

# ALON JACOVI

Curriculum Vitae

Tel Aviv, Israel ◊ alonjacovi@gmail.com

## EDUCATION

---

- Ph.D** in Natural Language Processing and Machine Learning *2019–2023*  
Bar Ilan University  
Advisor: Prof. Yoav Goldberg  
Research Topic: Explainable Artificial Intelligence
- 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–Mar 2024 (expected)*  
Consistency Team  
Hosts: Dr. Roei 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 Institute 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.

## SELECTED 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 by Evaluation Benchmarks.***

Alon Jacovi, Avi Caciularu, Omer Goldman, Yoav Goldberg.  
In EMNLP 2023. (*oral presentation*)

**A Comprehensive Evaluation of Tool-Assisted Generation Strategies.**

Alon Jacovi, Avi Caciularu, Jonathan Herzig, Roei Aharoni, Bernd Bohet, Mor Geva.  
In Findings of EMNLP 2023.

**Neighboring Words Affect Human Interpretation of Saliency Explanations.**

Alon Jacovi\*, Hendrik Schuff\*, Heike Adel, Ngoc Thang Vu, Yoav Goldberg.  
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.

**Aligning Faithful Interpretations with their Social Attribution.**

Alon Jacovi, Yoav Goldberg.  
In TACL 2021.

**Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?**

Alon Jacovi, Yoav Goldberg.  
In ACL 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.

**Understanding Convolutional Neural Networks for Text Classification.**

Alon Jacovi, Oren Sar Shalom, Yoav Goldberg.  
In BlackboxNLP @ EMNLP 2018. (*oral presentation*)

(\* ) - equal contribution.

***IN SUBMISSION***

**Unpacking Human-AI Interaction in Safety-Critical Industries: A Systematic Literature Review**

Tita A. Bach, Jenny K. Kristiansen, Aleksandar Babic, Alon Jacovi.

## PATENTS

---

**Computerized dialog system improvements based on conversation data.** *Published, 2022*  
Ofar 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, NAACL 2024

Assistant Reviewer: IJCAI 2018, IJCAI 2019

## INVITED TALKS

---

**DEEL & ANITI, Nov 2023:** *What would a good explanation even look like?*

**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.*

**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.*

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

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

## 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