Asma Ghandeharioun

Google NYC
111 8th Ave
New York, NY 10011

+1(857)303-1203
aghandeharioun@google.com
alum.mit.edu/www/asma_gh
@ghandeharioun

Interests

Interpretability, Language Models, Machine Learning
I am interested in interpreting (language) models and aligning AI with human values. In my research, I tackle questions like *why* models make certain predictions, *where* they store knowledge, and *how* to constrain their generations.

Education

Ph.D. in Media Arts and Sciences, Media Lab, **MIT** (2016 - 2021); *GPA: 5.0/5.0*
Advisor: Rosalind Picard
Thesis: Towards Human-Centered Optimality Criteria.

M.Sc. in Media Arts and Sciences, Media Lab, **MIT** (2014 - 2016); *GPA: 5.0/5.0*
Advisor: Rosalind Picard
Coursework: Machine learning, Statistics for neuroscience, Affective computing, Tools for wellbeing, Behavior change lab.
Thesis: BrightBeat: Effortlessly influencing breathing for cultivating calmness and focus.

B.Sc. in Computer Engineering, Sharif University of Tech. (2009 - 2014)
Thesis: Dr. Tick: An Android Application for Measuring Heart Rate and Respiratory Rate on Smart Phones.

Experience

Google DeepMind, Senior Research Scientist, May 2024 - Present.

Google Research, Research Scientist, Sep. 2021 - Apr. 2024.
Research Intern, Sep. 2019 - Jan. 2020.
Software Engineering Intern, Jun. 2018 - Aug. 2018.

Microsoft Research, Research Intern, Jun. 2017 - Aug. 2017.

MIT Media Lab, Research Assistant, Affective Computing group, Sep. 2014 - Jun. 2021.

Ecole Polyetechnique Federale de Lausanne (EPFL), Research Intern, Jul. 2013 - Sep. 2013.

Sharif University of Technology, Undergraduate Researcher, Sep. 2011 - Jul. 2014.

Publications

See a more complete publication list on google scholar. * equal contribution. ⋄ equal advising.

Preprints

1. **Ghandeharioun, A.***, Yuan, A., Guerard, M., Reif, E., Lepori, M. A., Dixon, L. (2024). Who's asking? User personas and the mechanics of latent misalignment. **Preprint**.
2. Yehudai, G., Kaplan, H., **Ghandeharioun, A.**, Geva, M., Globerson, A. (2024). When Can Transformers Count to n? **Preprint**.

Conference Papers

3. **Ghandeharioun, A.***, Caciularu, A.*, Pearce, A., Dixon, L., Geva, M. (2024). Patchscopes: A unifying framework for inspecting hidden representations of language models. **ICML**.
4. Friedman, D., Lampinen, A. K., Dixon, L., Chen, D., **Ghandeharioun, A.** (2024). Interpretability illusions in the generalization of simplified models. **ICML**.
5. Hase, P., Bansal, M., Kim, B., **Ghandeharioun, A.** (2023). Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models, **NeurIPS (Spotlight - top 3%)**.

6. Krishna, S., Ma, J., Slack, D., **Ghandeharioun, A.**, Singh, S., Lakkaraju, H. (2023). Post hoc explanations of language models can improve language models. **NeurIPS**.
7. **Ghandeharioun, A.**, Kim, B., Li, C., Jou, B., Eoff, B. , Picard, R. (2022). DISSECT: Disentangled Simultaneous Explanations via Concept Traversals, **ICLR**.
8. Jaques, N.*, Shen, J.*, **Ghandeharioun, A.**, Ferguson, C., Lapedriza, A., Jones, N., Gu, S., Picard, R. (2020). Human-centric dialog training via offline reinforcement learning. **EMNLP (Oral)**.
9. Saleh A.*, Jaques N.*, **Ghandeharioun, A.**, Shen, J., Picard, R. (2020). Hierarchical Reinforcement Learning for Open-Domain Dialog, **AAAI (Oral)**.
10. **Ghandeharioun, A.***, Shen, J.*, Jaques N.*, Ferguson, C., Jones, N., Lapedriza, A., Picard, R. (2019). Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems. **NeurIPS**.
11. **Ghandeharioun, A.**, McDuff, D., Czerwinski, M., Rowan, K. (2019). EMMA: An Emotion-Aware Wellbeing Chatbot. **ACII**, IEEE.
12. **Ghandeharioun, A.**, McDuff, D., Czerwinski, M., Rowan, K. (2019). Towards Understanding Emotional Intelligence for Behavior Change Chatbots. **ACII**, IEEE.
13. Leslie, G.*, **Ghandeharioun, A.***, Zhou, D., Picard, R. (2019). Engineering Music to Slow Breathing and Invite Relaxed Physiology. **ACII**, IEEE.
14. Saeedi, A., Hoffman, M., DiVerdi, S., **Ghandeharioun, A.**, Johnson, M., Adams, R. (2018). Multimodal Prediction and Personalization of Photo Edits with Deep Generative Models. **AISTATS**.
15. **Ghandeharioun, A.**, Fedor, S., Sangermano, L., Ionescu, D., Alpert, J., Dale, C., Sontag, D., & Picard, R. (2017). Objective assessment of depressive symptoms with machine learning and wearable sensors data. **ACII**. IEEE.
16. Jaques, N., Taylor, S., Azaria, A., **Ghandeharioun, A.**, Sano, A., & Picard, R. (2015). Predicting students' happiness from physiology, phone, mobility, and behavioral data. **ACII**. IEEE.

Workshop Papers

17. Hussein, N.*, **Ghandeharioun, A.***, Hussein, N., Mullins, R., Reif, E., Wilson, J., Thain, N.$^{\diamond}$, Dixon, L$^{\diamond}$. (2024). Can Large Language Models explain Their Internal Mechanisms? **IEEE VISxAI**
18. Pearce, A.*, **Ghandeharioun, A.***, Hussein, N., Thain, N., Wattenberg, M., Dixon, L. (2023). Do machine learning models memorize or generalize. **IEEE VISxAI (Best paper).**
19. Friedman, D., Lampinen, A. K., Dixon, L., Chen, D., **Ghandeharioun, A.** (2023). Comparing Representational and Functional Similarity in Small Transformer Language Models. *UniReps: the First Workshop on Unifying Representations in Neural Models*, **NeurIPS Workshop (Oral)**.
20. Jaques N., **Ghandeharioun, A.**, Shen, J., Ferguson, C., Jones, N., Lapedriza, A., Gu, S., Picard, R. (2019). Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog, *Conversational AI*, **NeurIPS workshop**.
21. Saleh A.*, Jaques N.*, **Ghandeharioun, A.**, Shen, J., Picard, R. (2019). Hierarchical Reinforcement Learning for Open-Domain Dialog, *Conversational AI*, **NeurIPS workshop (Best paper nominee)**.
22. **Ghandeharioun, A.**, Eoff, B., Jou, B., Picard, R. (2019). Characterizing Sources of Uncertainty for Improving Calibration and Disambiguating Annotator and Data Bias, **ICCV Workshop (Oral)**.
23. Saeedi, A., **Ghandeharioun, A.**, Hoffman, M. (2015). A simple hierarchical infinite HMM with efficient inference. **NeurIPS Workshop** (Bayesian Nonparametrics: The Next Generation).
24. Jones, N., Jaques, N., Pataranutaporn, P., **Ghandeharioun, A.**, Picard, R. (2019). Analysis of Online Suicide Risk with Document Embeddings and Latent Dirichlet Allocation, **ACII Workshop**.
25. Alizadehsani, R., Hosseini, M. J., Sani, Z. A., **Ghandeharioun, A.**, & Boghrati, R. (2012). Diagnosis of coronary artery disease using cost-sensitive algorithms. **ICDM Workshops**. IEEE.

Journal Papers

26. **Ghandeharioun, A.**, Azaria, A., Taylor, S., Picard, R. W. (2016). "Kind and grateful": a context-sensitive smartphone app utilizing inspirational content to promote gratitude. **Psychology of Well-being**.
27. Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., **Ghandeharioun, A.**, Bahadorian, B., Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. **Computer Methods and Programs in Biomedicine**, 111(1), 52-61.
28. Alizadehsani, R., Jafar Habibi, J., Sani, Z. A., Mashayekhi, H., Boghrati, R., **Ghandeharioun, A.**, Khozeimeh, F., Alizadeh-Sani, F. (2013). Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features. Research in cardiovascular medicine 2, no. 3: 133-139.
29. Alizadehsani R., Habibi J., Bahadorian B., Mashayekhi H., **Ghandeharioun, A.**, Boghrati, R., Sani Z.A. (2012). Diagnosis of coronary arteries stenosis using data mining. Journal of Medical Signals & Sensors. 2012 Jul 1;2(3):153-9.
30. Alizadehsani, R., Habibi, J., Sani, Z.A., Mashayekhi, H., Boghrati, R., **Ghandeharioun, A.**, Bahadorian, B. (2012). Diagnosis of coronary artery disease using data mining based on lab data and echo features. Journal of Medical and Bioengineering, 1(1).
31. Alizadehsani, R., Hosseini, M. J., Boghrati, R., **Ghandeharioun, A.**, Khozeimeh, F., Sani, Z. A. (2012). Exerting cost-sensitive and feature creation algorithms for coronary artery disease diagnosis. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 3(1), 59-79.

Extended Abstracts

32. **Ghandeharioun, A.**, Picard, R. (2017). BrightBeat: Effortlessly Influencing Breathing for Cultivating Calmness and Focus. **CHI Extended Abstracts**. ACM.
33. Howe, E., Nauphal, M., Shapero, B., Bentley, K., Mischoulon, D., **Ghandeharioun, A.**, Fedor, S., Picard, R. Pedrelli, P. (2018, November). Depression and Emotional Reactivity: A Closer Examination of Daily Variations in Affect. **Association for Behavioral and Cognitive Therapies Annual Convention (ABCT)**, Washington D.C.
34. Howe, E., **Ghandeharioun, A.**, Pedrelli, P., Mischoulon, D., Picard, R., Fedor, S. (2017, November). Location Patterns from Phone Sensors May Help Predict Depressive Symptoms: A Longitudinal Pilot Study. **Association for Behavioral and Cognitive Therapies Annual Convention (ABCT) – Tech SIG**, San Diego, CA.
35. Pedrelli, P., Howe, E., Mischoulon, D., Picard, R., **Ghandeharioun, A.**, Fedor, S. (2017, October). Integrating EMA, clinical assessment and wearable sensors to examine the association between MDD and alcohol use. **Connected Health Conference (CHC)**, Boston, MA.
36. **Ghandeharioun, A.**, Fedor, S., Sangermano, L., Alpert, J., Dale, C., Ionescu, D., Picard, R. (2017, May). Location Variability from Commodity Phone Sensors Is Negatively Associated with Self-Reported Depression Score: A Pilot Study. **Association for Psychological Science (APS)**, Boston, MA.
37. **Ghandeharioun, A.**, Sangermano, L., Picard, R., Alpert, J., Dale, C., Ionescu, D., Fedor, S. (2017, April). Objective vs. Subjective Reports of Sleep Quality in Major Depressive Disorder: A Pilot Study. **Anxiety and Depression Association of America (ADAA)**, San Fransisco.
38. Sangermano, L., **Ghandeharioun, A.**, Picard, R., Alpert, J., Dale, C., Fedor, S., Ionescu, D. (2017, April). Cell Phone Data as a Potential Predictor of Depression Severity: A Pilot Study. **Anxiety and Depression Association of America (ADAA)**, San Fransisco.
39. **Ghandeharioun, A.**, Azaria, A., Taylor, S., Maes, P., Picard, R. (2016), Promoting kindness and gratitude with a smartphone and triggers, **Annals of Behavioral Medicine**, 50 (Supp. 1), 266.
40. Taylor, S., Jaques, N., Sano, A., Azaria, A., **Ghandeharioun, A.**, Picard, R. (2016, June). Machine Learning of Sleep and Wake Behaviors to Classify Self-Reported Evening Mood, **SLEEP**.

Theses

41. **Ghandeharioun, A.** Towards Human-Centered Optimality Criteria. (2021, May). *MIT PhD Thesis*.

42. **Ghandeharioun, A.** BrightBeat: Effortlessly influencing breathing for cultivating calmness and focus. (2016, August). *MIT Master's Thesis.*
43. **Ghandeharioun, A.** Dr. Tick: An Android Application for Measuring Heart Rate and Respiratory Rate on Smart Phones. (2014, June). *Bachelor's Thesis.*

Patents

Fedor, S., **Ghandeharioun, A.**, & Picard, R., Ionescu, D. (2018, October) Methods and apparatus for assessing depression. *US 2019/0117143 A1.* Pending patent.

Honors and Awards

2.5M grant from **NIH**, 5R01MH118274, 2019-2023 (PIs: P. Pedrelli, R. Picard).

150K grant from **J-Clinic** for conducting machine learning in healthcare research, 2019 (PI: R. Picard).

D. E. Shaw Zenith Fellowship, 2021.

MIT Quest for Intelligence, MIT Stephen A. Schwarzman College of Computing, Machine Learning Across Disciplines Challenge, recipient of unlimited Google Cloud Platform credit, 2019.

Silver Medal in Iranian National Olympiad in Informatics, 2008.

Recipient of the grant for undergraduate studies from the Iranian **National Elites Foundation**, for outstanding academic success, 2009 - 2014.

Summer Internship fellowship from Human Computer Interaction group, Ecole Polyetechnique Federale de Lausanne, Lausanne, Switzerland, 2013.

Press

Wall Street Journal (2018, December). Once More With Feeling: Teaching Empathy to Machines.

Wired (2019, January). Wearable medical tech is about to become crucial for staying alive.

New Scientist (2019, July). Smartwatch app that soothes the nerves helps improve exam results.

Invited Talks and Panels

Speaker, Oxford Machine Learning Summer School (OxML 2024), AI Alignment via Interpretability: Illusions and Opportunities, 2024.

Speaker, ICML 2024 Workshop on Mechanistic Interpretability, Model Interpretability: from Illusions to Opportunities, 2024.

Panelist, Building trustworthy AI,Women in machine learning workshop at ICML 2024.

Speaker, Harvard University, Grounded understanding and controlling (language) models, 2024.

Speaker, Google, Patchscopes: A unifying framework for inspecting hidden representations of language models, 2024.

Speaker, Google, When interpretability does (not) lead to better control, 2023.

Speaker and Panelist, Future of Intelligent Communications, MIT Design Lab Workshop, 2019.

Speaker and Panelist, Duality's end conference: Computational psychiatry and the Cognitive Science of Representation, Improving Psychological Wellbeing Using Ubiquitous Technologies and AI, 2018.

Speaker, MGH-MIT Strategic Partnership Grand Challenge Grants presentation, Noninvasive Physiologic Sensors to Assess Depression, with Jonathan Alpert, M.D., 2016.

Speaker, MGH Depression Clinical and Research program, Affective computing and mental wellbeing. with Rosalind Picard Sc.D. and Sara Taylor S.M, 2016.

Professional service

Chair/Organizer:

- R2HCAI: Representation Learning for Responsible Human-Centric AI, AAAI, 2023.

Senior Area Chair, 2024-present:

- Conference on Neural Information Processing Systems **(NeurIPS)**.

Area Chair/Senior Program Committee, 2023-present:

- Conference on Neural Information Processing Systems **(NeurIPS)** **(Outstanding AC)**
- International Conference on Learning Representations **(ICLR)**
- Conference on Language Modeling **(COLM)**

- Affective Computing & Intelligent Interaction **(ACII)**

Reviewer/Program Committee, 2015-present:

- **Machine Learning**:
  - Conference on Neural Information Processing Systems **(NeurIPS)**
  - International Conference on Machine Learning **(ICML)**
  - International Conference on Learning Representations **(ICLR)**
  - AAAI Conference on Artificial Intelligence **(AAAI)**
  - International Joint Conference on Artificial Intelligence **(IJCAI)**
- **ML Applications**
  - Affective Computing & Intelligent Interaction (ACII)
  - IEEE journal of biomedical and health informatics (JBHI)
- **Human-Computer Interaction**:
  - ACM Conference on Human Factors in Computing Systems (CHI) **(Excellent Reviewer)**
  - ACM Designing Interactive Systems (DIS)
  - Transactions on Computer-Human Interaction (TOCHI)
  - Psychology of Well-Being Journal, Interactive, Mobile, Wearable and Ubiquitous Technologies journal (IMWUT)
  - Ubiquitous Computing (UbiComp)

Operations Officer, Ashdown graduate residency, technology subcommittee, 2018-2019.

(Officer of the Month, Aug. 2019)

Teaching and Mentorship

Mentor for the PhD Fellowship program at Google, 2024.

Mentor, interpretability and explainability round table, Women in machine learning workshop at ICML 2024.

gMentor Internal Google mentorship program, 2022-present.

Mentor for PhD interns, 2021-present.

Peter Hase, Dan Friedman, Michael Lepori.

Advisor MIT Alumni Advisors Hub, 2021-present.

Mentor for Master of Engineering (MEng) & Undergraduate Research Opportunities Program (UROP), 2016-2021:

Darian Bhathena, Alexander Lynch, Diane Zhou, Marek Subernat.

Guest lecturer, 2017-2018:

MAS.630 Affective Computing.

Students Offering Support: Assisting underrepresented students applying to the Media Lab, 2017.