ALON JACOVI

Curriculum Vitae

Tel Aviv, Israel ⋄ alonjacovi@gmail.com

EDUCATION

Ph.D in Natural Language Processing and Machine Learning *2019–2023 (expected)*
Bar Ilan University
Advisor: Prof. Yoav Goldberg
Research Topics: Explainable Artificial Intelligence, Learning from Imperfect Supervision

M.Sc in Natural Language Processing and Machine Learning *2017–2019*
Bar Ilan University
Advisor: Prof. Yoav Goldberg
Graduated *cum laude* with final grade 95.85.
Dissertation Topic: Understanding Convolutional Neural Networks for Text Classification

B.Sc in Computer Science *2014–2017*
Bar Ilan University
Graduated *cum laude* with final grade 95.2.
Final Project: Deep Reinforcement Learning Agent for the AI-BIRDS Angry Birds Competition

EXPERIENCE

Research Intern at Google, Israel *May 2022–Jan 2024 (expected)*
Consistency Team
Hosts: Dr. Roee Aharoni and Dr. Mor Geva
Topics: Evaluation of chain-of-thought verification methods, evaluation of tool-assisted complex reasoning strategies, and a user study of crowd-sourced evaluations of explanations.

Research Intern at Google, Germany *Fall 2021*
Trustworthy NLP Team
Hosts: Dr. Jasmijn Bastings and Dr. Katja Filippova
Topic: Diagnosing AI explanation methods with folk concepts of behavior.

Research Intern at the Allen Institue for Artificial Intelligence, USA *Winter 2020-2021*
MOSAIC Team
Host: Dr. Swabha Swayamdipta
Topic: Contrastive explanations for model interpretability.

Student Researcher at IBM Research, Israel *2016–2020*

- Conversation and Language Team
  Work on improving graph-based virtual dialogue assistants, based on in-production conversation logs that were escalated to a human agent.
- Machine Learning Technologies Team
  Work on integration of non-differentiable systems in neural pipelines (e.g., knowledge bases and program interpreters in semantic parsing) for end-to-end learning.

Research Intern at RIKEN, Japan *Spring 2019*
Imperfect Information Learning Team
Hosts: Dr. Gang Niu and Prof. Masashi Sugiyama
Topic: Scalable evaluation and improvement of Document Set Expansion via neural Positive-Unlabeled Learning.

PUBLICATIONS

Diagnosing AI Explanation Methods with Folk Concepts of Behavior.
Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, Katja Filippova.
In JAIR 2023.

Stop Uploading Test Data in Plain Text: **Practical Strategies for Mitigating Data Contamination of Your Evaluation Benchmark.**
Alon Jacovi, Avi Caciularu, Omer Goldman, Yoav Goldberg.
In EMNLP 2023.

A Comprehensive Evaluation of Tool-Assisted Generation Strategies.
Alon Jacovi, Avi Caciularu, Jonathan Herzig, Roee Aharoni, Bernd Bohet, Mor Geva.
In Findings of EMNLP 2023.

Neighboring Words Affect Human Interpretation of Saliency Explanations.
Alon Jacovi*, Hendrik Schuff*, Heike Adel, Yoav Goldberg, Ngoc Thang Vu.
In Findings of ACL 2023

Human Interpretation of Saliency-based Explanation Over Text.
Hendrik Schuff*, Alon Jacovi*, Heike Adel, Yoav Goldberg, Ngoc Thang Vu.
In ACM FAccT 2022.

Contrastive Explanations for Model Interpretability.
Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, Yoav Goldberg.
In EMNLP 2021.

Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI.
Alon Jacovi, Ana Marasović, Tim Miller, Yoav Goldberg.
In ACM FAccT 2021.

Scalable Evaluation and Improvement of Document Set Expansion via Neural Positive-Unlabeled Learning.
Alon Jacovi, Gang Niu, Yoav Goldberg, Masashi Sugiyama.
In EACL 2021.

Aligning Faithful Interpretations with their Social Attribution.
Alon Jacovi, Yoav Goldberg.
In TACL 2021.

Amnesic Probing: Behavioral Explanations with Amnesic Counterfactuals.
Yanai Elazar, Shauli Ravfogel, Alon Jacovi, Yoav Goldberg.
In TACL 2021.

Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data.
Shachar Rosenman, Alon Jacovi, Yoav Goldberg.
In EMNLP 2020.

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?
Alon Jacovi, Yoav Goldberg.
In ACL 2020.

Improving Task-Oriented Dialogue Systems in Production with Conversation Logs.
Alon Jacovi*, Ori Bar El*, Ofer Lavi, David Boaz, David Amid, Inbal Ronen, Ateret Anaby-Tavor.
In KDD Converse @ KDD 2020.

Neural Network Gradient-based Learning of Black-box Function Interfaces.
Alon Jacovi*, Guy Hadash*, Einat Kermany*, Boaz Carmeli*, Ofer Lavi, George Kour, Jonathan Berant.
In ICLR 2019.

Learning and Understanding Different Categories of Sexism Using Convolutional Neural Network Filters.
Sima Sharifirad, Alon Jacovi, Stan Matwin.
In Widening NLP @ ACL 2019. *(extended abstract)*

Understanding Convolutional Neural Networks for Text Classification.
Alon Jacovi, Oren Sar Shalom, Yoav Goldberg.
In BlackboxNLP @ EMNLP 2018. *(oral presentation)*

(∗) - equal contribution.

PATENTS

Computerized dialog system improvements based on conversation data. *Published, 2022*
Ofer Lavi, Alon Jacovi, David Amid, David Boaz, Inbal Ronen, Ateret Anaby-Tavor, Ori Bar El.
US Patent: US20220059097A1

ACADEMIC SERVICE

Workshop organizer: Data Contamination Workshop (CONDA) 2024

Area Chair: ACL Rolling Review 2021 and 2022, ACL 2023

Reviewer: ACL 2020, HAMLETS 2020, EACL 2021, NAACL 2021, ACL 2021 *(outstanding reviewer)*, EMNLP 2021, BlackboxNLP 2021, TRAIT 2022, BlackboxNLP 2022, HMCaT 2022, Open Mind 2022

Assistant Reviewer: IJCAI 2018, IJCAI 2019

TALKS

FAccT, Jun 2023: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

DNV, Mar 2023: *Formalizing Trust in Artificial Intelligence.*

Institute for Trusted Intelligence in Society Workshop, Sep 2022: *Formalizing Trust in Artificial Intelligence.*

Trustworthy ML Reading Group, Sep 2022: *Formalizing Trust in Artificial Intelligence.*

FAccT, Jun 2022: *Human Interpretation of Saliency-based Explanation Over Text.*

XAI Israel Meeting Group, May 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

Cambridge, May 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

Harvard, Mar 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

Oxford, Mar 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

NEC Labs Europe, Mar 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

EMNLP, Nov 2021: *Aligning Faithful Interpretations with their Social Attribution.*

Allen Institute for Artificial Intelligence, Jun 2021: *Formalizing Trust in Artificial Intelligence.*

Microsoft, Oct 2020: *Formalizing Properties of Interpretability in NLP.*

ACLi, Aug 2020: *Towards Faithfully Interpretable NLP Systems.*

BlackboxNLP 2018: *Understanding Convolutional Neural Networks for Text Classification.*

ISCOL, Sep 2018: *Understanding Convolutional Neural Networks for Text Classification.*

HONORS AND SCHOLARSHIPS

2021	Nadav Award for Excellence in Doctoral Studies
2019–2023	The President's Scholarship for Outstanding Doctoral Fellows
2016	The Miryam and Ezra Sofer Scholarship for Excellent Students
2016	Dean's Honors
2015	The Grace Shua and Jacob Ballas Scholarship
2015	Dean's Honors

LANGUAGES

Native	Hebrew
Fluent	English
Conversational	Japanese