Andrew Ilyas

ailyas@mit.edu | andrewilyas.com | andrewilyas

Education

Massachusetts Institute of Technology (PhD Candidate, EECS) **2018–Present**
Advisors: Aleksander Mądry and Constantinos Daskalakis
Thesis (tentative): From data, to models, and back—towards robust and reliable ML systems

Massachusetts Institute of Technology (B.S. EECS, B.S. Math, M.Eng. EECS) **2015-2018**
M.Eng. Advisor: Constantinos Daskalakis
M.Eng. Thesis: On the practical robustness of machine learning systems

Research Focus

My research pursues a precise empirical understanding of the entire ML pipeline, with an emphasis on data. My interests span **tracing predictions back to training data,** identifying and alleviating **data bias,** and studying machine learning **robustness.** I also like thinking more broadly about **trust in AI systems**, and have had the opportunity to contribute to writings on social media regulation and AI deployment.

Honors & Awards

Open Philanthropy Project AI Fellowship	2019-2023
Analog Devices Graduate Fellowship	2018-2019
M.Eng. Thesis Award, *MIT*	2018
SuperUROP Award, *MIT*	2017
Hackathon Winner, *Andreesen Horowitz*	2016, 2017

Peer-Reviewed Publications

* denotes equal (first-author) contribution, (α-β) alphabetical author ordering

1. Sung Min Park*, Kristian Georgiev*, **A Ilyas***, Guillaume Leclerc, Aleksander Madry. "TRAK: Understanding Model Predictions." International Conference on Machine Learning (ICML) 2023. **Oral.**

2. Harshay Shah*, Sung Min Park*, **A Ilyas*,** Aleksander Madry. "ModelDiff: A Framework for Comparing Learning Algorithms." ICML, 2023. (+ **Workshop Oral**, ICML Workshop on Spurious Correlations, Invariance and Stability)

3. Hadi Salman*, Alaa Khaddaj*, Guillaume Leclerc*, **A Ilyas**, and Aleksander Madry. Raising the Cost of Malicious AI-Powered Image Editing. ICML, 2023. **Oral.**

4. Alaa Khaddaj*, Guillaume Leclerc*, Alexander Makelov*, Kristian Georgiev*, Hadi Salman, **A Ilyas,** Aleksander Madry. Rethinking Backdoor Attacks. ICML, 2023.

5. (α-β) Yeshwanth Cherapanamjeri, Constantinos Daskalakis, **A Ilyas**, Manolis Zampetakis. "What Makes A Good Fisherman? Linear Regression under Self-Selection Bias." Symposium on Theory of Computation (STOC), 2023.

6. (α-β) Yeshwanth Cherapanamjeri, Constantinos Daskalakis, **A Ilyas**, Manolis Zampetakis. "Estimating Standard Auction Models." Economics and Computation (EC), 2022.

7. Guillaume Leclerc*, Hadi Salman*, **A Ilyas***, Sai Vemprala, Logan Engstrom, Vibhav Vineet, Kai Xiao, Pengchuan Zhang, Shibani Santurkar, Greg Yang, Ashish Kapoor, Aleksander Madry. "3DB: A Framework for Debugging Computer Vision Models." Neural Information Processing Systems (NeurIPS), 2022.

8. **A Ilyas***, Sung Min Park*, Logan Engstrom*, Guillaume Leclerc, Aleksander Madry. "Datamodels: Predicting Predictions from Training Data." ICML, 2022.

9. Mihaela Curmei*, **A Ilyas***, Owain Evans, Jacob Steinhardt. "Constructing and Adjusting Estimates for Household Transmission of SARS-CoV-2 from Prior Studies, Widespread-Testing and Contact-Tracing Data." International Journal of Epidemiology, 2021.

10. Hadi Salman*, **A Ilyas*,** Logan Engstrom*, Sai Vemprala, Aleksander Madry, Ashish Kapoor. "Unadversarial Examples: Designing Objects for Robust Vision." NeurIPS, 2021.

11. Kai Xiao, Logan Engstrom, **A Ilyas**, Aleksander Madry. "Noise or Signal: The Role of Image Backgrounds in Object Recognition." International Conference on Learning Representations (ICLR), 2021.

12. Hadi Salman*, **A Ilyas***, Logan Engstrom, Ashish Kapoor, Aleksander Madry. Do Adversarially Robust ImageNet Models Transfer Better? NeurIPS, 2020. **Oral.**

13. Logan Engstrom*, **A Ilyas***, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, Aleksander Madry. "Identifying Statistical Bias in Dataset Replication." ICML, 2020.

14. Dimitris Tsipras*, Shibani Santurkar*, Logan Engstrom, **A Ilyas**, Aleksander Madry. "From ImageNet to Image Classification: Contextualizing Progress on Benchmarks." ICML, 2020.

15. **A Ilyas***, Logan Engstrom*, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, Aleksander Madry. "A Closer Look at Deep Policy Gradients." ICLR, 2020. **Oral.**

16. Logan Engstrom*, **A Ilyas***, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, Aleksander Madry. "Implementation Matters in Deep RL: A Case Study on PPO and TRPO." ICLR, 2020. **Oral.**

17. **A Ilyas**, Emmanouil Zampetakis, Constantinos Daskalakis. "A Theoretical and Practical Framework for Regression and Classification from Truncated Samples." Conference on Artificial Intelligence and Statistics (AISTATS), 2020.

18. **A Ilyas*,** Shibani Santurkar*, Dimitris Tsipras*, Logan Engstrom*, Brandon Tran, Aleksander Madry. Adversarial Examples are not Bugs, they are Features. NeurIPS, 2019. **Spotlight.**

19. Shibani Santurkar*, **A Ilyas***, Dimitris Tsipras*, Logan Engstrom*, Brandon Tran*, Aleksander Madry. "Image Synthesis with a Single (Robust) Classifier." NeurIPS, 2019.

20. **A Ilyas***, Logan Engstrom*, Aleksander Madry. "Prior Convictions: Black-Box Adversarial Attacks with Bandits and Priors." ICLR, 2019.

21. **A Ilyas***, Logan Engstrom*, Anish Athalye*, Jessy Lin. Black-box Adversarial Examples with Limited Queries and Information. ICML, 2018.

22. Anish Athalye*, Logan Engstrom*, **A Ilyas***, Kevin Kwok. "Synthesizing Robust Adversarial Examples." ICML, 2018. (+ **Workshop Oral**, NeurIPS 2018 ML Security Workshop)

23. Shibani Santurkar*, Dimitris Tsipras*, **A Ilyas***, Aleksander Madry. "How does Batch Normalization help Optimization?" NeurIPS, 2018. **Oral.**

24. (α-β) Constantinos Daskalakis, **A Ilyas**, Vasilis Syrgkanis, Haoyang Zeng. "Training GANs with Optimism." ICLR, 2018.

25. **A Ilyas**, Joana MF da Trindade, Raul C. Fernandez, Samuel Madden. "Extracting Syntactical Patterns from Databases." International Conference on Data Engineering (ICDE), 2018.

26. **A Ilyas. "**MicroFilters: Harnessing Twitter for Disaster Management." IEEE Global Humanitarian Technology Conference (GHTC), 2014.

Working Papers & Other Writing

1. Sarah H. Cen*, **A Ilyas***, Jennifer Allen, Hannah Li, David Rand, Aleksander Madry. Measuring User Strategization on Data-Driven Recommender Systems (2023). Working paper. (+ **Oral,** Conference on Digital Experimentation/CODE).

2. Sarah H. Cen*, **A Ilyas***, Aleksander Madry. User Trust and Strategization on Data Driven Platforms (2023). Working paper. (+ **Oral,** ICML Workshop on Responsible Decision Making in Dynamic Environments).

3. Hadi Salman*, Saachi Jain*, **A Ilyas**, Logan Engstrom, Eric Wong, Aleksander Madry (2022). When Does Bias Transfer in Transfer Learning? arXiv preprint.

4. (α-β) Sarah H. Cen, Aspen Hopkins, **A Ilyas**, Aleksander Madry, Isabella Struckman, Luis Videgaray. Blog Series on AI Deployment (2023). https://aipolicy.substack.com/t/on-ai-deployment-series

5. Sarah H. Cen, **A Ilyas**, Aleksander Madry. Blog Series on Regulating Social Media (2022). https://aipolicy.substack.com/p/socialmediaseries

Selected Talks

Invited Talks

Microsoft Research, *Attributing model behavior at scale*	2023
TrustML Young Scientist Seminar, *Datamodels: predicting predictions from training data*	2023
Stanford MedAI Seminar, *Datamodels: predicting predictions from training data*	2022
Google Brain, *Datamodels: predicting predictions from training data*	2022
SIAM Mathematics of Data Science, *Datamodels: predicting predictions from training data*	2022
OpenAI, *Datamodels: predicting predictions from training data*	2022
Samsung AI Centre, *An empirical analysis of deep learning phenomena*	2020
MIT Vision Seminar, *Identifying bias in dataset replication*	2020
Berkeley CHAI, *A closer look at deep policy gradient algorithms*	2020
Microsoft Research, *How does batch normalization help optimization?*	2019
Simons Institute, *Adversarial examples are not bugs, they are features*	2019
Two Sigma, *A closer look at deep policy gradient algorithms*	2019
Two Sigma, *Robust adversarial examples*	2018
Intel Labs, *3D adversarial examples*	2018

Guest Lectures

University of Waterloo, *Course: Deep learning (graduate)*	2021
Harvard Law School, *Course: Ethics and Governance of AI*	2018
UT Austin, *Course: Machine learning (graduate)*	2018

Meetings and Symposia

CSAIL Imagination in Action, *Building AI we can trust*	2023
INFORMS 2022, *Estimating standard auction models*	2022
MSR-TRAC workshop, *A closer look at deep policy gradient algorithms*	2020
NY Academy of Sciences, *Training GANs with optimism (spotlight)*	2019
O'Reilly AI Summit, *Robust adversarial examples*	2018

Selected Open-Source Projects

[2600 ★] Fast Forward Computer Vision (FFCV): A library for accelerating machine learning model training by removing data loading bottlenecks (link to code).

[1800 ★] Falcon: A chrome extension that improves browser history search by allowing users to search for web page content and images (link to code).

[800 ★] Robustness library: A library for making adversarial training of machine learning models easy and reliable (link to code).

Selected Press (by Project)

Personal profiles

- MIT News (link), *by Kim Martineau*
    - *"Two longtime friends explore how computer vision systems go awry"*

Robust Adversarial Examples

- The Verge (link), *by James Vincent*
    - *"Google's AI thinks this turtle looks like a gun, which is a problem"*
- BBC News (link)
    - *"AI image recognition fooled by single pixel change"*
- The Guardian (link), *by Alex Hern*
    - *"Shotgun shell: Google's AI thinks this turtle is a rifle"*

Black-Box Adversarial Attacks

- MIT Technology Review (link), *by Jackie Snow*
    - *"Computer vision algorithms are still way too easy to trick"*
- IEEE Spectrum (link), *by Jeremy Hsu*
    - *"Hacked dog pics can play tricks on computer vision AI"*
- Fortune Magazine (link), *by David Z. Morris*
    - *"How Google AI was tricked into thinking this photo of machine guns was a helicopter"*

Adversarial Examples are Not Bugs, They are Features

- Science Magazine (link), *by Matthew Hutson*
    - *"Scientists help artificial intelligence outsmart hackers"*
- WIRED Magazine (link), *by Louise Matsakis*
    - *"Artificial Intelligence May Not Hallucinate After All"*

PhotoGuard

- MIT News (link), *by Rachel Gordon*
    - *"Using AI to protect against AI image manipulation"*
- VentureBeat (link), *by Victor Dey*
    - *"MIT CSAIL unveils PhotoGuard, an AI defense against unauthorized image manipulation"*
- Engadget (link), *by Andrew Tarantola*
    - *"MIT's 'PhotoGuard' protects your images from malicious AI edits"*

Professional Experience

Labs Intern, *Two Sigma Investments* Summer 2018

Researched the underpinnings deep RL algorithms. Based on our work, co-authored two papers, both oral presentations (top 1% of accepted papers) at ICLR 2020.

Undergraduate Research Assistant, *MIT, supervised by:* 2015-2018

Dr. Xavier Boix & Prof. Tomaso Poggio (Deep neural networks invariances)
Dr. Raul C. Fernandez & Prof. Sam Madden (Database structure extractions)
Carl Vondrick & Prof. Antonio Torralba (Predictive power of CNNs)
Prof. Constantinos Daskalakis (Last-iterate convergence of gradient descent)

Labs Intern, *Two Sigma Investments* Summer 2017

Studied robust online optimization with an application to portfolio selection.

Machine Learning Intern, *Twine Health (acquired by Fitbit/Google)* Summer 2016

Sole member of the ML team, responsible for all data science and ML initiatives.

WatchOS Intern, *Cambridge Mobile Telematics* Summer 2015

Built a (commercially-deployed) Apple Watch app from scratch.

Data Science Intern, *Cambridge Mobile Telematics* Summer 2014

Worked w/ Prof. Sam Madden on texting-while-driving detection from phone data.

Academic Service

Refereeing

Journal Reviewer, *Journal of Machine Learning Research (JMLR)*	2021, 2023
Journal Reviewer, *Transactions on Machine Learning Research (TMLR)*	2022-2023
Expert reviewer certification	2023
Reviewer, *Neural Information Processing Systems (NeurIPS)*	2018, 2019, 2020, 2021, 2022
Outstanding reviewer award	2021
Reviewer, *International Conference on Machine Learning (ICML)*	2019, 2020, 2021, 2022
Top reviewer award	2020
Reviewer, *International Conference on Learning Representations (ICLR)*	2022, 2023
Reviewer, *Computer Vision and Pattern Recognition (CVPR)*	2021
Reviewer, *Conference on Learning Theory (COLT)*	2019
Sub-reviewer, *Foundations of Computer Science (FOCS)*	2018, 2020

Organizing/Chairing

Area Chair, *Neural Information Processing Systems (NeurIPS)*	2023
Session Chair, *INFORMS*	2022, 2023
Organizer, *Workshop on Attributing Model Behavior at Scale (ATTRIB @ NeurIPS)*	2023
Organizer, *HackMIT*	2015-2017

Mentorship/Volunteering

Supervisor, eight undergraduate researchers & two masters students	2019-2023
Graduate Student Mentor, *Graduate Application Assistance Program*	2023
Technical Mentor & Volunteer Judge, *Blueprint (HS hackathon at MIT)*	2018,2019
Technical Mentor, *HackMIT*	2018,2019
High School Math/Physics Teacher, *IS Enrico Fermi, Mantova, Italy*	2017

Miscellanea/Extra-curricular interests

Instruments: Piano (recreationally/RCM 9), Violin (recreationally)
Sports: Soccer (intramural), Table Tennis (club/competitive), Cycling (recreationally)
Languages: English (native), French (proficient/DELF B2), Egyptian Arabic (spoken