# MAX KLEIMAN-WEINER

maxkw@uw.edu & web

## Education

| | | |
|---|---|---|
| Ph.D. | Massachusetts Institute of Technology<br>Brain and Cognitive Sciences<br>Thesis: *Computational Foundations of Human Social Intelligence*<br>Advisor: Josh Tenenbaum<br>Committee: Drazen Prelec, Rebecca Saxe, Fiery Cushman | 2012–2018 |
| 2x MSc | University of Oxford, Merton College<br>Applied Statistics & Experimental Psychology<br>Thesis: *Hierarchical Neural Network Models For Decision Making*<br>Advisor: Tim Behrens | 2010, 2012 |
| BS | Stanford University<br>Biological Sciences and honors in Neuroscience<br>Thesis: *Synergistic roles of $GABA_A$ receptors and SK channels in regulating thalamocortical oscillations*<br>Advisor: John Huguenard | 2009 |

## Academic Positions

University of Washington

| | |
|---|---|
| 2024 - | Assistant Professor, Foster School of Business |
| 2024 - | Adjunct Assistant Professor, Computer Science and Engineering |

Harvard University

| | |
|---|---|
| 2021-2024 | Associate, School of Engineering and Applied Sciences |
| 2018-2020 | Fellow in DSI, CRCS & MBB |

Massachusetts Institute of Technology

| | |
|---|---|
| 2018-2020 | Research Scientist |

## Work Experience

| | |
|---|---|
| 2020- | Common Sense Machines, Co-Founder & Chief Scientist |
| 2019 | Salesforce, Principal Researcher |
| 2012-2019 | Diffeo, Co-Founder & Chief Scientist [Acquired by Salesforce] |
| 2011-2012 | Chinese Academy of Sciences & REAP, Fulbright Fellow |
| 2010 | McKinsey & Company, Summer Associate |

## Awards

| | |
|---|---|
| 2020 | Best Paper Modeling Prize for Higher Cognition – Cognitive Science Society |
| 2020 | Best Paper Award – Cooperative AI Workshop, NeurIPS |
| 2019 | Glushko Dissertation Prize – Cognitive Science Society ($10,000 Prize) |
| 2017 | Best Paper (1st of 200+) – Reinforcement Learning and Decision Making |
| 2017 | William James Prize (best paper) – Society for Philosophy and Psychology |
| 2016 | MIT Pokerbots 1st of 22 ($10,000 prize) |
| 2016 | Angus MacDonald Award for Excellence in Undergraduate Teaching |
| 2015 | MIT Pokerbots 1st of 38 ($10,000 prize) |
| 2015,16,18 | Cognitive Science Society Glushko Student Travel Award (merit based) |
| 2015 | Psychonomics Poster Finalist |
| 2013 | Newman Entrepreneurial Initiative ($25,000 award) |
| 2009 | The Deans' Award (1 of 8, highest academic award) |
| 2009 | Firestone Medal (1 of 34, highest research award) |
| 2009 | Phi Beta Kappa and departmental distinction |

## Fellowships

| | |
|---|---|
| 2011-2017 | Hertz Foundation Graduate Fellowship |
| 2011-2012 | Fulbright Research Fellowship (China) |
| 2009-2014 | NSF Graduate Research Fellowship |
| 2009-2011 | Marshall Scholar |
| 2008 | Barry M. Goldwater Scholar |
| 2008 | Irene and Eric Simon Brain Research Foundation Student Fellow |

## Research Grants (Co-PI)

| | |
|---|---|
| 2020- | Templeton World Charity Foundation, The Cognitive Foundations of Social Minds (w/ Josh Tenenbaum, Francine Dolins, Richard Lewis, Josep Call), $1,000,000 |
| 2018-20 | DARPA, Ground Truth, Social MIND: Social Machine Intelligence for Novel Discovery (w/ Josh Tenenbaum & James Allen Evans, UChicago), $355,000 |
| 2018-20 | Future of Life Institute, Reverse-engineering fair cooperation (w/ Josh Tenenbaum), $150,000 |
| 2018-21 | Templeton World Charity Foundation, Diverse Intelligences Initiative, Reverse-engineering the moral mind (w/ Josh Tenenbaum), $228,250 |
| 2014-17 | DARPA, Memex, Maximizing Coreference Resolution with Efficient Human Input, $2,600,000 |

## Short Courses

| | |
|---|---|
| 2014 | Machine Learning Summer School (MLSS) |
| 2013 | Santa Fe Institute (SFI) - Complex Systems Summer School |
| 2011 | Inter-University Program (IUP) for Chinese Language Studies |

## Publications

Google Scholar Link: scholar.google.com/citations?hl=en&user=SACXQKYAAAAJ

Rong, F., **Kleiman-Weiner, M.** (2024). Value Internalization: Learning and Generalizing from Social Reward. *Reinforcement Learning Conference (RLC)* [CogSci Oral]

Levine, S., **Kleiman-Weiner, M.,** Chater, N., Cushman, F., Tenenbaum, J. (2024) When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *Cognition.*

McManus, R.M., Fong, H.P., **Kleiman-Weiner, M.**, Young, L. (2024) Most people do not "value the struggle": Tempted agents are judged as less virtuous than those who were never tempted. *Journal of Experimental Social Psychology*

Ma, M., Liu J., Sokota S., **Kleiman-Weiner, M.,** Foerster J. (2023). Learning Intuitive Policies Using Action Features. *International Conference on Machine Learning (ICML)*

Houlihan, S.D., **Kleiman-Weiner, M.**, Hewitt, L.B., Tenenbaum, J.B., Saxe R. (2013) Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A*

Kraft-Todd, G., **Kleiman-Weiner, M.,** Young, L. (2023) Observability Reduces Moral Actors' Perceived Virtue. *Open Mind*

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Zhiheng, L., Blin, K., Adauto, F., **Kleiman-Weiner, M.,**, Sachan, M., Schölkopf B. (2023) Cladder: Assessing causal reasoning in language models *Neural Information Processing Systems (NeurIPS)*

Kraft-Todd, G., **Kleiman-Weiner, M.,** Young, L. (2023) Assessing and dissociating virtues from the 'bottom up': A case study of generosity vs. fairness. *The Journal of Positive Psychology.*

Stacy, S., Parab, A., **Kleiman-Weiner, M.,** Gao, T., (2022) Overloaded Communication as Paternalistic Helping *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

*Wang, R.E., *Wu, S.A., Evans, J.A., Tenenbaum, J.B., Parkes, D.C., **Kleiman-Weiner, M.** (2021) Too many cooks: Coordinating multi-agent collaboration through inverse planning *Topics in Cognitive Science.* [paper prize]

Levine, S., **Kleiman-Weiner, M.**, Schulz, L., Tenenbaum, J.B., Cushman, F. (2020) The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences.*

McManus, R.M., **Kleiman-Weiner, M.**, Young, L. (2020) What we owe to family: The impact of special obligations on moral judgment. *Psychological Science.*

Awad, E., Levine, S., **Kleiman-Weiner, M.**, Dsouza, S., Tenenbaum, J.B., Shariff, A., Bonnefon, J., & Rahwan, I. (2020) Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behavior.*

***Kleiman-Weiner, M.**, *Sosa, F., Thompson, B., Opheusden, S., Griffiths, T., Gershman S., Cushman, F. (2020) Downloading Culture.zip: Social learning by program induction. *Proceedings of the 42th Annual Conference of the Cognitive Science Society.*

Lu, A.C., Kyuyoung, C.L., **Kleiman-Weiner, M.**, Truong, T., Wang, M., Huguenard, J.R., Beenhakker, M.P., (2020) Nonlinearities between inhibition and T-type calcium channel activity bidirectionally regulate thalamic oscillations. *Elife.*

*Serrino, J., ***Kleiman-Weiner, M.**, Parkes, C. D., & Tenenbaum, J. B. (2019) Finding friend and foe in multi-agent games. *Neural Information Processing Systems (NeurIPS)* (* indicates equal contribution) [spotlight, top 3%]

*Shum, M., ***Kleiman-Weiner, M.**, Littman, M. L., & Tenenbaum, J. B. (2019) Theory of Minds: Understanding Behavior in Groups Through Inverse Planning. *(AAAI)* (* indicates equal contribution) [oral]

Strouse, D., **Kleiman-Weiner, M.**, Tenenbaum, J.B., Botvinick, M., Schwab, D. (2018) Learning to share and hide intentions using information regularization. *Neural Information Processing Systems (NeurIPS).*

Cao, J., **Kleiman-Weiner, M.**, & Banaji, M.R. (2018). People make the Bayesian judgment they criticize in others. *Psychological Science.*

**Kleiman-Weiner, M.**, Tenenbaum, J. B., & Zhou, P. (2018). Non-parametric Bayesian inference of strategies in infinitely repeated games. *Econometrics Journal.*

Gerstenberg, T., Ullman, T. D., Nagel, J., **Kleiman-Weiner, M.**, Lagnado, D. A. & Tenenbaum, J. B. (2018). Lucky or clever? From changed expectations to attributions of responsibility. *Cognition.*

Kim R., **Kleiman-Weiner M.**, Abeliuk A., Awad E., Dsouza S., Tenenbaum J.B.. & Rahwan I. (2018). A Computational Model of Commonsense Moral Decision Making. *AAAI/ACM: AI, Ethics, and Society.*

Halpern, J.Y., **Kleiman-Weiner, M.** (2018). Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility. *AAAI.* [oral]

Cao, J., **Kleiman-Weiner, M.**, & Banaji, M.R. (2017). Statistically inaccurate and morally unfair judgments via base rate intrusion. *Nature Human Behavior,* 1(10), 738.

**Kleiman-Weiner, M.**, Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition.*

**Kleiman-Weiner, M.**, Shaw, A., & Tenenbaum, J. B. (2017). Constructing Social Preferences From Anticipated Judgments: When Impartial Inequity is Fair and Why? *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* [oral]

**Kleiman-Weiner, M.**, Ho, M., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* [oral]

Ho, M., MacGlashan, J., Greenwald, A., Littman, M. L., Hilliard, E. M., Trimbach, C., Stephen, B., Tenenbaum, J. B., **Kleiman-Weiner, M.**, & Austerweil, J. L. (2016). Feature-based joint planning and norm learning in collaborative games. *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

**Kleiman-Weiner, M.**, Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* [oral]

Allen, K., Jara-Ettinger, J., Gerstenberg, T., **Kleiman-Weiner, M.**, & Tenenbaum, J. B. (2015). Go fishing! responsibility judgments when cooperation breaks down. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Gerstenberg, T., Ullman, T. D., **Kleiman-Weiner, M.**, Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements *Proceedings of the 36th Annual Conference of the Cognitive Science Society.*

Frank, J.R., **Kleiman-Weiner**, M., Roberts, D.A., Voorhees, E., & Soboroff, I. (2014). Evaluating stream filtering for entity profile updates in TREC 2012, 2013, and 2014 *(KBA Track Overview, Notebook Paper)*

Frank, J. R., Bauer, S. J., **Kleiman-Weiner, M.**, Roberts, D. A., Tripuraneni, N., Zhang, C., Ré, C., Voorhees, E., & Soboroff, I. (2013). *Evaluating Stream Filtering for Entity Profile Updates for TREC 2013 (KBA Track Overview).*

**Kleiman-Weiner, M.**, Luo, R., Zhang, L., Shi, Y., Medina, A., & Rozelle, S. (2013). Eggs versus chewable vitamins: which intervention can increase nutrition and test scores in rural china? *China Economic Review*, 24, 165–176.

Zhang, L., **Kleiman-Weiner, M.**, Luo, R., Shi, Y., Martorell, R., Medina, A., & Rozelle, S. (2013). Multiple micronutrient supplementation reduces anemia and anxiety in rural China's elementary school children. *The Journal of Nutrition*, 143(5), 640– 647.

Frank, J. R., **Kleiman-Weiner, M.**, Roberts, D. A., Niu, F., Zhang, C., Ré, C., & Soboroff, I. (2012). Building an entity-centric stream filtering test collection for TREC 2012. *Proceedings of the Text Retrieval Conference (TREC).*

Cepeda, C., Cummings, D. M., Hickey, M. A., **Kleiman-Weiner, M.**, Chen, J. Y., Watson, J. B., & Levine, M. S. (2010). Rescuing the corticostriatal synaptic disconnection in the R6/2 mouse model of Huntington's disease: exercise, adenosine receptors and ampakines. *PLoS Currents*, 2.

Luo, R., **Kleiman-Weiner, M.**, Rozelle, S., Zhang, L., Liu, C., Sharbono, B., Shi, Y., & Lee, M. (2010). Anemia in rural China's elementary schools: prevalence and correlates in Shaanxi province's poor counties. *Ecology of Food and Nutrition*, 49(5), 357–372.

**Kleiman-Weiner, M.**, Beenhakker, M. P., Segal, W. A., & Huguenard, J. R. (2009). Synergistic roles of GABAA receptors and SK channels in regulating thalamocortical oscillations. *Journal of Neurophysiology*, 102(1), 203–213.

Schofield, C. M., **Kleiman-Weiner, M.**, Rudolph, U., & Huguenard, J. R. (2009). A gain in GABAa receptor synaptic strength in thalamus reduces oscillatory activity and absence seizures. *Proceedings of the National Academy of Sciences*, 106 (18), 7630– 7635.

Cepeda, C., André, V. M., Yamazaki, I., Wu, N., **Kleiman-Weiner, M.**, & Levine, M. S. (2008). Differential electrophysiological properties of dopamine D1 and D2 receptor-containing striatal medium-sized spiny neurons. *European Journal of Neuroscience*, 27(3), 671–682.

**Kleiman-Weiner, M.**, & Berger, J. (2006). The sound of one arm swinging: a model for multidimensional auditory display of physical motion. *Proceedings of the 12th International Conference on Auditory Display.*

## Preprint

Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., Talamadupula, K., Tenenbaum J. **Kleiman-Weiner, M.** (2022). When Is It Acceptable to Break the Rules? Knowledge Representation of Moral Judgement Based on Empirical Data. *arXiv.*

Stacy, S., Li, C., Zhao, M., Yun, Y., Zhao, Q., **Kleiman-Weiner, M.,** Gao T. (2021). Modeling Communication to Coordinate Perspectives in Cooperation. *arXiv.*

Kryven, M., Yu, S., **Kleiman-Weiner, M.,** Tenenbaum, J. (2021) Planning ahead in spatial search. *PsyArXiv.*

## Patents

Pavlini, E.B., Briggs J.R., **Kleiman-Weiner, Max**, Frank, F.R., "Systems and method for investigating relationships among entities," 2021. US Patent App 17/133,764

**Kleiman-Weiner, Max**, *et al.* "Knowledge operating system," 2020. US Patent 10,839,021.

Roberts, D.A., **Kleiman-Weiner, M.**, Frank, J.R., *et al.*, "Entity-centric knowledge discovery," 2016. US Patent 9,275,132.

## Teaching

| | |
|---|---|
| 2016 | MIT TA: Statistical Learning Theory and Applications |
| 2013, 14, 15 | MIT TA: Computational Cognitive Science |
| 2009 | Stanford TA: Economic Development of Greater China |
| 2008 | Stanford Lecturer: Current Debates in Neuroscience |

## Supervised Students

**PhD:** Essie (Suhyoun) Yu (2018-2022, Amazon)

**Masters of Engineering:** Sean Anderson (2021-2023, PhD Stanford) Sunayana Rane (2018-2020, PhD Princeton), Luana Lopes Lara (2018-2019, Founder Kalshi), Jack Serrino (2018-2019, Hudson River Trading), Michael Shum (2017-2018, Schwarzman Scholar), Lily Zhang (2017-2018, SWE GRAIL)

**Undergraduate:** Sarah Wu (2019-2020, PhD Stanford), Rose Wang (2019-2020, PhD Stanford), John Muchovej (2019-2020, PhD Yale), William Long (2018-2019, Founder Numinar Analytics), Suproteem Sarkar (2018-2019, PhD Harvard), Alyssa Dayan (2018, PhD Berkeley), Penghui Zhou (2015-2016, DE Shaw), Daniel Lerner (2015-2016), Suzanne A Mueller (2015-2016), Erwin Hilton (Summer 2015), Max Maybury (Spring 2015), Paul Masterson (Spring 2015), Alejandro Vientos (Summer 2014, 2016-2018, PhD Rutgers), Max Stein-Golenbock (Spring 2014), Drew Drechsler (Fall 2013)

| | | |
|---|---|---|
| Workshops Organized | 2024 | Mathematics of Intelligences (Long), IPAM, UCLA |
| | 2022 | Mathematics of Collective Intelligence (Short), IPAM, UCLA |
| | 2017 | Cooperative Social Intelligence Workshop, CogSci |
| | 2011-14 | Knowledge Base Acceleration (KBA), Organizer |

| | | |
|---|---|---|
| Invited Presentations (Selected) | 2024 | UW-UBC Marketing |
| | 2023 | Max Planck Research Group Dynamics of Social Behavior |
| | 2022 | Foundations and Frontiers in Cognitive Science (U Michigan) |
| | 2022 | Beneficial AI Seminar, Berkeley |
| | 2021 | Center for Human-Compatible AI Conference (CHAI), Berkeley |
| | 2021 | Evolution of Social Complexity Colloquium, ASU |
| | 2020 | Center for Human-Compatible AI Conference (CHAI), Berkeley |
| | 2020 | Machine Learning Special Interest Group, Lincoln Laboratory |
| | 2019 | Diverse Intelligences Summit, St. Andrews |
| | 2019 | EconCS Seminar, Harvard |
| | 2019 | Center for Research on Computation and Society, Harvard |
| | 2019 | MI21 Human Like Technologies, London |
| | 2019 | Center for Human-Compatible AI Conference (CHAI), Berkeley |
| | 2018 | Leading Integrity, Warwick Business School, London |
| | 2018 | O'Reilly Artificial Intelligence Conference, NYC |
| | 2018 | Distinguished Speaker, Accelerated Discovery Forum, IBM Research (Almaden) |
| | 2018 | Boston College, Carroll School of Management, JDM Day |
| | 2018 | Lee Lab (Prof. Daeyeol Lee), Yale |
| | 2017 | Facebook AI (FAIR), New York |
| | 2017 | Human Cooperation Lab (Prof. David Rand), Yale |
| | 2017 | Morality, Language and Thought Workshop, Institut Jean Nicod |
| | 2017 | Boston University, Questrom School of Business, JDM Day |
| | 2017 | Scalable Cooperation Group (Prof. Iyad Rahwan), MIT Media Lab |
| | 2017 | Social Cognitive Neuroscience Lab (Prof. Rebecca Saxe), MIT BCS |
| | 2017 | MIT Cognitive Lunch |
| | 2016 | Workshop on Physical & Social Scene Understanding, CogSci |
| | 2016 | Workshop on Learning, Inference and Control of Multi-Agent Systems, NIPS |
| | 2016 | Organizational Economics Lunch, MIT Sloan |
| | 2016 | Cooperation and Self-Control Workshop, London |
| | 2016 | London Judgement and Decision Making Seminar, UCL |
| | 2016 | DeepMind, London |
| | 2016 | Boston College, Carroll School of Management, JDM Day |
| | 2016 | Morality Lab (Prof. Liane Young), Boston College |
| | 2016 | Computational Cognitive Neuroscience Lab (Prof. Sam Gershman), Harvard |
| | 2016 | Computation & Cognition Lab (Prof. Noah Goodman), Stanford |
| | 2015 | Brown University, CLPS, Cognition Seminar Series |
| | 2015 | Shaw Lab (Prof. Alex Shaw), University of Chicago |
| | 2015 | Boston Area Moral Cognition Group |
| | 2015 | Affective Brain Lab (Prof. Tali Sharot), UCL/MIT |
| | 2015 | Scalable Cooperation Group (Prof. Iyad Rahwan), MIT Media Lab |
| | 2015 | MIT Cognitive Lunch |
| | 2015 | Moral Psychology Research Lab (Prof. Fiery Cushman & Josh Greene), Harvard |
| | 2015 | Computation & Cognition Lab (Prof. Noah Goodman), Stanford |
| | 2015 | Northeastern Undergraduate Researchers of Neuroscience |
| | 2009 | Achauer Honors Symposium (Stanford) |

| | |
|---|---|
| Reviewer | Psychological Review, Proceedings of the National Academy of Sciences (PNAS), Cog- |

nition, Cognitive Science, Open Mind, PLOS Computational Biology, Artificial Intelligence, ACR, NeurIPS, CogSci

MISC

Citizenship: USA
Languages: English, Mandarin Chinese