Dylan Slack

Website: dylanslacks.website
Email: dslack@uci.edu
Scholar: scholar.google.com/dylanslack
GitHub: github.com/dylan-slack

Employment

Google DeepMind
Research Scientist 2024 - Present

Scale AI
Senior Machine Learning Research Scientist 2023 - 2024

Google AI 2021
Research Intern

Amazon Web Services (AWS) 2020
Applied Scientist Intern

Education

UC Irvine 2019 - 2023
Ph.D. Computer Science
Advisors: Sameer Singh (UC Irvine), Himabindu Lakkaraju (Harvard University)
- Dissertation: *Robust Interactions With Machine Learning Models*

Haverford College 2015 - 2019
B.S. Computer Science (High Honors, Magna Cum Laude)
Advisor: Sorelle Friedler

Publications

Learning Goal-Conditioned Representations for Language Reward Models. Vaskar Nath*, **Dylan Slack***, Jeff Da, Yuntao Ma, Hugh Zhang, Spencer Whitehead, and Sean Hendryx. NeurIPS, 2024.

A Careful Examination of Large Language Model Performance on Grade School Arithmetic. Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, **Dylan Slack**, Qin Lyu, Sean Hendryx, Russell Kaplan, Summer Yue. NeurIPS D&B, 2024 (Spotlight).

Post Hoc Explanations of Language Models Can Improve Language Models. Satyapriya Krishna, Jiaqi Ma, **Dylan Slack**, Asma Ghandeharioun, Sameer Singh, Himabindu Lakkaraju. NeurIPS, 2023.

TalkToModel: Understanding Machine Learning Models With Open Ended Dialogues. **Dylan Slack**, Satyapriya Krishna, Hima Lakkaraju*, and Sameer Singh*. Nature Machine Intelligence, 2023.

Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. **Dylan Slack**, Sophie Hilgard, Sameer Singh, and Hima Lakkaraju. NeurIPS, 2021.

Counterfactual Explanations Can Be Manipulated. **Dylan Slack**, Sophie Hilgard, Hima Lakkaraju, and Sameer Singh. NeurIPS, 2021.

Active Meta-Learning for Predicting and Selecting Perovskite Crystallization Experiments. Venkateswaran Shekar, Gareth Nicholas, Mansoor Ani Najeeb, Margaret Zeile, Vincent Yu, Xiaorong Wang, **Dylan Slack**, Zhi Li, Philip Nega, Emory Chan, Alexander Norquist, Joshua Schrier, Sorelle Friedler. The Journal of Chemical Physics, 2021.

On the Lack of Robust Interpretability of Neural Text Classifiers. Muhammad Bilal Zafar, Michele Donini, **Dylan Slack**, Cedric Archambeau, Sanjiv Das, Krishnaram Kenthapadi. Findings of ACL, 2021.

Context, Language Modeling, and Multimodal Data in Finance. Sanjiv Das, Connor Goggins, John He, George Karypis, Sandeep Krishnamurthy, Mitali Mahajan, Nagpurnanand Prabhala, **Dylan Slack**, Rob van Dusen, Shenghua Yue, Sheng Zha, Shuai Zheng. The Journal of Financial Data Science, 2021.

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. **Dylan Slack***, Sophie Hilgard*, Emiliy Jia, Sameer Singh, and Himabindu Lakkaraju. AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 2020.

Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data. **Dylan Slack**, Sorelle Friedler, and Emile Givental. ACM Conference on Fairness, Accountability and Transparency (FAccT), 2020.

Workshop

Rethinking Explainability as a Dialogue: A Practitioner's Perspective. Himabindu Lakkaraju*, **Dylan Slack***, Yuxin Chen, Chenhao Tan, and Sameer Singh, NeurIPS HCAI Workshop, 2022.

SAFER: Data-Efficient and Safe Reinforcement Learning via Skill Acquisition. **Dylan Slack**, Yinlam Chow, Bo Dai, and Nevan Wichers, ICML DARL Workshop, 2022.

Defuse: Training More Robust Models through Creation and Correction of Novel Model Errors. **Dylan Slack**, Nathalie Rauschmayr, Krishnaram Kenthapadi. NeurIPS XAI 4 Debugging Workshop 2021.

Feature Attributions and Counterfactual Explanations Can Be Manipulated. **Dylan Slack**, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI, 2021.

Differentially Private Language Models Benefit from Public Pre-training. Gavin Kerrigan*, **Dylan Slack***, and Jens Tuyls*. EMNLP PrivNLP Workshop, 2020.

Assessing the Local Interpretability of Machine Learning Models. **Dylan Slack**, Sorelle A. Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy. NeurIPS Workshop on Human-Centric Machine Learning, 2019.

Technical Reports

A Holistic Approach For Test and Evaluation of Large Language Models. **Dylan Slack***, Jean Wang*, Denis Semenenko*, Kate Park, Daniel Berrios, Sean Hendryx. 2023.

Selected Awards

Honorable Mention Outstanding Paper, NeurIPS 2022 TSRML Workshop
NeurIPS Outstanding Reviewer, 2021/2022
ICLR Outstanding Reviewer, 2021
Hasso Plattner Institute Fellow, 2021 (Full Ph.D. Funding)

Preprints

TABLET: Learning From Instructions For Tabular Data. **Dylan Slack**, Sameer Singh. arXiv, 2023.

Presentations

Invited Talks & Presentations

- Stanford MedAI, 2022. *Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations*
- Imperial College, 2022. *Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations*
- Meta, 2022. *Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations*
- UCI CML, 2022. *Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations*
- Harvard University, 2021. *Reliable Post Hoc Explanations*

- Aggregate Intellect, 2021. *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods*

Patents

Automatic Failure Diagnosis and Correction in Machine Learning Models
Nathalie Rauschmayr, Krishnaram Kenthapadi, and **Dylan Slack**
Patent Application Filed

Academic Service

Community

- KDD Deep Learning Day, Organizer, 2021.

Program Committee Member

- NeurIPS (2019, 2020, 2021, 2022), FAccT (2021), ICLR (2021), ICML (2020), AAAI (2020, 2021), KDD (2019).