Aida Mostafazadeh Davani

Google Research
555 SW Morrison Street, Suite 500
Portland, OR 97204 USA

aidamd@google.com
https://aidamd.github.io/
Cell Phone: +1 (323) 449 - 4538

Research Interests

- ◇ Fairness in Machine Learning
- ◇ Natural Language Processing
- ◇ Computational Social Science

Work & Research Experience

- ◇ **Research Scientist** at Google LLC — 2022 – Present
  Technology, AI, Society, and Culture (TASC) team
- ◇ **Research Assistant** at University of Southern California — 2017 – 2022
  Computational Social Science lab
- ◇ **Research Intern** at Google — 2021
  Ethical AI team
- ◇ **Organizer** of the Workshop on Online Abuse and Harm — 2021 – 2024
  NAACL 2024, ACL 2023, NAACL 2022, and ACL 2021
- ◇ **Research Assistant** at Sharif University of Technology — 2014 – 2016
  Ambient Intelligence lab

Education

- ◇ **Ph.D. Computer Science** — 2017 – 2022
  University of Southern California, Los Angeles, USA
- ◇ **M.Sc. Software Engineering** — 2014 – 2017
  Sharif University of Technology, Tehran, Iran
- ◇ **B.Sc. Software Engineering** — 2009 – 2014
  Sharif University of Technology, Tehran, Iran

Peer-Reviewed Publications

- ◇ **Davani, A.**, Díaz, M., Baker, D., Prabhakaran, V. *"D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation"*, EMNLP (2024).
- ◇ **Davani, A.***, Gubbi, S.,* Dev, S., Dave, S., Prabhakaran, V. *"GeniL: A multilingual dataset on generalizing language"*, COLM (2024).
- ◇ **Davani, A.**, Díaz, M., Baker, D., Prabhakaran, V. *"Disentangling perceptions of offensiveness: Cultural and moral correlates"*, FAccT (2024).
- ◇ Prabhakaran, V., Homan, C. M., Aroyo, L., **Davani, A.**, Parrish, A., Taylor, A., Díaz, M., Wang, D., Serapio-García, G. *"GRASP: A disagreement analysis framework to assess group associations in perspectives"*, NAACL (2024).
- ◇ Prabhakaran, V., **Davani, A.**, Ferguson, M. J., Atir, S. *"Distinguishing address vs. reference mentions of personal names in text"*, ACL Findings (2023).
- ◇ Jha, A., **Davani, A.**, Dave, S., Reddy, C., Dev, S., Prabhakaran, V. *"A stereotype benchmark with broad geo-cultural coverage leveraging generative models"*, ACL (2023).
- ◇ Kennedy, B., Golazizian, P., Trager, J., Atari, M., Hoover, J., **Davani, A.**, Dehghani, M. *"The (moral) language of hate"*, PNAS Nexus (2023).
- ◇ Atari, M., Mehl, M. R., Graham, J., Doris, J. M., **Davani, A.**, Omrani, A., ..., Dehghani, M. *"The paucity of morality in everyday talk"*, Scientific Reports (2023).
- ◇ **Davani, A.**, Atari, M., Kennedy. B., Dehghani, M. *"Hate speech classifiers learn normative social stereotypes"*, TACL (2022).
- ◇ Atari, M., Reimer, N. K., Graham, J., Hoover, J., Kennedy, B., **Davani, A.**, Karimi-Malekabadi, F., Birjandi, S., Dehghani, M. *"Pathogens are linked to human moral systems across time and space"*, Current Research in Ecological and Social Psychology (2022).

- Kennedy, B., Atari, M., **Davani, A.**, Yeh, L., Omrani, A., Kim, Y., ..., Hoover, J. *"Introducing the Gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale"*, Language Resources and Evaluation (2022).
- **Davani, A.**, Díaz, M., Prabhakaran, V. *"Dealing with disagreements: Looking beyond the majority vote in subjective annotations"*, TACL (2021).
- Prabhakaran, V.*, **Davani, A.***, Díaz, M. *"On releasing annotator-level labels and information in datasets"*, The 15th Linguistic Annotation & 3rd Designing Meaning Representations Joint Workshop (2021).
- **Davani, A.**, Omrani, A., Kennedy, B., Atari, M., Ren, X., Dehghani, M. *"Improving counterfactual generation for fair hate speech detection"*, WOAH (2021).
- Atari, M., **Davani, A.**, Kogon, D., Kennedy, B., Saxena, N. A., Anderson, I., Dehghani, M. *"Morally homogeneous networks and radicalism"*, Social Psychological and Personality Science (2021).
- Hoover, J., Atari, M.*, **Davani, A.***, Kennedy, B.*, Portillo-Wightman, G., Yeh, L., Dehghani, M. *"Investigating the role of group-based morality in extreme behavioral expressions of prejudice"*, Nature Communications (2021).
- Kennedy, B., Atari, M., **Davani, A.**, Hoover, J., Omrani, A., Graham, J., Dehghani, M. *"Moral concerns are differentially observable in language"*, Cognition (2021).
- Jin, X., Barbieri, F., Kennedy, B., **Davani, A.**, Neves, L., Ren, X. *"On transferability of bias mitigation effects in language model fine-tuning"*, ACL (2021).
- **Davani, A.**, Atari, M., Kennedy, B., Havaldar, S., Dehghani, M. *"Hatred is in the eye of the annotator: Hate speech classifiers learn human-like social stereotypes"*, Conference of the Cognitive Science Society (2020).
- Atari, M., **Davani, A.**, Dehghani, M. *"Body maps of moral concerns"*, Psychological Science (2020).
- Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., **Davani, A.**, Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Chin, J., Leong, C., Leung, J. Y., Mirinjian, A., Dehghani, M. *"Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment"*, Social Psychological and Personality Science (2020).
- Kennedy, B.*, Jin, X.*, **Davani, A.**, Dehghani, M., Ren, X. *"Contextualizing hate speech classifiers with post-hoc explanation"*, ACL (2020).
- **Davani, A.**, Yeh, L., Atari, M., Kennedy, B., Wightman, G. P., Gonzalez, E., Delong, N., Bhatia, R., Mirinjian, A., Ren, X., Dehghani, M. *"Reporting the unreported: Event extraction for analyzing the local representation of hate crimes"*, EMNLP (2019).
- Courtland, M., **Davani, A.**, Reyes, M., Yeh, L., Leung, J., Kennedy, B., Dehghani, M., Zevin, J. *"Modeling performance differences on cognitive tests using LSTMs and skip-thought vectors trained on reported media consumption"*, NLP-CSS (2019).
- **Davani, A.**, Shirehjini, A. A. N., Daraei, S. *"Towards interacting with smarter systems"*, Journal of Ambient Intelligence and Humanized Computing (2018).

Pre-prints

- Trager, J., Ziabari, A., **Mostafazadeh Davani, A.**, Golazazian, P., Karimi-Malekabadi, F., Omrani, A., Li, Z., Kennedy, B., ..., Morteza Dehghani *"The Moral Foundations Reddit Corpus"*, (in preparation)
- Vial, A. C.*, **Mostafazadeh Davani, A.***, Havaldar, S., Chestnut, E. K., Dehghani, M., Cimpian, A. *"Syntactic and semantic gender biases in the language on children's television: Evidence from a corpus of 95 shows from 1960 to 2018"*, (in preparation).
- Goodwin, R. D., Dodson, S. J., Chambers, M., **Mostafazadeh Davani, A.**, Dehghani, M., Graham, J., Diekmann, K. A. *"Twitter observers' moral language reveals how sexual harassment denials condemn #MeToo victims"*, (in preparation).

Skills

- Programming: Python, Java, C++, C#, R
- Deep Learning: Tensorflow, PyTorch, Keras
- Statistics: Hierarchical Modeling, Time Series Analysis