Riccardo Lazzeretti, Silvia Bonomi, and Fernando Kuipers. "SoK: Harmonising Attack Graphs and Intrusion Detection Systems". Under submission
- **Andrea Agiollo**, Young In Kim, and Rajiv Khanna. "SAMOSA: Sharpness Aware Minimization for Open Set Active Learning". Under submission
- **Andrea Agiollo** and Andrea Omicini. "Integrating Machine Learning into Belief-Desire-Intention Agents: Current Advances and Open Challenges". Under submission

Software & Artifacts

- **EneA-FL**: energy-aware orchestration framework for serverless federated learning.
- **MIABench**: modular, scalable framework for benchmarking membership inference attacks, with distributed training and sharpness-aware optimizers (e.g., SAM, ASAM, GSAM).
- **PSyKI**: Python library for symbolic knowledge injection into neural networks, released on PyPI (second largest contributor).
- **SAMIS**: memorization estimation from loss-surface geometry for dataset pruning and summarization.
- **DETONAR** and the **RADAR** dataset: intrusion detection for routing attacks in RPL-based IoT, with RADAR, a public dataset of simulated RPL traffic covering 14 routing attacks.
- **GNN4IFA** and the **SPOTIFAI** dataset: graph-neural-network detection of Interest Flooding Attacks in Named-Data Networking, with SPOTIFAI, a public dataset of attacks simulated in ndnSIM on topologies of up to 163 nodes.

Invited Talks & Scientific Events

Invited Participation:

Invited Participant at Dagstuhl Seminar 26421 **Wadern, Germany** *10/2026*
Invitation-only seminar on "*Advancing Testability and Verifiability of CPS with Neurosymbolic and Large Language Models*", gathering international experts to explore how neuro-symbolic methods and large language models can be combined to test and formally verify learning-enabled cyber-physical systems.

Seminars & Invited Talks:

- Talk: "The Geometry of Memorization: When Better Generalization Increases Privacy Risk"; Venue: 3rd ACCSS Security & AI Workshop, Apeldoorn, NL (09/2026).
- Invited Talk: "The Best of Both Worlds: Fixing Data-Driven NIDS with Neuro-Symbolic Logic"; Venue: TU Eindhoven (03/2026).
- Invited Talk: "On the Quest for Resource-Efficient AI: Lessons Learnt from Machine Learning and Neuro-Symbolic Integration"; Venue: Vrije Universiteit Amsterdam (04/2025).
- Seminar: "Neuro-Symbolic AI: The What and Why"; Venue: TU Delft (06/2025).
- Seminar: "Resource-Friendly Artificial Intelligence"; Venue: Delft University of Technology (01/2023).
- Seminar: "Graph Neural Networks - Theory & Applications"; Venue: University of Bologna (04/2021).

Grants & Funding Applications

Applicant for an NWO Veni Grant **Delft, NL**
- *Funding scheme*: NWO Talent Programme – Veni (Dutch Research Council), personal grant for early-career researchers.
- *Budget*: up to €320,000 over 3 years.
- *Status*: Submitted, decision pending.

Supervision & Teaching

Student Supervision:

- Master Thesis Supervisor for student Huib Sprangers working on thesis entitled "Self-supervised Learning for Privacy preserving Graph Neural Networks" - A.Y. 2026/2027.
- Master Thesis Supervisor for student Sophie Schaaf working on thesis entitled "Trajectory-Based Analysis of Backdoored Samples in Neural Networks" - A.Y. 2026/2027.
- CSE3000 Research Project (Bachelor Thesis) Supervisor involving 5 Bachelor students – A.Y. 2026/2027.
- Master Thesis Supervisor for student Raoul Saad, thesis entitled "Towards Neuro-Symbolic AI in Network Intrusion Detection Systems" - A.Y. 2025/2026, **graduated cum laude**; the thesis led to a first-author paper under submission.
- Master Thesis Supervisor for student Sofia Burkow, thesis entitled "DeepChronos: Neuro-Symbolic Reasoning for Multi-Step Attack Detection" - A.Y. 2025/2026, **graduated cum laude**; the thesis led to a first-author paper under submission.
- Master Thesis Supervisor for student Anna Visman, thesis entitled "A Neuro-Symbolic Framework for Alert Fatigue Reduction in Security Operation Center" - A.Y. 2025/2026, graduated.
- Master Thesis Supervisor for student Davide Alpi working on thesis entitled "Conversational Non-Player Characters: A Study on the Derailment of Large Language Models" - A.Y. 2022/2023.
- CSE3000 Research Project (Bachelor Thesis) Supervisor involving 5 Bachelor students – A.Y. 2024/2025.
- Supervisor of Hacking Lab Course Project involving 5 Master students - A.Y. 2024/2025.

Teaching Activities:

- Lecture: "Introduction to Natural Language Processing"; Course: Engineering Intelligent Systems; Venue: University of Bologna (05/2023).
- Teacher Assistant (Tutor) for the "Fondamenti di Informatica" course – A.Y. 2022/2023.
- Teacher Assistant (Tutor) for the "Fondamenti di Informatica" course – A.Y. 2021/2022.

Teaching Development:

- Following University Teaching Qualification (UTQ) courses at TU Delft, including the Empathy trainings on disengaged students, students' personal situations, and gender imbalance (ongoing).

Academic Service & Leadership

Institutional Service:

Committee Member of the PostDoc Network TU Delft **Delft, NL** *03/2026 - Ongoing*
The PostDoc Network represents the postdoctoral researchers of TU Delft. As a committee member, I work to give postdocs a voice in their departments, faculties, and the higher governing bodies of the university, bringing their perspective into the decisions that shape their positions and careers. In June 2026, I represented the network at the National Postdoc Day organized by PostdocNL in Utrecht.

Participant in the TU Delft Generative AI Workgroup **Delft, NL** *11/2026*
The workgroup, part of the ELLIS Delft Unit programmes, helps shape the university-wide stance and agenda on generative AI in research.

Organization:

Web-Chair at ECAI 2025 **Bologna, Italy**
I served as the Web Chair for ECAI 2025, the premier European conference on Artificial Intelligence, where I led the design, development, and management of the conference website.

Organizing Committee Member of the 1st HotDiSec Workshop **Rome, Italy**
Held in conjunction with ESORICS 2026. The HotDiSec workshop provides a platform for researchers and practitioners working in the distributed ML and security domain to share high-impact research, rigorous critiques, and forward-looking discussion papers.

Organizing Committee Member of the 1st ANSyA Workshop **Bologna, Italy**
Held in conjunction with ECAI 2025, the workshop's first edition was a success, gathering *30 submissions*. The ANSyA workshop provides a platform for applied neuro-symbolic systems engineering, fostering collaborations between academia and industry, and encouraging innovative thinking. ANSyA revitalized the applied dimension of neuro-symbolic AI, previously underestimated.

Organizing Committee Member of the 1st HotDiML Workshop **Glasgow, Scotland, UK**
Held in conjunction with ICDCS 2025, the workshop's first edition was a success, gathering *11 submissions*. HotDiML serves as a platform for critical thinking, sharing both successes and failures, and fostering constructive discussions about distributed machine learning applications. HotDiML encourages novel directions, accepting papers that offer fresh perspectives or critical analyses, even if not fully developed.

Publicity Chair of the 6th EXTRAAMAS Workshop **Auckland, New Zealand**
Held in conjunction with AAMAS 2024. I served as the Publicity Chair for the EXTRAAMAS 2024 workshop, one of the premier venues on the explainability and transparency of intelligent agent systems.

Program Committees & Reviewing:

Regular reviewer for top-tier venues in machine learning, artificial intelligence, and security, completing **40+ reviews per year**. Recognized as Gold, Outstanding, or Top Reviewer at ICML (2025, 2026), KDD (2025), and UAI (2023).

- **AI & ML conferences (PC member):** NeurIPS (2025–2026), ICML (2025–2026), ICLR (2026), KDD (2024–2027), AAAI (2023–2027), IJCAI (2027), UAI (2023, 2024, 2026).
- **Security conferences (PC member):** ESORICS (2025), ACM CODASPY (2025–2027).
- **Other conferences:** ACM SAC (PC member, 2025, 2027), IEEE INFOCOM (external reviewer, 2025).
- **Workshops (PC member):** NEmo @ NeurIPS (2026), EXTRAAMAS (2023–2025), WOA (2022–2024).
- **Journals (reviewer):** Transactions on Machine Learning Research (2025–2026), ACM Computing Surveys, ACM Transactions on Artificial Intelligence Security and Privacy, IEEE Transactions on Network and Service Management, IEEE Transactions on Services Computing, Future Generation Computer Systems, Knowledge-Based Systems, Computer Networks, AI and Ethics.

Skills

- **Programming:** Python, Java, C++, LaTeX (advanced); C, C#, MATLAB, Kotlin, Bash (intermediate); HTML, JavaScript, Prolog (basic).
- **ML/DL frameworks:** PyTorch, TensorFlow, Scikit-Learn (advanced); HuggingFace, Keras (intermediate); DL4J, ONNX (basic).
- **Tools & OS:** git, Unittest, JUnit, IntelliJ, PyCharm, Visual Studio; macOS, Windows, Linux, Android, iOS.

Languages

Italian [Native] • **English** [Proficient] - C1 • **Spanish** [Basic] - A1