Shahaf Bassan

Ph.D. Student, Interests: Explainable AI (XAI), Interpretability, ML Verification, ML Theory

Hebrew University of Jerusalem

shahaf.bassan@mail.huji.ac.il | bassanshahaf.github.io/

Education

Hebrew University of Jeruslaem — *Jerusalem, Israel*

Ph.D. in Computer Science, Topic: Explainability & Interpretability — *January 2023 - Current*

- Advisor: Prof. Guy Katz
- Recipient of the Federman Scholarship for Excellent Ph.D. Students in both 2023 and 2024

Hebrew University of Jerusalem — *Jerusalem, Israel*

M.Sc. in Computer Science — *2021 - 2022*

- Advisor: Prof. Jeff Rosenschein co-advised with Dr. Yossi Adi

Hebrew University of Jerusalem — *Jerusalem, Israel*

B.Sc. in Computer Science and Physics — *2017 - 2020*

Publications

Itamar Hadad, Guy Katz, **Shahaf Bassan**. Formal Mechanistic Interpretability: Automated Circuit Discovery with Provable Guarantees *[**ICLR**, 2026, Rio De Janiero]*[Paper]

Shahaf Bassan*, Xuanxiang Huang*, Guy Katz. Unifying Formal Explanations: A Complexity-Theoretic Perspective *[**ICLR**, 2026, Rio De Janiero]*[Paper]

Shahaf Bassan*, Yizhak Elboher*, Tobias Ladner*, Volkan Sahin, Jan Kretinsky, Matthias Althoff, Guy Katz. Provably Explaining Neural Additive Models *[**ICLR**, 2026, Rio De Janiero]*[Paper]

Ryma Boumazouza, Raya Elsaleh, Melanie Ducoffe, **Shahaf Bassan**, Guy Katz. FAME: Formal Abstract Minimal Explanation for Neural Networks *[**ICLR**, 2026, Rio De Janiero]*[Paper]

Shahaf Bassan, Michal Moshkovitz, Guy Katz. Additive Models Explained: A Computational Complexity Approach *[**NeurIPS**, 2025, San-Diego]*[Paper]

Reda Marzouk*, **Shahaf Bassan*** (equal contribution), Guy Katz. SHAP Meets Tensor Networks: Provably Tractable Explanations with Parallelism *[**NeurIPS**, 2025, San Diego]*[Paper]

Shahaf Bassan, Guy Amir, Meirav Zehavi, Guy Katz. What makes an Ensemble (Un) Interpretable? *[**ICML**, 2025, Vancouver]*[Paper]

Shahaf Bassan*, Yizhak Elboher*, Tobias Ladner* (equal contribution), Matthias Althoff, Guy Katz. Explaining, Fast and Slow: Abstraction and Refinement of Provable Explanations *[**ICML**, 2025, Vancouver]*[Paper]

Shahaf Bassan, Ron Eliav, Shlomit Gur. Explain Yourself, Briefly! Self-Explaining Neural Networks with Concise Sufficient Reasons *[**ICLR**, 2025, Singapore]*[Paper]

Reda Marzouk*, **Shahaf Bassan*** (equal contribution), Guy Katz, Colin de la Higuera. On the Computational Tractability of the (Many) Shapley Values *[**AISTATS**, 2025, Mai Khao, Thailand]*[Paper]

Shahaf Bassan, Guy Amir, Guy Katz. Local vs. Global Interpretability: A Computational Complexity Perspective *[**ICML**, 2024, Vienna, Austria]* (***Spotlight***) [Paper]

Ron Eliav, Arie Cattan, Eran Hirsch, **Shahaf Bassan**, Elias Stengel-Eskin, Mohit Bansal, Ido Dagan. CLATTER: Comprehensive Entailment Reasoning for Hallucination Detection *[preprint, 2025]*[Paper]

Shahaf Bassan*, Shlomit Gur*, Sergey Zeltyn*, Konstantinos Mavrogiorgos*, Ron Eliav, Dimosthenis Kyriazis. Self-Explaining Neural Networks for Business Process Monitoring *[**ICSBT**, 2025, Balboa, Spain, Best Paper Runnerup]*[Paper]

Guy Amir*, **Shahaf Bassan*** (equal contribution), Guy Katz. Hard to Explain: On the Computational Hardness of In-Distribution Model Interpretation *[**ECAI**, 2024, Santiago, Spain]*[Paper]

Haoze Wu, Omri Isac, Aleksandar Zeljić, Teruhiro Tagomori, Matthew Daggitt, Wen Kokke, Idan Refaeli, Guy Amir, **Shahaf Bassan**, Ori Lahav, Min Wu, Min Zhang, Ekaterina Komendantskaya, Guy Katz, and Clark Barrett.
Marabou 2.0: A Versatile Formal Analyzer of Neural Networks *[**CAV** 2024, Montreol, Canada [Paper]*

Shahaf Bassan*, Guy Amir*, Davide Corsi, Idan Refaeli, Guy Katz. Formally Explaining Neural Networks within Reactive Systems *[**FMCAD**, 2023, Iowa, USA, **Best Paper Runnerup**][Paper]*

Shahaf Bassan, Guy Katz. Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks. *[**Tacas**, 2023, Paris, France][Paper]*

Shahaf Bassan, Yossi Adi, Jeffrey S. Rosenschein. Symbolic Music Segmentation using Ensemble Temporal Prediction Errors *[**Interspeech**, 2022, Incheon, South Korea] [Paper]*

Teaching Experience

2021-2023 **Data Structures and Algorithms**, Teaching Assistant.
2021, 2022, 2023: Chosen for the list of outstanding teachers at the Hebrew University.
2022, 2023: Ranked **first** in student surveys among all computer science and engineering faculty members at the Hebrew University, out of hundreds of teachers and lecturers.

Invited Talks

December 2025 **Invited Keynote Speaker** at the Theory of Explainable ML Workshop, Copenhagen *[Workshop link]*

October 2025 **Invited Speaker** at the Theory of Interpretable AI Seminar: "The Computational Complexity of Explaining ML Models" *[Talk]*

June 2025 **Invited Speaker** at the IARCS Verification Seminar: "Formal XAI": Can we formally explain ML models? *[Slides]*

March 2025 **Invited Speaker** at the Interpretability Research Group, Tel Aviv: "Explain Yourself, Briefly! Self-Explaining Neural Networks with Concise Sufficient Reasons"

February 2025 **Invited Speaker** at Kings College, London: "Towards Formal Explainability of Neural Networks"

July 2024 **Invited Speaker** at Bosch Research Center for AI, Haifa: "Formal XAI: Formally Explaining the Decisions of ML Models"

September 2023 **Invited Speaker** at IBM Research, Haifa: "Formal XAI: Can we Formally Explain ML Models?"

June 2023 **Keynote Speaker** at Mobilit.AI, Toulouse, France: "Towards Formally Verifying and Explaining Deep Neural Networks" *[Talk]*

April 2023 **Invited Speaker** at TADM Workshop, Paris, France: "Formal Verification of Large Language Models"

April 2023 **Invited Speaker** at Live Workshop, Paris, France: "Formal Explainability of Deep Neural Networks"

Professional Experience

2023 **Explainable AI Research Intern**, IBM Research (Summer Internship)
Research Topic: Explainability during training. Two papers published during the internship, one at ICLR (see publications)

2019-2020 **Software Engineer**, Fiverr (Key involvement in building a new major feature: "Fiverr for Business")

Volunteer Experience

2020-2024 **Alumni Forum Co-Leader and Key Contributor**, Tech2Peace
Tech2Peace is an NGO with a noble mission to bring Israelis and Palestinians together through technology and entrepreneurship.

2021-Present **Programming Teacher and Lecturer**, Tech2Peace

Community Service

2025-Present **Theory of Interpretable AI seminar (an international monthly seminar on the theoretical foundations of interpretability)**, Co-Organizer

2025 **Selected as an outstanding reviewer**, ICML

2024-Present **ICLR, ICML, NeurIPS, Tacas**, Reviewer