idanshenfeld@gmail.com \ idanshen@mit.edu
857-2051549
www.idanshenfeld.com

# IDAN SHENFELD

## PROFILE

EECS PhD at MIT, working on reinforcement learning algorithms and their applications, mainly in NLP and robotics. 3+ years of industrial research experience in various global companies.

## EDUCATION

**PH.D. IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE, MIT**

- September 2022 – present. (GPA 5.00/5.00)
- advised by **Professor Pulkit Agrawal**.

**B.S. IN COMPUTER ENGINEERING, TECHNION**

- Class of 2021, Summa Cum Laude (GPA 94.8/100).
- Expedited course of study (3 years instead of 4); Rothschild scholarship; Apple Award for Excellence BSc Students; Dean's or Rector's List, all semesters.
- Conducted research in Reinforcement Learning under the supervision of **Professor Aviv Tamar**, and in Geometrical Learning under the supervision of **Professor Ron Kimmel**.

## PROFESSIONAL EXPERIENCE

**APPLIED RESEARCHER, GENERAL MOTORS AV PROJECT**
August 2021- September 2022

- Performing research as part of the Perception group in GM autonomous vehicle project.
- Focus on problems such as General Obstacle Detection, Road Segmentation, Sensor Fusion and more.

**MACHINE LEARNING ALGORITHM ENGINEER, SAMSUNG FLASH SOLUTIONS RESEARCH LAB (AFSL)**
October 2017- October 2018

- Research of machine learning based algorithms for storage systems. Development and incorporation of machine learning techniques into storage controllers.
- Lead the research of innovative error correction modules for storage systems that integrate classic ECC techniques and DL algorithms.

**DATA AND INTELLIGENCE ANALYST DEPARTMENT LEADER, UNIT 8200, ISRAELI DEFENCE FORCE**
July 2015- October 2017

- Military service at Unit 8200, the Israeli equivalent of the NSA. Finished the service as an officer at the rank of First Lieutenant.
- Managed 4 multidisciplinary teams, with a total of approximately 40 employees.
- Award of excellence given to me on the exemplary department functioning.

**DATA ANALYST \ SCIENTIST, UNIT 8200, ISRAELI DEFENCE FORCE**
May 2013-January 2015

- Intelligence reports and data analyses, many times under time-sensitive conditions.
- Research and Development of classification model for valuable network segments. The project won the “Chief of Intelligence prize for Outstanding Projects”.

## PUBLICATIONS

- **Guidance Functions Meets RL – Aligning Black-Box Large Language Models.**
  **Idan Shenfeld**, Seungwook Han, Akash Srivastava, Yoon Kim, Pulkit Agrawal. (Under review).

- **Curiosity-driven Red-teaming for Large Language Models.**
  Zhang-Wei Hong, **Idan Shenfeld**, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James R. Glass, Akash Srivastava, Pulkit Agrawal. ICLR, 2024 (Under review).

- **TGRL: An Algorithm for Teacher Guided Reinforcement Learning.**
  **Idan Shenfeld**, Zhang-Wei Hong, Aviv Tamar, Pulkit Agrawal. ICML, 2023.

- **Offline Meta Reinforcement Learning - Identifiability Challenges and Effective Data Collection Strategies.**
  Ron Dorfman, **Idan Shenfeld**, Aviv Tamar. NeurIPS, 2021.