

# 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 – Present  
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

- ◇ Prabhakaran, V., **Mostafazadeh Davani, A.**, Ferguson, M. J., Atir, S. “*Distinguishing Address vs. Reference Mentions of Personal Names in Text*”, ACL Findings (2023).
- ◇ Jha, A., **Mostafazadeh 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., **Mostafazadeh Davani, A.**, Dehghani, M. “*The (Moral) Language of Hate*”, PNAS Nexus (2023).
- ◇ Atari, M., Mehl, M. R., Graham, J., Doris, J. M., **Mostafazadeh Davani, A.**, Omrani, A., Kennedy, B., ..., Dehghani, M. “*The paucity of morality in everyday talk*”, Scientific Reports (2023).
- ◇ **Mostafazadeh 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., **Mostafazadeh 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., **Mostafazadeh 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).
- ◇ **Mostafazadeh Davani, A.**, Díaz, M., Prabhakaran, V. “*Dealing with disagreements: Looking beyond the majority vote in subjective annotations*”, TACL (2021).
- ◇ Prabhakaran, V.\*, **Mostafazadeh Davani, A.\***, Díaz, M. “*On releasing annotator-level labels and information in datasets*”, The 15th Linguistic Annotation & 3rd Designing Meaning Representations Joint Workshop (2021).

- ◇ **Mostafazadeh Davani, A.**, Omrani, A., Kennedy, B., Atari, M., Ren, X., Dehghani, M. “*Improving counterfactual generation for fair hate speech detection*”, Proceedings of the 5th Workshop on Online Abuse and Harms (2021).
- ◇ Atari, M., **Mostafazadeh 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.\*, **Mostafazadeh 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., **Mostafazadeh 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., **Mostafazadeh Davani, A.**, Neves, L., Ren, X. “*On transferability of bias mitigation effects in language model fine-tuning*”, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (2021).
- ◇ **Mostafazadeh Davani, A.**, Atari, M., Kennedy, B., Havaladar, 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., **Mostafazadeh Davani, A.**, Dehghani, M. “*Body maps of moral concerns*”, Psychological Science (2020).
- ◇ Hoover, J., Portillo-Wightman, G., Yeh, L., Havaladar, S., **Mostafazadeh 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.\*, **Mostafazadeh Davani, A.**, Dehghani, M., Ren, X. “*Contextualizing hate speech classifiers with post-hoc explanation*”, In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020).
- ◇ **Mostafazadeh 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*”, In the Proceedings of Empirical Methods in Natural Language Processing (2019).
- ◇ Courtland, M., **Mostafazadeh 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*”, Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science (2019).
- ◇ **Mostafazadeh Davani, A.**, Nazari Shirehjini, A. A., Daraei, S. “*Towards interacting with smarter systems*”, Journal of Ambient Intelligence and Humanized Computing (2018).
- ◇ **Mostafazadeh Davani, A.**, Nazari Shirehjini, A. A. and Daraei, S. “*A Meta user interface for understandable and predictable interaction in AAL.*”, Human Aspects of IT for the Aged Population. Design for Everyday Life. Springer International Publishing, (2015).
- ◇ **Mostafazadeh Davani, A.**, Nazari Shirehjini, A. A., Daraei, S., Khojasteh, N., and Shirmohammadi, S. “*A Meta user interface for interaction with mixed reality environments.*”, Haptic, Audio and Visual Environments and Games (HAVE), 2015 IEEE International Symposium on, IEEE, (2015).

## Skills

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